# Modelos de Regressão e Aplicações

Departamento de Estatística e Matemática Aplicada/CC/UFC

Laboratório de Inovação em Estatística - StatLab

**Grupo de pesquisa Modelos de Regressão e aplicações**

**Versão Preliminar**
**- Em Elaboração -**

# Índice

# Informações Legais e Declaração de Uso de Inteligência Artificial

## Direitos Autorais e Uso da Obra



## Declaração sobre o Uso de Ferramentas de Inteligência Artificial

Durante a elaboração deste livro foram utilizadas ferramentas de Inteligência Artificial, em especial sistemas de apoio à escrita e organização textual, como recurso complementar no processo de produção acadêmica.

O uso de IA teve caráter **estritamente assistivo**, não substituindo o desenvolvimento conceitual, a formulação matemática, a interpretação estatística nem as decisões pedagógicas da obra. Todo conteúdo foi cuidadosamente revisado, validado e adaptado pelos autores, que assumem integral responsabilidade científica e intelectual pelo texto final.

Não foram delegadas à IA decisões metodológicas, inferenciais ou conclusões analíticas. A elaboração teórica, a seleção de exemplos, a validação matemática e a coerência didática resultam da experiência acadêmica e da atuação dos autores na área de Modelos de Regressão.

Reafirmamos que o uso responsável de tecnologias emergentes deve sempre preservar o rigor científico, a integridade acadêmica e o protagonismo intelectual humano.

# Prefácio

Este livro resulta da experiência acumulada no ensino, na orientação de estudantes e no desenvolvimento de pesquisas em modelos de regressão. Somos docentes pesquisadores do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará (DEMA/UFC), com atuação contínua em regressão linear, modelagem estatística e inferência. Integramos o Laboratório de Inovação em Estatística — StatLab/UFC (https://statlab.quarto.pub/) e o Grupo de Pesquisa *Modelos de Regressão e Aplicações* do CNPq (http://dgp.cnpq.br/dgp/espelhogrupo/6344659534806942), espaços nos quais articulamos ensino, pesquisa e aplicação a problemas reais.

O material tem como base as disciplinas Modelos de Regressão I (CC0290) e Modelagem Estatística (CC0452), momentos centrais na formação do estudante, em que probabilidade, inferência e álgebra matricial se integram em uma estrutura unificada de modelagem. Ao longo dos anos, consolidamos a convicção de que ensinar regressão exige equilíbrio entre rigor teórico e clareza interpretativa.

O livro desenvolve o Modelo de Regressão Linear Simples e o Modelo de Regressão Linear Múltipla de forma progressiva, enfatizando interpretação, hipóteses, propriedades dos estimadores, inferência clássica e diagnóstico. A implementação é realizada integralmente no **R**, utilizado como ambiente de experimentação conceitual e não apenas como ferramenta computacional.

Os fundamentos matemáticos mais densos são apresentados em apêndices específicos, preservando o fluxo didático do texto principal e permitindo diferentes níveis de aprofundamento. Acreditamos que a regressão deve ser ensinada como modelo estatístico, com atenção às suas suposições, limitações e implicações práticas.

Esperamos que este livro contribua para a formação de profissionais capazes de utilizar modelos de regressão com rigor técnico, pensamento estatístico crítico e responsabilidade analítica.

# Parte I

# Parte I — Modelagem

# 1 Introdução e Panorama dos Modelos de Regressão

Este livro é dedicado aos estudos de **Modelos de Regressão**, desde a formulação clássica até extensões modernas. O material apoia as disciplinas Modelos de Regressão I (CC0290) e Modelagem Estatística (CC0452), ambas com implementação no software R. Ao longo do semestre, o estudante terá contato tanto com a fundamentação matemática e estatística quanto com aplicações práticas em diferentes áreas, utilizando ferramentas computacionais para explorar dados reais.

O curso está estruturado de forma progressiva: parte-se da regressão linear simples, como porta de entrada ao raciocínio de modelagem, avança-se para a regressão linear múltipla, inclusão de variáveis categóricas e técnicas de seleção de variáveis e análise de diagnóstico. Tópicos adicionais como modelos lineares generalizados (MLGs), extensões não lineares e métodos de regularização também serão apresentados.

Mais do que aprender procedimentos técnicos, o objetivo é desenvolver a capacidade de interpretar resultados, avaliar a adequação dos modelos e comunicar conclusões de forma clara. A ênfase está tanto na teoria quanto na prática, de modo que o estudante seja capaz de aplicar a modelagem estatística em contextos multidisciplinares.

## 1.1 A centralidade da regressão

A análise de regressão ocupa posição central na Estatística e, em especial, na Econometria. Como afirma Hoffmann (2016):

> "A análise de regressão é o método mais importante da econometria."

Essa afirmação reflete o fato de que praticamente toda modelagem econométrica, seja para estimar elasticidades, avaliar políticas públicas, medir impactos ou testar teorias, passa, de alguma forma, por um modelo de regressão.

Mas essa centralidade não é exclusiva da economia. Na saúde, a regressão mede risco e associações; na engenharia, modela desempenho; nas ciências ambientais, estima impactos; nas ciências sociais, investiga relações estruturais; na ciência de dados, permanece como ferramenta interpretável diante de modelos mais complexos.

A regressão tornou-se, portanto, uma linguagem universal para responder a uma pergunta fundamental:

> Como varia uma quantidade quando outra varia?

Essa pergunta é simples. A resposta exige matemática, probabilidade, inferência e interpretação.

## 1.2 Objetivos do livro

- Contextualizar historicamente os modelos de regressão;
- Compreender a lógica da modelagem estatística: componente sistemático, componente aleatório e relação entre estes;
- Apresentar e discutir os principais modelos abordados na disciplina;
- Conectar a teoria com aplicações em economia, saúde, engenharia, ciências sociais e ambientais;
- Discutir potenciais e limitações de cada abordagem, reconhecendo as hipóteses subjacentes;
- Desenvolver a capacidade de usar ferramentas computacionais para ajuste, diagnóstico e interpretação de modelos.

Este material é concebidao como uma jornada pela família dos modelos de regressão. Partimos de um problema simples, como relacionar uma variável resposta a uma variável explicativa, e avançamos gradualmente até modelos capazes de lidar com múltiplos fatores, variáveis categóricas, dados de contagem, proporções e situações em que as hipóteses clássicas do modelo deixam de ser válidas.

A ideia central é que, ao final do semestre, o estudante seja capaz de compreender não apenas **como ajustar** um modelo, mas também **quando e por que** usá-lo, avaliando sua adequação e reconhecendo seus limites.

## 1.3 Breve História da Regressão

A regressão, como hoje conhecemos, é fruto de mais de um século de evolução teórica e prática. Seu ponto de partida remonta ao século XIX, quando estudos empíricos sobre hereditariedade começaram a revelar regularidades que poderiam ser descritas matematicamente. Desde então, a regressão deixou de ser apenas uma curiosidade biológica e tornou-se um dos pilares da estatística moderna, sustentando análises nas mais diversas áreas.

### 1.3.1 Francis Galton (1822–1911)

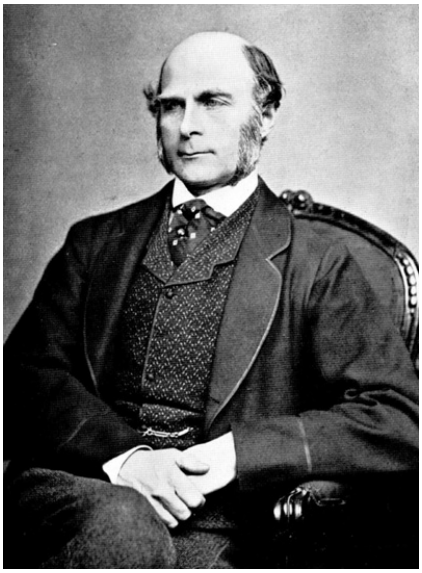


Figura 1.1: Francis Galton

A história da regressão começa com uma inquietação simples, quase doméstica. No final do século XIX, Francis Galton observava famílias inglesas e fazia uma pergunta que parecia trivial, mas que mudaria a estatística para sempre: **filhos de pais muito altos serão igualmente altos?**

Ao analisar dados de estaturas familiares, Galton percebeu algo intrigante: filhos de pais muito altos tendiam a ser altos, mas não tanto quanto seus pais; filhos de pais muito baixos tendiam a ser baixos, mas não tão baixos quanto seus progenitores. Havia uma força invisível puxando as medidas extremas de volta ao centro. Em 1886 (Galton (1886b); Galton (1886a)), ele chamou esse fenômeno de *regression toward mediocrity*.

Sem perceber completamente, Galton havia introduzido quatro pilares da regressão moderna: uma variável resposta, uma variável explicativa, uma média condicional e uma variabilidade em torno dessa média.

A matemática ainda era rudimentar. Mas a ideia estava lançada.

### 1.3.2 Karl Pearson (1857–1936)

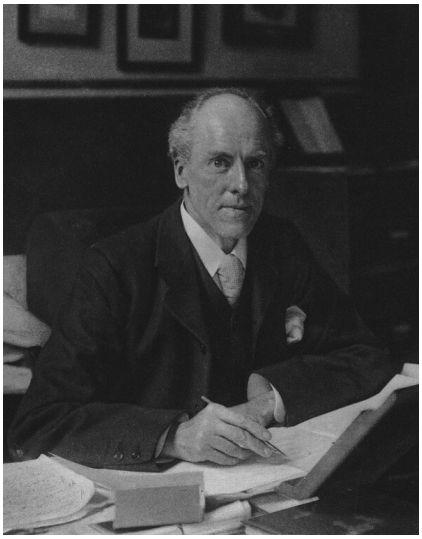


Figura 1.2: Karl Pearson

Se Galton teve a intuição, Karl Pearson deu forma matemática ao fenômeno. Discípulo e colaborador de Galton, Pearson transformou observações empíricas em estrutura formal. Desenvolveu o coeficiente de correlação linear, sistematizou métodos de ajuste e ajudou a fundar o primeiro departamento de estatística do mundo, no University College London. Criou também a revista *Biometrika*, marco na consolidação da estatística como disciplina científica.

A regressão deixava de ser apenas uma regularidade observada em dados biológicos. Tornava-se parte da teoria da probabilidade e da estatística matemática (Pearson e Lee (1903)).

O que antes era descrição começava a se tornar método.

### 1.3.3 Ronald A. Fisher (1890–1962)

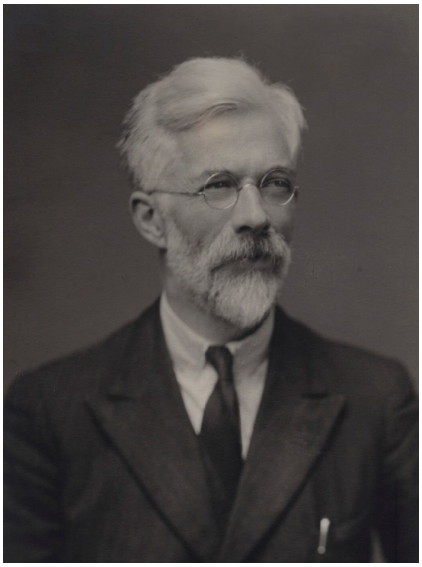


Figura 1.3: Ronald Fisher

Nas décadas seguintes, a estatística enfrentava um novo desafio: não bastava ajustar curvas; era preciso decidir. Ronald A. Fisher foi o arquiteto dessa virada. Na década de 1920, incorporou fundamentos de inferência à regressão e redefiniu o papel da estatística na ciência.

Com Fisher surgem a análise de variância (ANOVA), o método da máxima verossimilhança e os princípios modernos de planejamento experimental. A regressão passa a permitir testar hipóteses, construir intervalos de confiança e quantificar incertezas.

A técnica deixa de ser apenas descritiva. Torna-se ferramenta de decisão científica.

### 1.3.4 Jerzy Neyman (1894–1981) & Egon Pearson (1895–1980)

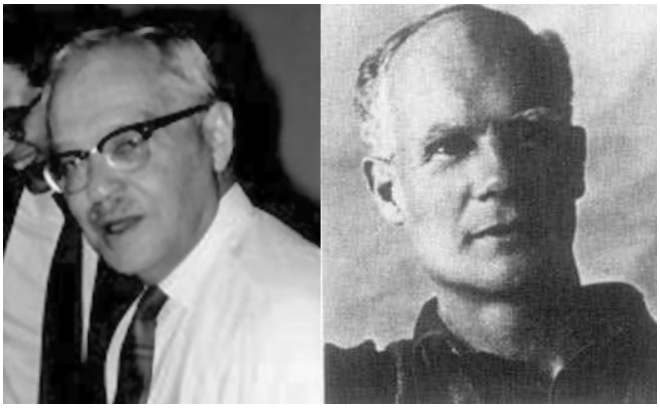


Figura 1.4: Jerzy Neyman & Egon Pearson

A década de 1930 marca outro salto conceitual. Jerzy Neyman e Egon Pearson, filho de Karl Pearson, estruturam formalmente os testes de hipóteses. Definem os erros do tipo I e tipo II, introduzem critérios de decisão e consolidam o conceito de intervalo de confiança.

A regressão, agora, não apenas estima relações: ela fornece regras claras para aceitar ou rejeitar hipóteses. A incerteza passa a ser quantificada de maneira sistemática.

O método ganha rigor lógico.

### 1.3.5 John Tukey (1915–2000)

Figura 1.5: John Tukey

Mas a estatística não evolui apenas por formalização. Na década de 1950, John Tukey propõe algo radical: antes de testar, é preciso explorar. Defende que os dados devem ser interrogados visualmente. Populariza gráficos de resíduos, diagnósticos e técnicas exploratórias.

A regressão passa a ser acompanhada por perguntas práticas:
os pressupostos fazem sentido? há pontos influentes? o modelo realmente representa os dados?

A estatística recupera o diálogo com a realidade empírica.

### 1.3.6 Peter McCullagh (1952–) & John Nelder (1934–2010)

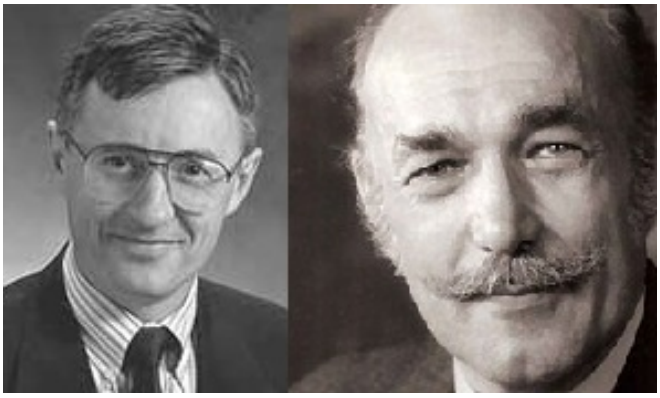


Figura 1.6: Peter McCullagh & John Nelder

Com o avanço da computação nas décadas de 1960 e 1970, a regressão amplia suas fronteiras. A álgebra matricial viabiliza modelos com múltiplos preditores, e a necessidade de analisar dados não normais torna-se evidente.

Em 1983, McCullagh e Nelder publicam *Generalized Linear Models*, obra que sistematiza os Modelos Lineares Generalizados. A regressão deixa de ser restrita a variáveis contínuas normalmente distribuídas. Passa a modelar contagens, proporções, tempos de ocorrência.

A ideia que nasceu da altura de pais e filhos expande-se para praticamente todos os campos científicos.

A regressão não surgiu pronta. Foi construída por inquietações, formalizações, rupturas e expansões. Cada geração acrescentou uma camada: intuição, estrutura, inferência, decisão, diagnóstico e generalização.

O que começou como uma pergunta sobre estatura tornou-se uma das ferramentas centrais da ciência moderna.

## 1.4 O contexto histórico: eugenia

É necessário reconhecer que parte do desenvolvimento inicial da estatística ocorreu em um contexto marcado por ideias eugenistas. Francis Galton cunhou o termo "eugenia" em 1883. A eugenia defendia o "melhoramento" da raça humana por meio de seleção artificial.

Karl Pearson foi defensor ativo dessas ideias e dirigiu o laboratório Francis Galton para a Eugenia Nacional na University College London. Ronald Fisher também expressou posições eugenistas em seus escritos.

As ideias eugenistas foram posteriormente utilizadas para justificar políticas discriminatórias e violações graves de direitos humanos no século XX. Após a Segunda Guerra Mundial, consolidou-se um consenso internacional de condenação à eugenia como ideologia racista e incompatível com princípios éticos fundamentais.

Essa contextualização histórica não diminui a importância dos avanços metodológicos produzidos por esses autores. Pelo contrário, reforça a necessidade de compreender que métodos estatísticos são ferramentas cujo uso exige responsabilidade ética.

A regressão é instrumento científico poderoso. O modo como é utilizada depende do pesquisador.

## 1.5 Consolidação da regressão como pilar da Estatística moderna

A trajetória da regressão pode ser compreendida como uma sequência evolutiva:

1. Observação empírica da regressão à média (Galton);
2. Formalização matemática e correlação (Pearson);
3. Inferência e planejamento experimental (Fisher);
4. Estrutura formal de testes de hipóteses (Neyman-Pearson);
5. Diagnóstico e análise exploratória (Tukey);
6. Generalização para modelos lineares generalizados (McCullagh e Nelder);
7. Regularização moderna (Ridge e LASSO);
8. Integração com ciência de dados e aprendizagem de máquina.

Apesar das transformações metodológicas, a pergunta central permanece:

> Qual é a relação média entre uma variável resposta e seus preditores?

## 1.6 Panorama dos Modelos de Regressão

Tendo em mente os objetivos, estruturas e suposições discutidos anteriormente, é possível visualizar o **panorama dos modelos de regressão** que compõem esta disciplina. A ideia é seguir uma trajetória progressiva: partir de modelos simples e intuitivos, que permitem enxergar diretamente a relação entre duas variáveis, e avançar gradualmente para estruturas

mais sofisticadas, capazes de lidar com múltiplos fatores, respostas não normais e bases de dados complexas.

O percurso do curso está organizado em quatro blocos principais:

1. **Fundamentos** — regressão linear simples (MRLS), que introduz a lógica de média condicional, estimação por mínimos quadrados e análise de resíduos.

2. **Modelos com múltiplos fatores** — regressão linear múltipla (MRLM), inclusão de variáveis categóricas e interpretação de efeitos parciais.

3. **Validação e escolha de modelos** — métodos de diagnóstico, análise de resíduos, critérios de seleção de variáveis e regularização.

4. **Extensões** — modelos lineares generalizados (GLMs), regressão polinomial, modelos não lineares e métodos modernos de regularização (Ridge e LASSO), fundamentais na era dos grandes volumes de dados.

Em cada etapa, o modelo será apresentado de forma integrada: sua **fundamentação teórica**, as **suposições** que o tornam válido, exemplos de **interpretação prática** e **aplicações reais** em áreas como economia, saúde, engenharia, ciências ambientais e ciência de dados.

Assim, o panorama da disciplina funciona como um **mapa de navegação**: consolidamos primeiro os alicerces da regressão, depois exploramos aplicações mais complexas e, por fim, avançamos até modelos modernos que dialogam diretamente com os desafios atuais da análise de dados.

# 2 Modelagem Estatística e Regressão

## 2.1 Introdução à Modelagem

Antes de estudarmos a regressão em si, é importante entender que ela faz parte de um campo mais amplo chamado **modelagem matemática/estatística**. Modelar é o ato de construir representações formais de fenômenos reais com o objetivo de descrevê-los, explicá-los, prever seu comportamento ou orientar decisões.

Em termos conceituais, modelar significa responder à seguinte pergunta:

> Como posso representar, de maneira estruturada, um fenômeno complexo do mundo real por meio de variáveis e relações formais?

De forma geral, a modelagem matemática/estatística consiste na tentativa de traduzir um problema do mundo real em termos matemáticos ou estatísticos, estruturando hipóteses, equações e relações que permitam analisar e responder à pergunta proposta. Ela está presente em diversas áreas, como física, química, biologia, economia, engenharia e ciências sociais.

É fundamental compreender que **um modelo não é a realidade**. Ele é uma aproximação útil. Todo modelo envolve simplificações, suposições e escolhas: quais variáveis considerar, quais ignorar, que tipo de relação assumir, qual grau de precisão é necessário. Assim, modelar é sempre um exercício de equilíbrio entre realismo e simplicidade.

Alguns exemplos de fenômenos que podem ser descritos por modelos matemáticos incluem:
- Crescimento populacional;
- Reações químicas;
- Sistemas mecânicos e eletrônicos;
- Previsão do clima;
- Dinâmica do tráfego e da logística;
- Estratégias econômicas e financeiras;
- Mudanças ambientais e climáticas.

### 2.1.1 Modelo matemático versus modelo estatístico

Uma distinção conceitual importante é a diferença entre:

- **Modelo matemático (determinístico)**: descreve uma relação funcional exata entre variáveis. Para cada valor de entrada, existe um único valor de saída.

- **Modelo estatístico (estocástico)**: descreve uma relação média ou probabilística, reconhecendo a presença de variabilidade não explicada pelas variáveis observadas.

No modelo determinístico, escreve-se algo como:

$$Y = 2X$$

Aqui, se $X = 5$, então necessariamente $Y = 10$. Não há variação possível.

Já em um modelo estatístico, reconhecemos que fenômenos reais sofrem influência de fatores não observados, erros de mensuração, flutuações naturais ou variáveis omitidas. Assim, escreve-se:

$$Y = 2X + \varepsilon,$$

em que $\varepsilon$ representa a componente aleatória (ruído).

A diferença entre essas duas estruturas pode ser visualizada na simulação abaixo.

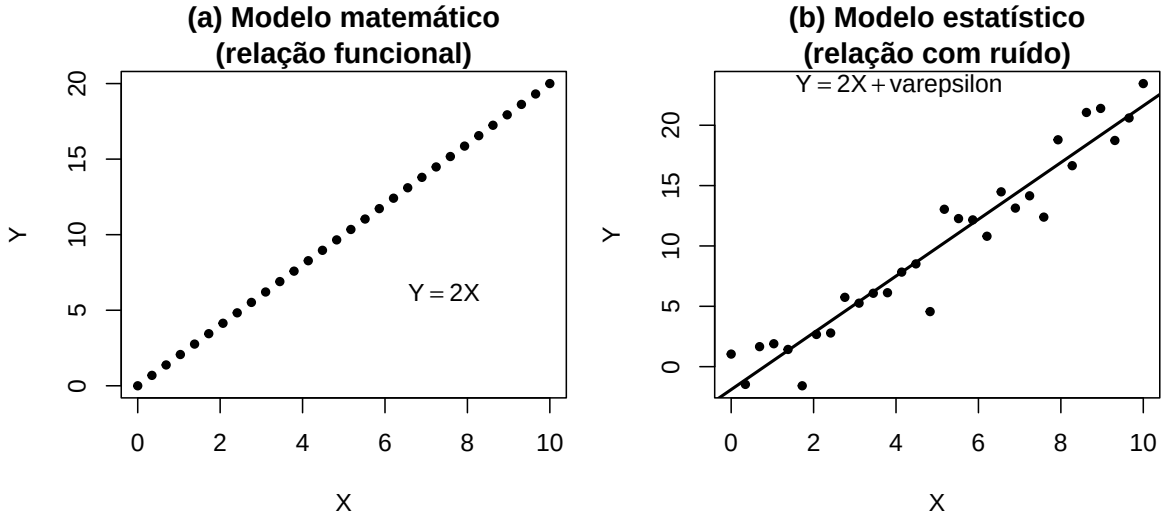



Figura 2.1: Modelo matemático (determinístico) versus modelo estatístico (com ruído): no primeiro, a relação é funcional; no segundo, observa-se uma tendência média com variabilidade aleatória.

Observe que, no modelo estatístico, não buscamos uma igualdade exata, mas uma **tendência média** em torno da qual os dados se distribuem.

Essa diferença é central para compreender a regressão.

### 2.1.2 Classificações didáticas de modelos

De forma didática, modelos matemáticos/estatísticos podem ser classificados em:

1. **Determinísticos ou estocásticos (estatísticos)**
   Dependendo se o acaso está explicitamente presente na formulação.
2. **Discretos ou contínuos**
   Dependendo da natureza das variáveis envolvidas. Contagens são discretas; altura, peso e temperatura são contínuos.
3. **Dinâmicos ou estáticos**
   Dependendo se o tempo é incorporado explicitamente na estrutura do modelo.

Essas classificações são discutidas em livros clássicos de modelagem matemática, como Meerschaert (2013) e Giordano, Fox, e Horton (2013). Elas ajudam a organizar o tipo de pergunta que estamos fazendo e a estrutura matemática adequada para respondê-la.

### 2.1.3 O ciclo da modelagem

O processo de modelagem geralmente envolve as seguintes etapas:

1. Formular a pergunta em termos matemáticos.
2. Selecionar uma abordagem de modelagem.
3. Construir o modelo com base nas variáveis e hipóteses do problema.
4. Resolver o modelo matemático (ou ajustá-lo aos dados, no caso estatístico).
5. Interpretar a solução em termos do fenômeno real.

Esse ciclo é **iterativo**: muitas vezes o modelo precisa ser ajustado ou refinado conforme novas informações surgem. Ao confrontar o modelo com dados, podem surgir questões como:

- A forma funcional escolhida faz sentido?
- Variáveis importantes foram omitidas?
- O comportamento dos resíduos está de acordo com as hipóteses?
- O modelo mantém desempenho em novos dados?

Modelagem não é um procedimento linear; é um processo de aproximações sucessivas.

Além disso, três elementos costumam acompanhar o ciclo de modelagem:

- **Estimação ou calibração**: ajuste de parâmetros com base em dados observados.

- **Validação**: verificação da qualidade do ajuste.
- **Análise de sensibilidade**: avaliação do impacto de pequenas mudanças nas hipóteses ou nos dados.

### 2.1.4 Exemplos de modelagem

- **Problema de otimização**: determinar o momento ideal de venda de um animal de criação, considerando ganho de peso, custo de manutenção e preço de mercado.

- **Problema de crescimento populacional**: prever a evolução da população de um país ou espécie a partir de dados censitários.

Dentro desse panorama mais amplo, a regressão ocupa um papel específico: ela é uma técnica estatística voltada para modelar a **média condicional** de uma variável resposta em função de variáveis explicativas. Em outras palavras, ela formaliza matematicamente a ideia de que parte da variação observada pode ser explicada sistematicamente, enquanto outra parte permanece como ruído.

Modelagem é, portanto, a linguagem que conecta teoria, dados e inferência. A regressão é uma de suas expressões mais importantes.

#### 2.1.4.1 Exemplos de modelagem em diferentes áreas

A modelagem matemática e estatística é utilizada para descrever e prever o comportamento de fenômenos em uma grande variedade de contextos. Alguns exemplos ilustrativos incluem:

- **Epidemiologia**: modelos SIR (Susceptíveis–Infectados–Recuperados) para descrever a propagação de doenças infecciosas ao longo do tempo.
- **Engenharia de Pesca**: relação entre esforço de pesca (dias de mar, número de embarcações) e o estoque pesqueiro disponível.
- **Ciência de Dados**: previsão da demanda de energia elétrica a partir de temperatura, hora do dia e perfil de consumo.
- **Economia**: modelos de oferta e demanda que relacionam preços e quantidades em equilíbrio de mercado.
- **Climatologia**: simulações que conectam emissão de gases de efeito estufa, temperatura média global e regimes de precipitação.
- **Educação**: análise da relação entre investimento em ensino e desempenho em exames padronizados.
- **Oceanografia**: modelos que descrevem correntes marinhas em função de gradientes de temperatura e salinidade.

Esses exemplos mostram que **modelagem é uma linguagem comum em ciência e tecnologia**. A regressão estatística é uma forma particular de modelagem que foca em quantificar relações entre variáveis observáveis, buscando separar o **sinal** (estrutura determinística) do **ruído** (componente aleatória).

## 2.2 Introduçao à Regressão

Regressão é um **modelo matemático-estatístico** que busca relacionar uma **variável resposta** ($Y$) com uma ou mais **variáveis explicativas**. De forma conceitual, a regressão parte da ideia de que o comportamento médio de $Y$ pode ser descrito condicionalmente às variáveis explicativas $X$.

Em termos mais precisos, o que a regressão modela não é simplesmente $Y$, mas a **média condicional**:

$$E(Y \mid X)$$

Essa perspectiva é central nos textos clássicos de regressão e econometria (Charnet et al. (2008); Hoffmann (2006); Gujarati (2006)). O objetivo não é afirmar que $X$ determina exatamente $Y$, mas que existe uma **estrutura sistemática média** associada às variáveis explicativas.

A ideia fundamental é que a variação observada em um fenômeno pode ser decomposta em duas partes:

1. Uma **estrutura determinística** (ou componente sistemática), que representa o sinal da relação entre as variáveis;
2. Uma **componente aleatória**, representada pelo ruído $\varepsilon$, que captura variações não explicadas pelo modelo.

Essa decomposição aparece de forma recorrente na literatura de regressão aplicada (Draper e Smith (1998); Kutner et al. (2005); Montgomery, Peck, e Vining (2021)) como a base conceitual do modelo linear.

De forma geral, existem duas representações usuais para essa ideia. - **Modelo Aditivo** (o mais utilizado):

$$Y = \mu(X) + \varepsilon$$

- **Modelo Multiplicativo** (útil em contextos onde a variabilidade é proporcional ao nível médio):

$$Y = \mu(X) \cdot \varepsilon$$

com:

- $mu(X) \rightarrow$ parte determinística (estrutura média);
- $\varepsilon \rightarrow$ componente aleatória (ruído), que pode ser aditivo ou multiplicativo.

O sinal $\mu(X)$ também é denominado de função de regressão.

No modelo aditivo clássico, assume-se que

$$E(\varepsilon \mid X) = 0,$$

isto é, o erro não carrega informação sistemática adicional sobre $Y$ além da já contida em $X$. Essa condição garante que $E(Y \mid X) = \mu(X)$, permitindo interpretar $\mu(X)$ como a **média condicional de $Y$ dado $X$**. Essa hipótese é central para a interpretação dos coeficientes como efeitos médios condicionais (Hoffmann (2006); Gujarati (2006)).

No caso do **modelo multiplicativo**, a condição análoga é

$$E(\varepsilon \mid X) = 1.$$

Nesse caso, temos $E(Y \mid X) = \mu(X)\, E(\varepsilon \mid X) = \mu(X)$, de modo que $\mu(X)$ continua representando a média condicional de $Y$ dado $X$.

A diferença fundamental entre as duas formulações é estrutural. No modelo aditivo, o erro representa um **deslocamento absoluto** em torno da média condicional: a variação aleatória soma-se ao valor esperado, produzindo oscilações cuja magnitude, em princípio, não depende do nível médio. Já no modelo multiplicativo, o erro atua como um **fator proporcional**, ampliando ou reduzindo o valor médio de acordo com sua própria intensidade. Nesse caso, a variabilidade está intrinsecamente ligada ao nível esperado da variável resposta.

Essa distinção tem implicações relevantes. Em estruturas aditivas, concebe-se a variabilidade como relativamente independente da escala do fenômeno; em estruturas multiplicativas, a dispersão tende a crescer ou decrescer proporcionalmente ao valor médio. Fenômenos econômicos, biológicos e ambientais frequentemente apresentam esse comportamento proporcional, por exemplo, renda, produção, crescimento populacional ou biomassa, o que torna a formulação multiplicativa conceitualmente mais adequada em muitos contextos.

A escolha entre uma estrutura aditiva ou multiplicativa envolve considerações teóricas e empíricas. Modelos econômicos frequentemente sugerem relações proporcionais, enquanto a inspeção gráfica dos dados pode revelar padrões de variância crescente. Assim, a definição da forma funcional e da natureza da variação depende da teoria utilizada, da escala de mensuração

das variáveis e do comportamento empírico observado nos dados (Gujarati (2006); Hoffmann (2006)).

No caso do modelo aditivo clássico, a hipótese de que o erro não carrega informação sistemática adicional além da contida nas variáveis explicativas é central para a interpretação dos coeficientes como efeitos médios condicionais. Essa perspectiva, segundo a qual a regressão modela a média condicional e não valores individuais, constitui um dos pilares da regressão aplicada e da econometria tradicional (Hoffmann (2006); Gujarati (2006)).

### 2.2.1 Interpretação conceitual: sinal e ruído

A regressão parte do reconhecimento de que fenômenos reais são influenciados por múltiplos fatores, muitos dos quais não são observáveis ou mensuráveis. Assim, mesmo que exista uma relação estrutural entre $X$ e $Y$, os dados observados não estarão perfeitamente alinhados sobre uma curva.

O modelo assume que:

$$Y = \text{sinal} + \text{ruído}$$

O sinal representa a parte explicável pelas variáveis incluídas no modelo; o ruído representa:

- variáveis omitidas,
- erros de mensuração,
- flutuações naturais,
- fatores aleatórios não controlados.

Essa separação entre componente sistemática e erro é um dos pilares da regressão linear clássica (Draper e Smith (1998)).

#### 2.2.1.1 Exemplo ilustrativo: Sinal + Ruído

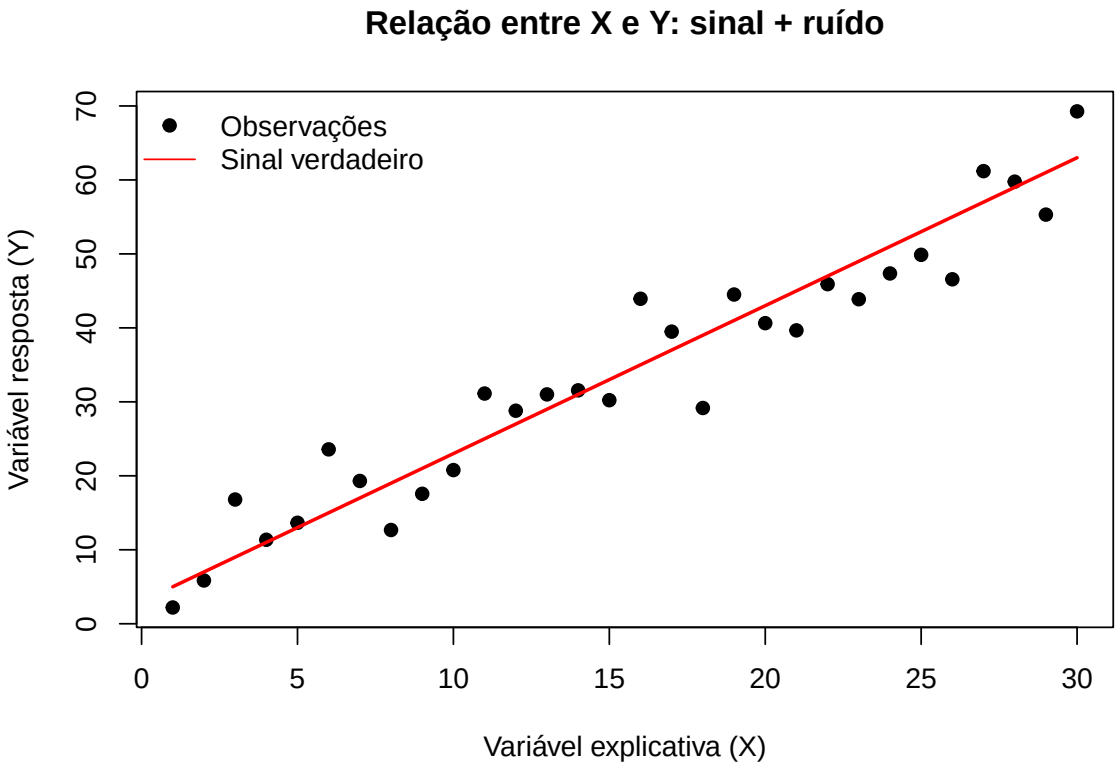
A seguir, simulamos um conjunto de dados artificiais para ilustrar a lógica central da regressão: **um sinal determinístico mais um componente de ruído**.

Considere a relação:

$$Y = 3 + 2X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 5^2),$$

com $X \in \{1, 2, \ldots, 30\}$. A reta $Y = 3 + 2X$ representa o **sinal verdadeiro**, enquanto os pontos observados $(X, Y)$ incluem a variabilidade aleatória do ruído $\varepsilon$.

O gráfico abaixo mostra a dispersão dos dados simulados juntamente com a curva verdadeira. Essa visualização reforça a ideia de que a regressão busca recuperar a estrutura determinística do fenômeno em meio à aleatoriedade introduzida pelos ruídos (erros ou fonte de variação).




#### 2.2.1.2 Exemplo ilustrativo: Sinal + Ruído (homoscedástico vs heteroscedástico)

Neste exemplo, comparamos dois cenários de regressão linear que diferem apenas na **dispersão dos ruídos**:

- **Homoscedástico**: a variância dos erros é constante em todos os valores de $X$. Nesse caso, a nuvem de pontos se distribui de forma aproximadamente uniforme em torno da reta verdadeira, independentemente da posição no eixo $X$.

- **Heteroscedástico**: a variância cresce com $X$. Assim, para valores pequenos de $X$ os pontos estão mais concentrados, enquanto para valores grandes de $X$ a dispersão aumenta.

Esse contraste é fundamental: se a homoscedasticidade não for respeitada, os estimadores de mínimos quadrados continuam **não viesados**, mas deixam de ser eficientes (ou seja, não são os

de menor variância). Isso motiva o uso de métodos alternativos, como os mínimos quadrados ponderados (WLS) ou transformações nos dados.

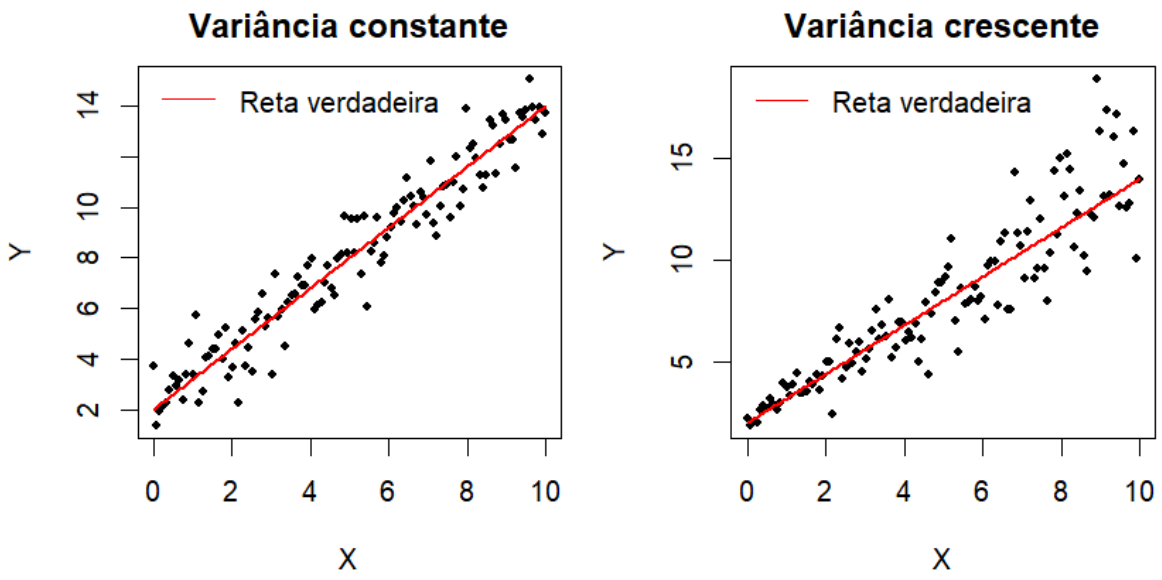



Figura 2.2: Sinal + ruído: (esq.) variância constante; (dir.) variância crescente (heteroscedasticidade).

#### 2.2.1.3 Exemplo ilustrativo: Possíbildiades de relações ente X e Y

Observe que nem toda relação entre duas variáveis é **linear e forte**. Considere algums possibilidades comuns:

- **(a) Sem correlação**: não existe padrão claro entre $X$ e $Y$; conhecer $X$ não ajuda a prever $Y$.
- **(b) Correlação linear fraca**: existe uma tendência positiva, mas com grande dispersão ao redor da reta.
- **(c) Correlação linear forte**: os pontos seguem de perto uma tendência linear; $X$ *explica* grande parte da variabilidade em $Y$.
- **(d) Relação não linear**: $X$ e $Y$ se relacionam, mas a forma não é bem descrita por uma reta (por exemplo, uma parábola).

Esse exemplo mostra que a **correlação linear** é apenas um caso particular dentro de uma variedade de possíveis dependências entre variáveis. Por isso, ao analisar dados, é sempre

importante, quando possível, **visualizar os diagramas de dispersão** antes de ajustar um modelo, evitando conclusões equivocadas sobre linearidade.

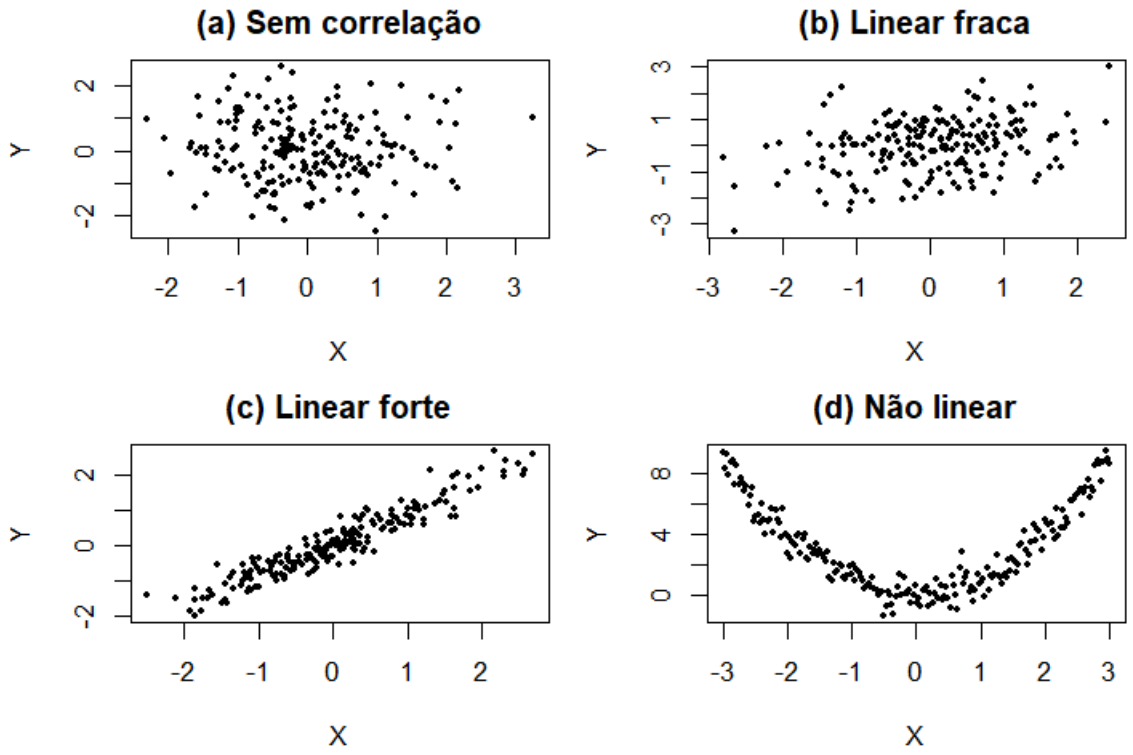



Figura 2.3: Correlação entre X e Y: (a) nenhuma, (b) linear fraca, (c) linear forte, (d) não linear.

#### 2.2.1.4 Exemplo ilustrativo: Relação não linear entre $X$ e $Y$

Até agora, discutimos a regressão sob a perspectiva de relações lineares. No entanto, nem todo fenômeno real apresenta comportamento aproximadamente linear. Em muitos contextos, como crescimento biológico, processos de absorção, liberação acumulada de substâncias ou resposta a doses, a relação entre as variáveis pode apresentar **curvatura**, **saturação** ou **pontos de inflexão**.

Neste exemplo, construiremos artificialmente uma base de dados cujo **sinal verdadeiro não é linear**. O comportamento adotado será do tipo **crescimento com saturação**: inicialmente, pequenos aumentos em $X$ produzem grandes aumentos em $Y$; à medida que $X$ cresce, o efeito marginal diminui e a curva tende a estabilizar.

Essa estrutura pode ser representada genericamente por uma função do tipo:

$$Y = f(X) + \varepsilon$$

em que $f(X)$ é não linear e $\varepsilon$ representa o componente aleatório.

O objetivo aqui é didático: comparar dois cenários sobre os mesmos dados observados:

1. Ajustar um **modelo linear**, mesmo sabendo que o sinal não é linear;
2. Comparar esse ajuste com a **curva verdadeira** que gerou os dados.

Essa comparação permite visualizar um ponto conceitual central da modelagem:

> Um modelo pode estar corretamente estimado sob suas hipóteses e, ainda assim, estar incorretamente especificado.

Ou seja, o problema pode não estar na estimação, mas na **forma funcional escolhida**.

Ao observar os gráficos, procure refletir:

- O modelo linear captura adequadamente o padrão médio?
- Há evidências visuais de curvatura?
- O erro parece sistemático ao longo de $X$?
- O que aconteceria com os resíduos nesse caso?

Esse tipo de análise visual é uma etapa importante antes da formalização do modelo. A regressão linear é uma ferramenta poderosa, mas sua adequação depende da coerência entre a estrutura assumida e o comportamento real dos dados.

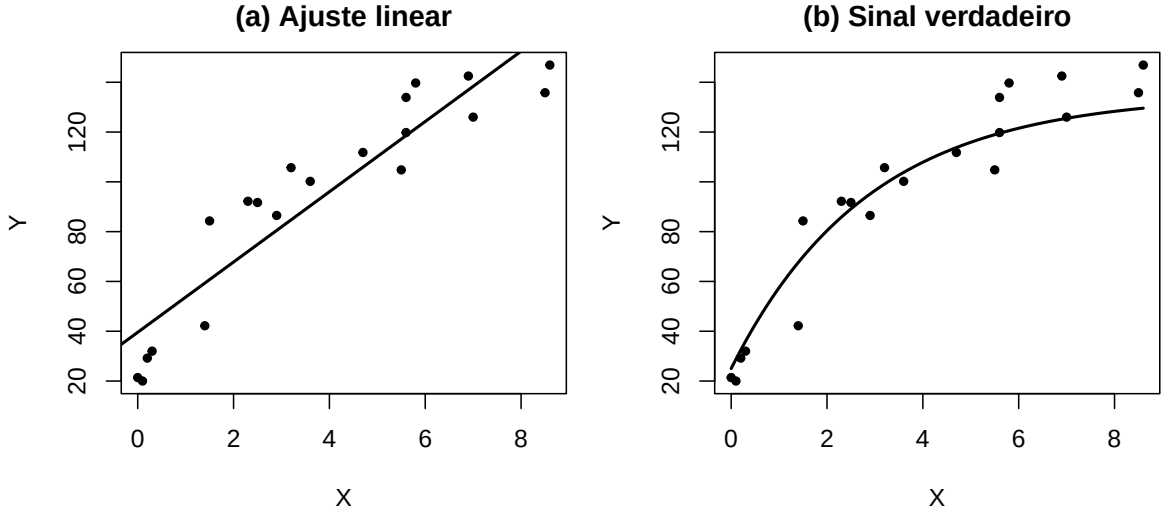



Figura 2.4: Exemplo com sinal não linear: (esq.) ajuste linear; (dir.) sinal verdadeiro que gerou os dados.

### 2.2.2 Objetivos e Estruturas do Modelo da Regressão

Depois de explorar exemplos visuais de associação entre variáveis, é importante consolidar uma visão mais conceitual sobre a regressão. A regressão não é simplesmente uma técnica de ajuste de curvas; ela é uma estrutura formal para modelar a **relação média entre variáveis observáveis**.

Em termos conceituais, a regressão procura compreender como o comportamento médio de $Y$ varia em função de $X$. Isto é, ela organiza a seguinte pergunta:

> Como a variável resposta se comporta, em média, quando as variáveis explicativas variam?

#### 2.2.2.1 Objetivos centrais

A regressão pode ser compreendida a partir de três grandes finalidades:

1. **Explicar**
   Identificar quais fatores estão associados à variável resposta e quantificar a magnitude dessas associações.
2. **Predizer**
   Utilizar o modelo ajustado para estimar valores futuros ou não observados de $Y$.
3. **Controlar**
   Avaliar o efeito de uma variável mantendo as demais constantes, permitindo interpretações condicionais.

Esses objetivos aparecem de maneira combinada na prática. Em econometria, frequentemente busca-se compreender relações estruturais entre variáveis macroeconômicas (Gujarati (2006); Hoffmann (2006)). Em engenharia e estatística aplicada, pode haver maior ênfase na capacidade preditiva do modelo (Montgomery, Peck, e Vining (2021)). Em estudos científicos em geral, explicar e prever caminham juntos.

#### 2.2.2.2 Observações conceituais importantes

Antes de avançarmos para a formalização do modelo linear simples, duas observações merecem destaque.

**1. Correlação significativa e ajuste linear**

Em geral, quando existe correlação linear significativa entre variável explicativa e variável resposta, é razoável esperar que o Modelo de Regressão Linear Simples produza um ajuste satisfatório. Isso ocorre porque a correlação mede o grau de associação linear entre duas

variáveis. Se essa associação é forte, a reta ajustada tende a capturar uma parcela substancial da variabilidade observada.

Entretanto, correlação elevada não garante que o modelo esteja corretamente especificado. Pode haver curvatura, variância não constante ou outros padrões estruturais não capturados por uma reta. A inspeção gráfica continua sendo essencial.

**2. Associação estatística não implica causa e efeito**

Uma relação estatística entre variáveis não implica automaticamente causalidade.Mesmo que um coeficiente estimado seja estatisticamente significativo, isso não significa que uma variável cause a outra. A interpretação causal:

- não pode se basear apenas na amostra considerada;
- deve estar apoiada em teoria ou conhecimento empírico;
- pode exigir desenho experimental ou hipóteses estruturais adicionais.

Exemplos clássicos ilustram essa distinção:

- **Despesas de consumo pessoal e renda pessoal disponível** — aqui, a teoria econômica sustenta a direção causal.
- **Ganho de peso e consumo de calorias** — a evidência empírica e experimental apoia a interpretação causal.

Sem essa base teórica ou experimental, a regressão revela associação, não mecanismo causal.

#### 2.2.2.3 Duas características fundamentais do modelo de regressão

Do ponto de vista probabilístico, um modelo de regressão possui duas características essenciais:

1. Para cada nível fixado de $X$, existe uma **distribuição de probabilidade de** $Y$.
2. As médias dessas distribuições variam de forma sistemática com $X$.

Essa segunda característica é o coração da regressão: modelar como a média de $Y$ se altera quando $X$ varia.

Visualmente, isso significa que, para cada valor de $X$, não há um único valor possível de $Y$, mas sim uma distribuição de valores possíveis. O modelo descreve o comportamento médio dessas distribuições.

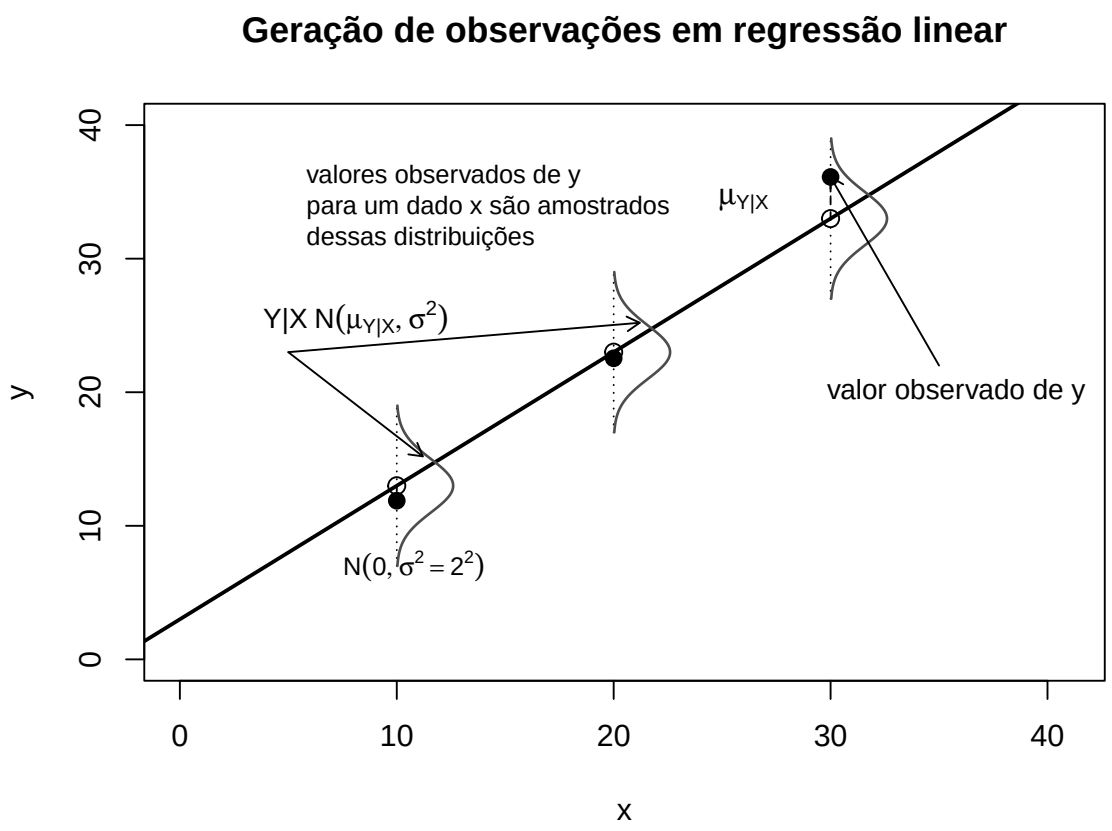



Figura 2.5: Como as observações são geradas em regressão linear.

Essa perspectiva probabilística será formalizada no próximo capítulo, quando introduzirmos o Modelo de Regressão Linear Simples (MRLS).

#### 2.2.2.4 Estrutura conceitual da regressão

Independentemente da área, a regressão parte de uma ideia fundamental:

$$Y = \text{parte sistemática} + \text{parte não explicada}$$

A parte sistemática representa o padrão médio associado às variáveis explicativas. A parte não explicada representa variações adicionais que não são capturadas pelo modelo, seja por fatores não observados, limitações de mensuração ou variabilidade inerente ao fenômeno.

A regressão organiza essa decomposição de maneira formal e mensurável. Ela não elimina a variabilidade; ela a estrutura.

#### 2.2.2.5 Alguns exemplos que mostram a importância prática

A utilidade da regressão torna-se mais clara quando observamos aplicações concretas.

- **Economia**
  Modelos de regressão são usados para analisar como variáveis como taxa de juros, câmbio e nível de consumo se associam à inflação. O objetivo pode ser compreender o mecanismo econômico (explicação) ou projetar cenários futuros (previsão).
- **Saúde**
  Em estudos clínicos, a regressão permite avaliar a relação entre tratamento e resposta terapêutica, controlando por idade, sexo ou comorbidades. Aqui, a regressão organiza a comparação entre grupos e ajuda a quantificar diferenças médias.
- **Engenharia**
  Na modelagem da resistência de materiais, regressões relacionam tensão aplicada e deformação observada, permitindo prever limites operacionais.
- **Esportes**
  Pode-se modelar o desempenho de uma equipe em função de variáveis como investimento, tempo de posse de bola ou eficiência ofensiva, identificando padrões associados ao resultado final.
- **Pesca e aquicultura**
  Relações entre esforço de pesca e biomassa capturada, ou entre tempo de cultivo e ganho de peso, podem ser analisadas por regressão para apoiar decisões produtivas.
- **Políticas públicas**
  Avaliações de impacto utilizam regressão para investigar como programas sociais se associam a indicadores como renda, escolaridade ou emprego.

Esses exemplos mostram que regressão não é apenas um instrumento matemático; é uma ferramenta de organização do raciocínio quantitativo em contextos reais.

#### 2.2.2.6 Potenciais e limites

A regressão possui algumas virtudes que explicam sua ampla utilização:

- Permite quantificar efeitos médios;
- Oferece interpretação relativamente direta dos coeficientes;
- Estrutura a análise de dados de forma sistemática;
- Serve como base para extensões mais sofisticadas.

Ao mesmo tempo, é importante reconhecer que:

- A qualidade da conclusão depende da qualidade dos dados;
- A forma funcional escolhida influencia os resultados;
- Modelos podem ser mal especificados;
- Associação estatística não é automaticamente causalidade.

Reconhecer essas dimensões faz parte da maturidade estatística.

A regressão é, portanto, uma ferramenta que conecta **teoria, dados e decisão**. Ela organiza a variabilidade observada em uma estrutura interpretável e mensurável.

No próximo capítulo, iniciaremos o estudo formal do **Modelo de Regressão Linear Simples (MRLS)**, que constitui o ponto de partida para compreender, com rigor matemático, como essa estrutura é estimada e quais propriedades possui.

# 3 Exercícios e atividades

## 3.1 Exercícios conceituais

1. Ao transformar um problema do mundo real em um modelo matemático ou estatístico, é necessário definir variáveis e relações formais.

   a) O que significa “traduzir um problema real” em linguagem matemática?

   b) Como a escolha das variáveis influencia o tipo de resposta que o modelo pode oferecer?

   c) Dê um exemplo em que a escolha inadequada das variáveis leve a conclusões limitadas ou equivocadas.

2. Considere um problema de crescimento populacional ao longo do tempo.

   a) Explique a diferença conceitual entre um modelo estático e um modelo dinâmico nesse contexto.

   b) Por que a inclusão explícita do tempo altera a estrutura matemática do modelo?

3. Em muitos fenômenos reais, diferentes modelos podem ser propostos para descrever o mesmo problema.

   a) Quais critérios podem ser utilizados para escolher entre dois modelos concorrentes?

   b) Explique como o critério da parcimônia atua nesse processo de escolha.

   c) Por que modelos excessivamente complexos podem ser problemáticos, mesmo quando ajustam melhor os dados?

4. Explique a afirmação: “um modelo não é a realidade; é uma aproximação útil”.

   a) O que significa equilíbrio entre realismo e simplicidade?

   b) Por que todo modelo envolve escolhas e simplificações?

c) Dê um exemplo em que um modelo necessariamente ignora parte da complexidade do fenômeno.

5. Considere as duas estruturas apresentadas no capítulo:

$$Y = 2X$$

e

$$Y = 2X + \varepsilon.$$

a) Diferencie conceitualmente modelo determinístico e modelo estatístico.

b) Explique o papel de $\varepsilon$ na segunda equação.

c) Liste três possíveis fontes para $\varepsilon$.

d) Por que, no modelo estatístico, buscamos uma tendência média e não uma igualdade exata?

6. O capítulo afirma que a regressão modela a média condicional:

$$E(Y \mid X).$$

a) Explique o que significa média condicional em linguagem intuitiva.

b) Mostre como a ideia de $E(Y \mid X)$ está relacionada à decomposição

$$Y = \text{sinal} + \text{ruído}.$$

c) É possível que um valor observado de $Y$ esteja distante de $E(Y \mid X)$ e, ainda assim, o modelo esteja adequado? Justifique.

7. O texto apresenta duas formas gerais de modelagem:

$$Y = f(X) + \varepsilon$$

e

$$Y = f(X) \cdot \varepsilon.$$

a) Diferencie erro aditivo e erro multiplicativo.

b) Em que tipo de fenômeno o erro tende a ser proporcional ao nível médio?

8. Sobre o ciclo da modelagem:

a) Explique por que o processo de modelagem é iterativo.

b) Diferencie estimação, validação e análise de sensibilidade.

c) Dê um exemplo de situação em que, após ajustar o modelo, seria necessário voltar e modificar hipóteses ou forma funcional.

9. O capítulo mostra que nem toda relação entre $X$ e $Y$ é linear.

a) Por que correlação linear elevada não garante especificação correta do modelo?

b) Dê um exemplo conceitual de relação não linear em que o coeficiente de correlação de Pearson possa ser próximo de zero.

c) O que significa dizer que há "erro sistemático ao longo de $X$"?

10. Sobre associação e causalidade:

a) Explique a diferença entre associação estatística e causalidade.

b) Liste duas razões pelas quais um coeficiente estimado pode ser estatisticamente significativo e ainda não representar um efeito causal.

c) Que tipo de informação adicional (teórica ou experimental) seria necessária para sustentar uma interpretação causal?

11. Sobre previsão e extrapolação:

a) Diferencie interpolação e extrapolação.

b) Por que extrapolar pode ser arriscado mesmo quando o ajuste parece bom no intervalo observado?

c) Dê um exemplo aplicado em que extrapolação seria particularmente problemática.

12. Escolha um dos contextos aplicados citados no capítulo.

a) Defina uma variável resposta $Y$ e pelo menos três variáveis explicativas $X_1, X_2, X_3$.

b) Indique se o modelo conceitual seria mais plausivelmente aditivo ou multiplicativo e justifique.

c) Liste duas suposições que você consideraria críticas para interpretar os resultados.

13. Reflita sobre a seguinte estrutura geral:

$$Y = f(X) + \varepsilon.$$

a) O que significa assumir que $E(\varepsilon \mid X) = 0$?

b) Por que essa hipótese é central para interpretar os coeficientes como efeitos médios condicionais?

c) O que pode acontecer se essa hipótese não for satisfeita?

As respostas devem ser redigidas de forma argumentativa, conectando explicitamente os conceitos apresentados no capítulo.

## 3.2 Atividade de Simulação e Regressão

Esta atividade tem como objetivo consolidar os conceitos estudados ao longo do capítulo 2 por meio de **simulações controladas**. A proposta é investigar, de forma sistemática, como diferentes estruturas funcionais (linear, quadrática, exponencial e potência) se comportam sob ruído e como o ajuste por mínimos quadrados responde a essas situações.

Em todos os exercícios:

- Gere os dados conforme indicado no código.
- Produza os gráficos solicitados.
- Ajuste os modelos especificados.
- Responda às questões de forma **argumentativa**, conectando os resultados aos conceitos de sinal, ruído, forma funcional e especificação do modelo.

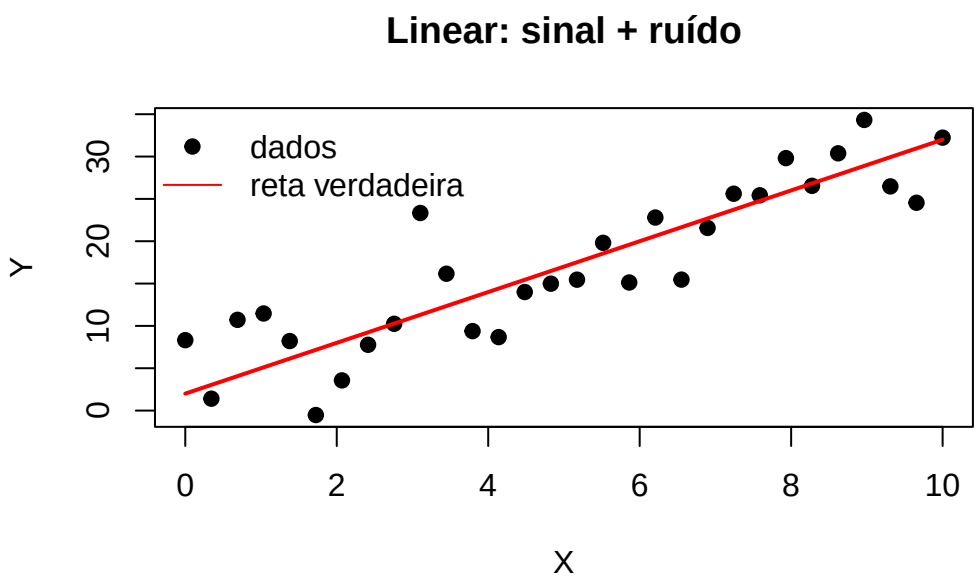
**Objetivo:** simular dados sob diferentes modelos, visualizar dispersões, calcular correlações e ajustar modelos via **MQO** (OLS), comparando a **curva verdadeira** que gerou os dados com o **ajuste estimado**.

1. Preparação do Ambiente

```
# Preparação do ambiente (simples e reprodutível)
set.seed(0)
```

2. Simular dados lineares simples

```r
# 2) Simular dados lineares simples
n <- 30
x <- seq(0, 10, length.out = n)
beta0 <- 2
beta1 <- 3
sigma <- 5

sinal <- beta0 + beta1*x
y <- sinal + rnorm(n, mean = 0, sd = sigma)

# Visualização: pontos + reta verdadeira
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Linear: sinal + ruído")
lines(x, sinal, col = "red", lwd = 2)
legend("topleft", legend = c("dados", "reta verdadeira"),
       pch = c(19, NA), lty = c(NA, 1), col = c("black", "red"), bty = "n")
```




```r
# Correlação
cor_xy <- cor(x, y)
cor_xy
```

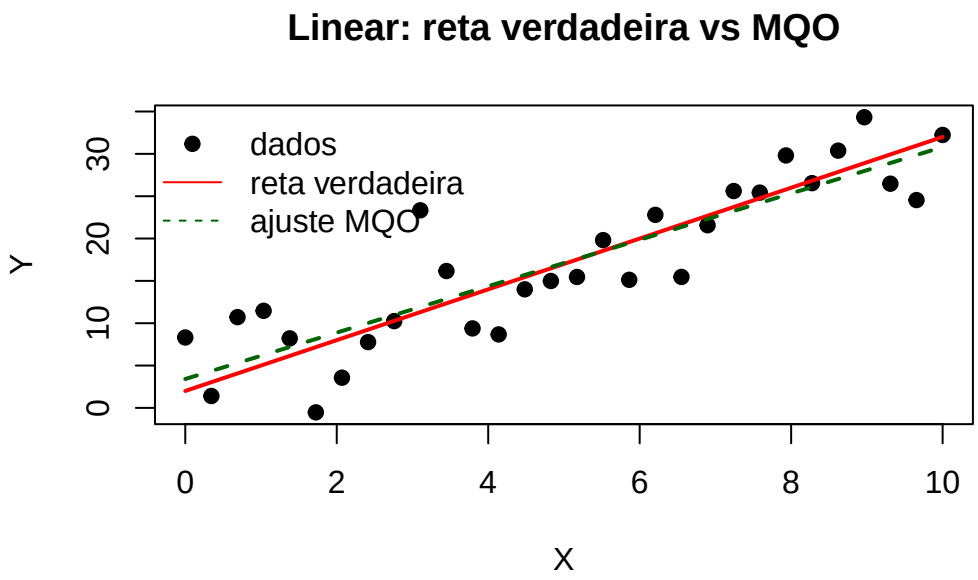
```
[1] 0.8794497
```

3. Ajustar reta por MQO

```r
# 3) Ajustar reta por MQO (OLS)
mod_lin <- lm(y ~ x)

# Resumo do ajuste
summary(mod_lin)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6607 -2.8240 -0.1325  3.0743 11.4207

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4090     1.6321   2.089   0.0459 *
x             2.7402     0.2803   9.777 1.58e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.582 on 28 degrees of freedom
Multiple R-squared:  0.7734,    Adjusted R-squared:  0.7653
F-statistic: 95.58 on 1 and 28 DF,  p-value: 1.581e-10
```

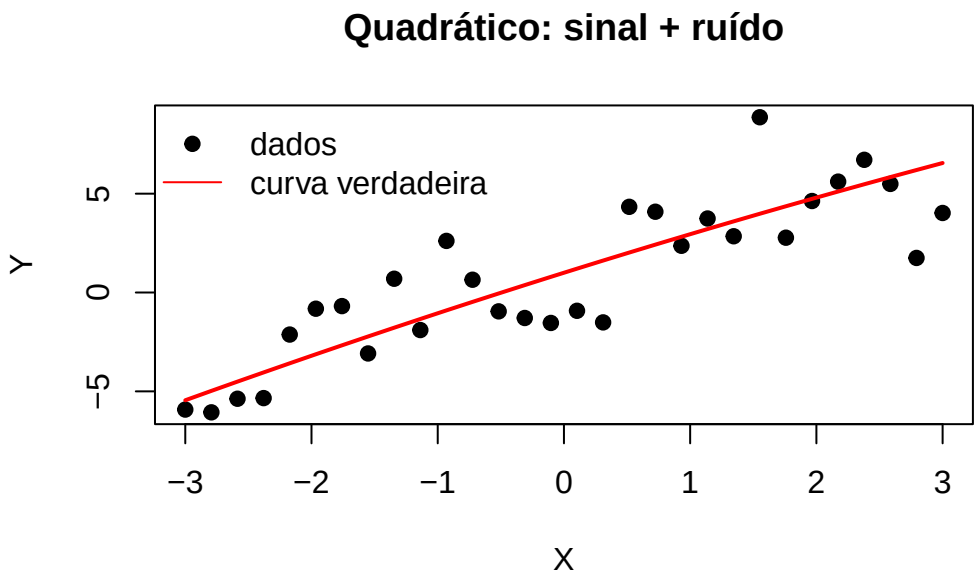
```r
# Visualização: dados + reta verdadeira + reta ajustada
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Linear: reta verdadeira vs MQO")
lines(x, sinal, col = "red", lwd = 2)
lines(x, fitted(mod_lin), col = "darkgreen", lwd = 2, lty = 2)

legend("topleft",
       legend = c("dados", "reta verdadeira", "ajuste MQO"),
       pch = c(19, NA, NA),
       lty = c(NA, 1, 2),
       col = c("black", "red", "darkgreen"),
       bty = "n")
```




Perguntas – linear a) A correlação está próxima de 1? Por quê?
b) O coeficiente estimado da inclinação $\beta_1$ ficou próximo do valor verdadeiro?
c) Experimente:
- Aumente o ruído para `rnorm(30, mean = 0, sd = 10)` e veja o que acontece com a correlação e o ajuste.
- Reduza o ruído para `rnorm(30, mean = 0, sd = 2)` e observe a diferença.

4. Modelo polinomial quadrático

```
# 4) Modelo polinomial quadrático (sinal não linear)
n <- 30
x <- seq(-3, 3, length.out = n)
sigma <- 2

sinal <- 1 + 2*x - 0.05*x^2
y <- sinal + rnorm(n, mean = 0, sd = sigma)

# Visualização: pontos + curva verdadeira
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Quadrático: sinal + ruído")
lines(x, sinal, col = "red", lwd = 2)
legend("topleft", legend = c("dados", "curva verdadeira"),
       pch = c(19, NA), lty = c(NA, 1), col = c("black", "red"), bty = "n")
```




```
# Correlação (linear) pode enganar em relação não linear
cor(x, y)
```

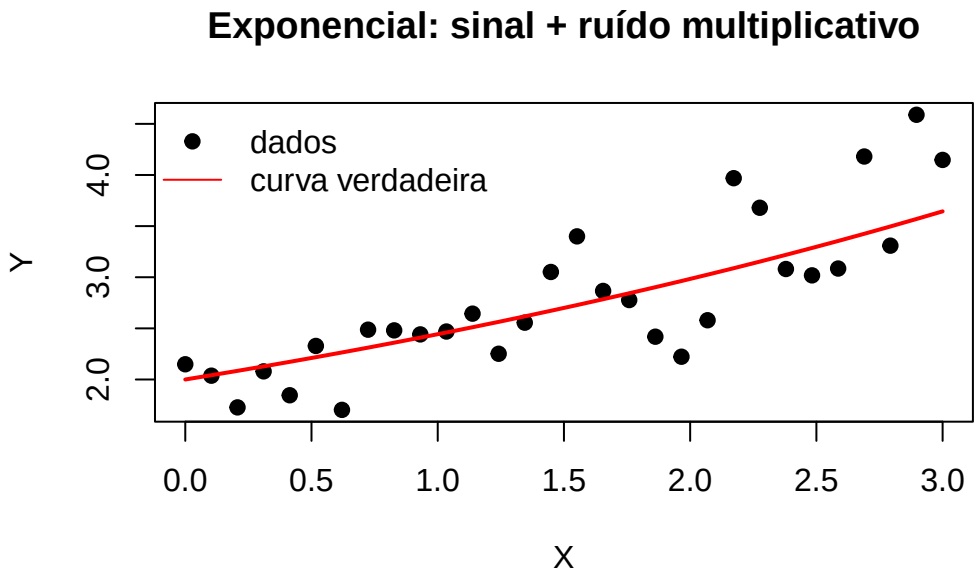
```
[1] 0.8480377
```

Perguntas – Quadrático a) A correlação de Pearson reflete bem a relação entre `x` e `y` neste caso? é linear?
b) Aumente o coeficiente do termo quadrático: troque `-0.05` por `-0.5`. E agora?
c) Altere o sinal do termo quadrático: use `+0.5`. Como fica a concavidade da curva?
d) Aumente o ruído (ex.: `sigma <- 4`). O que acontece com a correlação e a visualização da curva?

5. Modelo Exponencial

```
# 5) Modelo exponencial com ruído multiplicativo
n <- 30
x <- seq(0, 3, length.out = n)

a <- 2
b <- 0.2
sigma_log <- 0.2  # ruído no log (multiplicativo em Y)
```

```r
sinal <- a * exp(b*x)
y <- sinal * exp(rnorm(n, mean = 0, sd = sigma_log))

# Visualização
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Exponencial: sinal + ruído multiplicativo")
lines(x, sinal, col = "red", lwd = 2)
legend("topleft", legend = c("dados", "curva verdadeira"),
       pch = c(19, NA), lty = c(NA, 1), col = c("black", "red"), bty = "n")
```



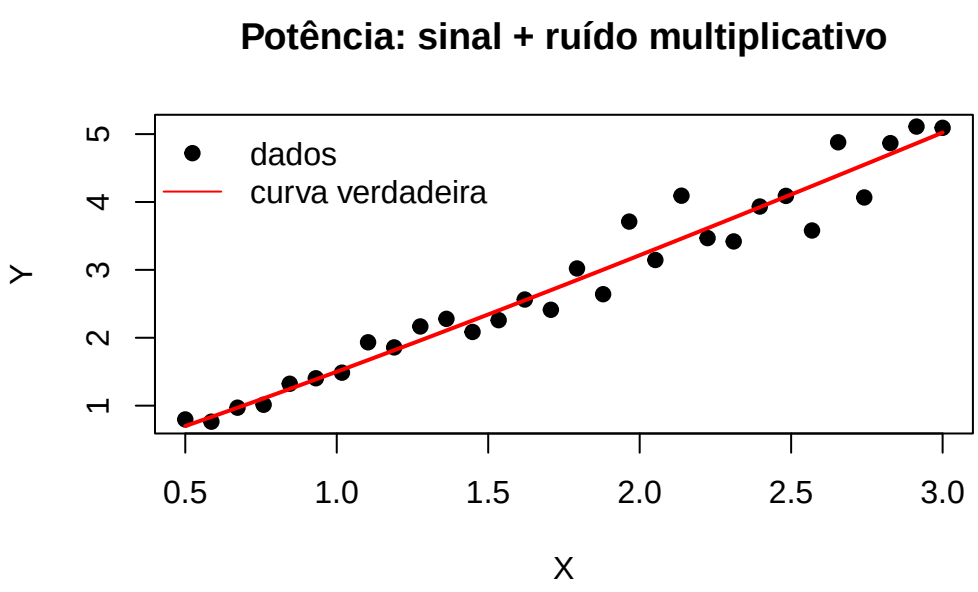

```r
# Correlação
cor(x, y)
```

```
[1] 0.8255609
```

Perguntas – Exponencial a) A curva gerada com $b = 0.2$ parece linear? A correlação confirma isso?
b) Aumente o parâmetro de crescimento para $b = 0.8$. A curva agora se afasta de uma reta?
c) Aumente ainda mais para $b = 1.5$. Como fica a forma da curva e a correlação de Pearson?
d) Reduza o ruído (`0.2 → 0.05`). O que muda na dispersão e na clareza da forma exponencial?

6. Modelo Potência

```r
# 6) Modelo potência com ruído multiplicativo
n <- 30
x <- seq(0.5, 3, length.out = n)

alpha <- 1.5
expoente <- 1.1
sigma_log <- 0.1

sinal <- alpha * (x^expoente)
y <- sinal * exp(rnorm(n, mean = 0, sd = sigma_log))

# Visualização
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Potência: sinal + ruído multiplicativo")
lines(x, sinal, col = "red", lwd = 2)
legend("topleft", legend = c("dados", "curva verdadeira"),
       pch = c(19, NA), lty = c(NA, 1), col = c("black", "red"), bty = "n")
```



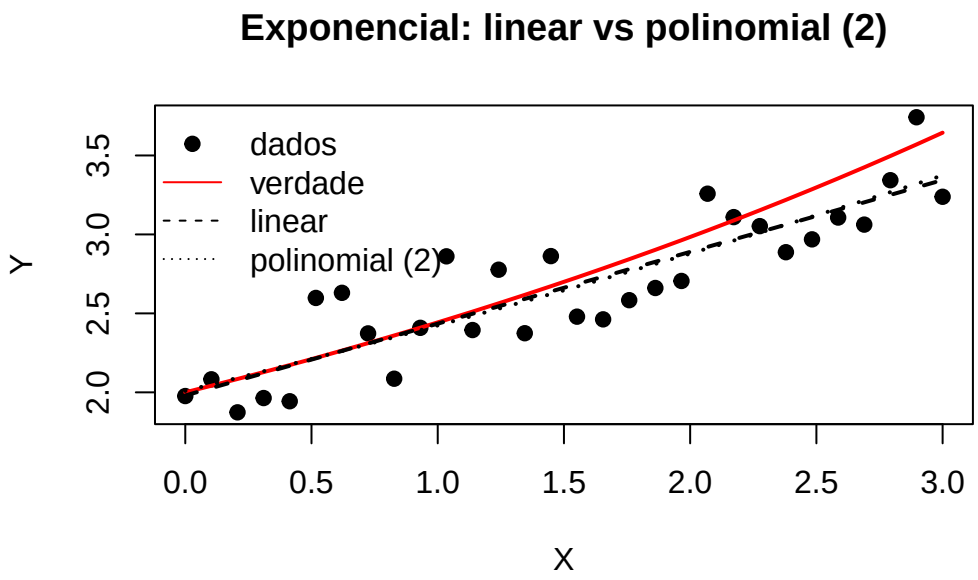

```r
# Correlação
cor(x, y)
```

```
[1] 0.9769569
```

Perguntas – Potência a) Com expoente 1.1, a curva parece linear? A correlação confirma isso?
b) Aumente o expoente para 2.5. A curva agora se distancia de uma reta?
c) Teste com um expoente ainda maior, por exemplo 3.5. O que muda na curvatura e na dispersão dos pontos?
d) Dobre o ruído (`0.1 → 0.2`). O que acontece com a clareza da curva e com a correlação de Pearson?

7. Exponencial — linear vs. polinomial

```r
# 7) Exponencial - comparar ajuste linear vs polinomial (grau 2)
n <- 30
x <- seq(0, 3, length.out = n)

a <- 2.0
b <- 0.2        # depois teste 0.8 e 1.5
sigma_log <- 0.1

sinal <- a * exp(b*x)
y <- sinal * exp(rnorm(n, mean = 0, sd = sigma_log))

# Ajustes
mod_lin <- lm(y ~ x)
mod_poly <- lm(y ~ x + I(x^2))

# Comparação visual
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Exponencial: linear vs polinomial (2)")
lines(x, sinal, col = "red", lwd = 2)
lines(x, fitted(mod_lin), lwd = 2, lty = 2)
lines(x, fitted(mod_poly), lwd = 2, lty = 3)

legend("topleft",
       legend = c("dados", "verdade", "linear", "polinomial (2)"),
       pch = c(19, NA, NA, NA),
       lty = c(NA, 1, 2, 3),
       col = c("black", "red", "black", "black"),
       bty = "n")
```




```
# R² (comparação rápida)
c(R2_linear = summary(mod_lin)$r.squared,
  R2_polinomial2 = summary(mod_poly)$r.squared)
```

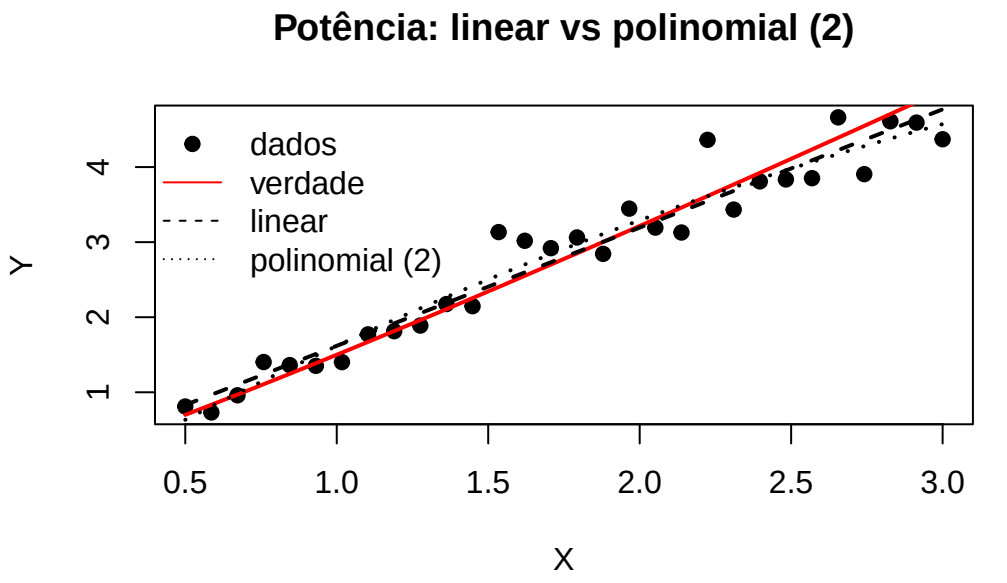
```
R2_linear R2_polinomial2
0.7751012      0.7760694
```

Perguntas — Exponencial (começando quase reta) a) Com **b = 0.2**, os ajustes **linear** e **polinomial (2)** parecem semelhantes? O $R^2$ confirma? b) Aumente **b** para **0.8** e depois **1.5**. Como mudam o gráfico e os $R^2$? Qual ajuste passa a representar melhor a curva? c) Reduza o ruído para `sigma_log <- 0.05`. Fica mais fácil perceber a diferença entre o linear e o polinomial quando **b** é maior?

8. Potência — linear vs. polinomial

```
# 8) Potência - comparar ajuste linear vs polinomial (grau 2)
n <- 30
x <- seq(0.5, 3, length.out = n)

alpha <- 1.5
expoente <- 1.1   # depois teste 2.5 e 3.5
sigma_log <- 0.1
```

```r
sinal <- alpha * (x^expoente)
y <- sinal * exp(rnorm(n, mean = 0, sd = sigma_log))

# Ajustes
mod_lin <- lm(y ~ x)
mod_poly <- lm(y ~ x + I(x^2))

# Comparação visual
plot(x, y, pch = 19, xlab = "X", ylab = "Y",
     main = "Potência: linear vs polinomial (2)")
lines(x, sinal, col = "red", lwd = 2)
lines(x, fitted(mod_lin), lwd = 2, lty = 2)
lines(x, fitted(mod_poly), lwd = 2, lty = 3)

legend("topleft",
       legend = c("dados", "verdade", "linear", "polinomial (2)"),
       pch = c(19, NA, NA, NA),
       lty = c(NA, 1, 2, 3),
       col = c("black", "red", "black", "black"),
       bty = "n")
```



```r
# R² (comparação rápida)
c(R2_linear = summary(mod_lin)$r.squared,
  R2_polinomial2 = summary(mod_poly)$r.squared)
```

```
R2_linear R2_polinomial2
0.9422860      0.9487419
```

Perguntas — Potência (começando quase reta)

a) Com **expoente = 1.1**, os ajustes **linear** e **polinomial (2)** parecem semelhantes? O $R^2$ confirma?

b) Aumente o **expoente** para **2.5** e depois **3.5**. Como mudam o gráfico e os $R^2$? Qual ajuste passa a representar melhor a curva?

c) Dobre o ruído para `sigma_log <- 0.2`. Fica mais difícil perceber a diferença entre o linear e o polinomial quando o expoente é maior?

# Parte II

# Parte II — Modelo de Regressão Linear Simples (MRLS)

# 4 O MRLS como Modelo para a Média Condicional

A compreensão do **Modelo de Regressão Linear Simples (MRLS)** é essencial para o estudo dos modelos de regressão. Sua importância não se limita à simplicidade algébrica, mas repousa no fato de que ele estabelece as bases conceituais para toda a teoria de modelagem estatística. Ao assumir que a variação média de uma variável resposta $Y$ pode ser explicada por uma única variável explicativa $X$, o MRLS introduz a noção central de **média condicional**, isto é, a ideia de que existe uma estrutura determinística que organiza o comportamento médio dos dados, à qual se sobrepõe uma componente aleatória que representa o ruído *inevitável* das observações empíricas.

Essa leitura "pela média condicional" é a forma mais precisa de entender o que a regressão linear simples afirma: para cada valor fixado de $X$, existe uma distribuição de $Y$, e o modelo especifica como a média dessa distribuição varia com $X$. (ver Montgomery, Peck, e Vining (2021); Hoffmann (2016))

O intercepto e a inclinação da reta de regressão apresentadas anteriormente traduzem a parte sistemática do fenômeno, enquanto o erro agrega fatores não observados, variações aleatórias ou imprecisões de medição. É nessa combinação entre regularidade e aleatoriedade que se encontra a força do modelo: a regressão linear simples oferece uma linguagem matemática capaz de quantificar associações e, ao mesmo tempo, de reconhecer que o mundo real não se comporta de maneira perfeitamente determinística.

Em particular, o termo "erro" não deve ser lido como "falha": ele representa a parcela de variabilidade de $Y$ que permanece mesmo quando $X$ é conhecido e o componente médio $E(Y \mid X)$ foi especificado. (ver Kutner et al. (2005))

O MRLS pode ser usado em diferentes perspectivas. Em um primeiro plano, ele ajuda a compreender como uma variável se relaciona com outra, permitindo isolar a contribuição média de $X$ sobre $Y$. Em seguida, oferece meios de previsão, já que a reta ajustada pode ser utilizada para estimar valores futuros ou não observados de $Y$. Finalmente, ele fornece um instrumento de controle, pois ao quantificar a variação esperada em $Y$ para uma mudança em $X$, torna-se possível avaliar de forma objetiva a influência de um fator específico mantendo os demais aspectos fixos ou controlados no desenho do estudo. Essa tríade, nomeadamente; explicação, predição e controle, sustenta a relevância prática do modelo e justifica sua centralidade tanto no ensino quanto na aplicação da estatística.

Um cuidado conceitual importante é distinguir "prever o valor médio" de "prever uma observação individual": mesmo que a média condicional seja bem descrita, observações individuais ainda variam ao redor dessa média por causa do erro aleatório. (ver Montgomery, Peck, e Vining (2021))

A intuição do MRLS pode ser visualizada em gráficos de dispersão: os pontos $(X, Y)$ representam as observações empíricas e, sobre esse conjunto, a reta de regressão traduz a tendência média. As distâncias verticais entre cada ponto e a reta correspondem aos resíduos, isto é, às variações não capturadas pelo modelo. Adicionalmente, sabendo que a relação entre $X$ e $Y$ pode assumir diferentes intensidades e direções, mesmo dentro de um modelo linear simples, considere os quatro cenários a seguir, todos baseados em uma reta verdadeira perturbada por erros aleatórios:

- **(a) Correlação positiva forte**: a inclinação da reta é positiva e os pontos se distribuem próximos a ela. O sinal determinístico domina, e o ruído é pequeno em relação à estrutura média.

- **(b) Correlação positiva fraca**: a reta mantém inclinação positiva, mas a dispersão em torno dela é elevada. O sinal ainda existe, mas é encoberto por grande variabilidade aleatória.

- **(c) Correlação negativa forte**: a inclinação é negativa e os pontos se alinham de forma clara em torno da reta decrescente. A média condicional é bem definida e o erro exerce papel secundário.

- **(d) Correlação negativa fraca**: a inclinação é negativa, porém os pontos apresentam grande dispersão em torno da reta. A variabilidade do erro é tão relevante quanto a estrutura determinística, tornando a associação menos evidente.

Essas quatro situações destacam a essência do **MRLS**: independentemente da direção ou da força da associação, o modelo parte da ideia de que a média condicional de $Y$ pode ser descrita por uma função linear em $X$, à qual se soma um ruído $\varepsilon_i$ com $E[\varepsilon_i \mid X_i] = 0$.

Vale enfatizar por que essa condição é conceitualmente importante: ela expressa que, uma vez fixado $X_i$, o termo de erro não tem *tendência sistemática* (em média, não empurra $Y$ para cima nem para baixo), de modo que toda a variação média de $Y$ com $X$ fica concentrada no termo $E(Y \mid X)$. Quando essa condição falha, o que se interpreta como "efeito de $X$" pode estar contaminado por fatores omitidos que variam com $X$ (situação que, em econometria, está relacionada à ideia de endogeneidade). O gráfico, portanto, antecipa de forma intuitiva a formulação matemática que será detalhada na próxima seção.

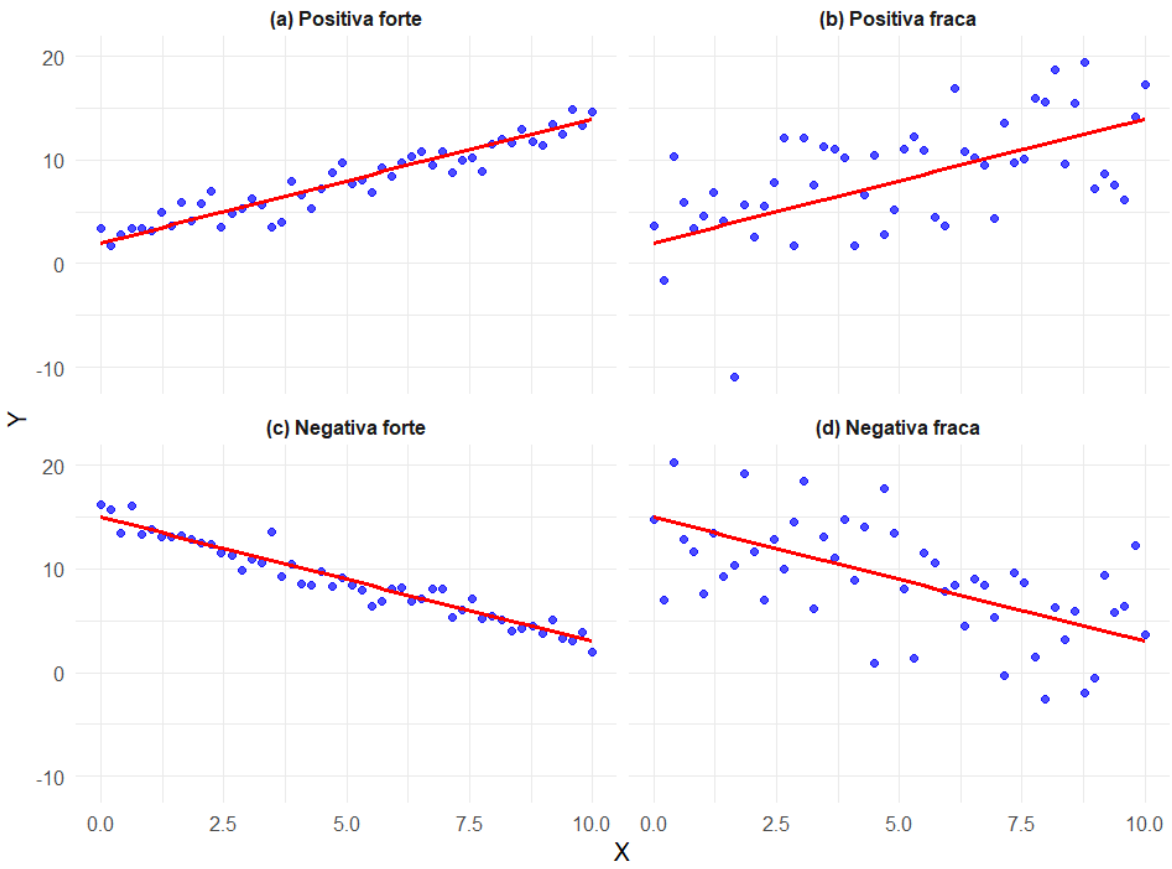



Figura 4.1: Quatro cenários de associação linear: (a) positiva forte, (b) positiva fraca, (c) negativa forte, (d) negativa fraca.

Essa representação gráfica simples, mas poderosa, revela a lógica do modelo: a regressão não busca explicar cada observação em particular, mas descrever o comportamento médio de $Y$ em função de $X$. É nesse sentido que ela constitui uma primeira aproximação, um alicerce sobre o qual se constroem modelos mais complexos. (ver Montgomery, Peck, e Vining (2021))

## 4.1 Formulação Matemática do MRLS

A formulação do modelo de regressão linear simples parte da ideia de que cada observação $Y_i$ pode ser decomposta em duas partes: uma componente sistemática, que expressa o valor médio de $Y$ condicionado a um dado valor de $X_i$, e uma componente aleatória, que traduz as variações não explicadas. Em notação formal, escrevemos

$$Y_i = \mu_i + \varepsilon_i, \quad \text{com} \quad \mu_i = E[Y_i \mid X_i],$$

em que $\mu_i$ é a média condicional de $Y$ dado $X_i$, e $\varepsilon_i$ é o erro aleatório associado à observação $i$.

Para que essa decomposição tenha interpretação estatística clara, é conveniente explicitar as **suposições** (ou hipóteses) que conectam o termo aleatório à componente sistemática. A hipótese central é que o erro, em média, não carrega informação adicional além de $X_i$, de modo que

$$E[\varepsilon_i \mid X_i] = 0.$$

Essa condição de exogeneidade fraca garante que a parte sistemática do modelo seja, de fato, uma descrição da média condicional, e não uma mistura entre efeito sistemático e ruído (ver Gujarati (2006)). Um modo equivalente (e útil) de ler essa hipótese é: ao fixar $X_i$, a média de $Y_i$ é exatamente $\mu_i$, pois

$$E(Y_i \mid X_i) = E(\mu_i + \varepsilon_i \mid X_i) = \mu_i + E(\varepsilon_i \mid X_i) = \mu_i.$$

Além da condição de média nula, frequentemente se acrescenta uma hipótese sobre a dispersão do erro, que relaciona o modelo à variância condicional de $Y$:

$$Var(\varepsilon_i \mid X_i) = \sigma^2.$$

Por fim, para que as observações tragam informação "nova" umas em relação às outras e para que os resultados usuais de estimação e inferência sejam válidos, costuma-se assumir ausência de dependência linear entre erros de unidades distintas. Uma forma padrão de expressar isso é impor **covariância nula** entre erros diferentes:

$$Cov(\varepsilon_i, \varepsilon_j \mid X_i, X_j) = 0, \quad \forall i \neq j,$$

isto é, condicionando ao conjunto de regressores, os termos de erro não apresentam associação linear entre observações distintas. Em muitos textos, essa hipótese aparece na forma mais forte de independência entre os erros; a condição de covariância nula é a expressão mínima necessária para várias propriedades algébricas clássicas do modelo linear. (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021))

Sob essas suposições, podemos derivar **propriedades imediatas** do modelo:

1) **Média condicional**

$$E(Y_i \mid X_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i \mid X_i) = \beta_0 + \beta_1 X_i.$$

2) **Variância condicional**

$$Var(Y_i \mid X_i) = E\left\{[Y_i - E(Y_i \mid X_i)]^2 \mid X_i\right\} = E(\varepsilon_i^2 \mid X_i) = Var(\varepsilon_i \mid X_i) = \sigma^2.$$

3) **Covariância condicional entre observações distintas**

Para $i \neq j$, e condicionando ao conjunto de regressores), temos

$$Cov(Y_i, Y_j \mid X_i, X_j) = Cov(\beta_0 + \beta_1 X_i + \varepsilon_i,\ \beta_0 + \beta_1 X_j + \varepsilon_j \mid X_i, X_j) = Cov(\varepsilon_i, \varepsilon_j \mid X_i, X_j).$$

Assim, sob a hipótese $Cov(\varepsilon_i, \varepsilon_j \mid X_i, X_j) = 0$ para $i \neq j$, segue que

$$Cov(Y_i, Y_j \mid X_i, X_j) = 0, \quad i \neq j.$$

A função de regressão do modelo é, portanto, a própria média condicional (ou média do componente sistemático):

$$\mu(X_i; \beta_0, \beta_1) = E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i.$$

No caso linear, supõe-se que a média condicional é uma função linear nos parâmetros, de forma que

$$E[Y_i \mid X_i] = \mu(X_i; \beta_0, \beta_1) = \beta_0 + \beta_1 X_i,$$

o que leva à formulação completa do modelo:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Uma vez especificada a função $\mu(X_i; \beta_0, \beta_1)$ e as hipóteses sobre $\varepsilon_i$, a tarefa estatística passa a ser **estimar** $E(Y_i \mid X_i)$ (isto é, a função de regressão) a partir dos dados, tipicamente via métodos como mínimos quadrados (MMQ) e, quando apropriado, máxima verossimilhança (MMV). (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021))

## 4.2 Interpretação dos Parâmetros e Hipóteses do Modelo

Nessa estrutura,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

$\beta_0$ representa o valor esperado de $Y$ quando $X = 0$, isto é,

$$E(Y \mid X = 0) = \beta_0,$$

enquanto $\beta_1$ indica a taxa média de variação de $Y$ a cada incremento unitário em $X$:

$$E(Y \mid X = x + 1) - E(Y \mid X = x) = \beta_1.$$

Assim, $\beta_1$ mede o efeito médio de uma variação unitária em $X$ sobre a média condicional de $Y$. O erro $\varepsilon_i$ traduz tanto a variabilidade natural dos fenômenos quanto fatores não observados, assumindo sempre que sua esperança condicional a $X_i$ seja nula.

É essencial notar que "$X = 0$" pode não ter significado em alguns contextos; ainda assim, $\beta_0$ permanece necessário como parâmetro de localização da reta. Quando 0 não pertence ao intervalo observado de $X$, ou quando não possui interpretação prática, pode-se redefinir a variável explicativa por centralização (por exemplo, $X_i^* = X_i - \bar{X}$), de modo que o novo intercepto represente o valor esperado de $Y$ em um ponto de referência mais informativo. (ver Montgomery, Peck, e Vining (2021))

## 4.3 Hipóteses do MRLS

Para que o modelo tenha propriedades estatísticas bem definidas, é conveniente explicitar as **hipóteses clássicas do MRLS**, usualmente apresentadas na literatura de modelos lineares (ver Kutner et al. (2005)). Sob a formulação clássica, assumimos:

1) **Linearidade nos parâmetros**

A função de regressão é linear nos parâmetros:

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

Observe que a linearidade refere-se aos parâmetros $\beta_0, \beta_1$, e não necessariamente à variável $X$ em si.

2) **Valores de $X$ fixos (ou condicionais)**

Os valores $X_i$ são considerados fixos pelo planejamento do estudo, ou, alternativamente, a análise é conduzida condicionalmente aos valores observados de $X$. Essa hipótese garante que toda a aleatoriedade do modelo esteja concentrada no termo de erro.

3) **Erro com média zero**

$$E(\varepsilon_i \mid X_i) = 0.$$

Essa condição assegura que o componente sistemático do modelo coincide com a média condicional de $Y$.

4) **Homoscedasticidade**

A variância do erro é constante para todos os valores de $X$:

$$Var(\varepsilon_i \mid X_i) = \sigma^2.$$

Consequentemente,

$$Var(Y_i \mid X_i) = \sigma^2.$$

5) **Ausência de correlação entre erros**

Para $i \neq j$,

$$Cov(\varepsilon_i, \varepsilon_j \mid X) = 0.$$

Sob essa hipótese, segue que

$$Cov(Y_i, Y_j \mid X) = 0, \quad \forall i \neq j.$$

Essas cinco condições compõem o conjunto clássico de hipóteses do modelo linear simples conforme apresentado em textos de modelos lineares aplicados (ver Kutner et al. (2005)).

### 4.3.1 Hipóteses adicionais para inferência

Até este ponto, não foi necessário supor nenhuma distribuição específica para os erros. As hipóteses acima são suficientes para garantir propriedades como não-viesamento dos estimadores de mínimos quadrados e expressões fechadas para suas variâncias.

Para a construção de intervalos de confiança exatos e testes de hipóteses com distribuição conhecida em amostras finitas, acrescenta-se frequentemente a suposição de **normalidade dos erros**:

$$\varepsilon_i \sim N(0, \sigma^2).$$

Sob essa condição, o vetor de respostas possui distribuição normal multivariada condicional a $X$, o que permite derivar resultados exatos para estatísticas $t$ e $F$.

É conceitualmente importante distinguir:

- **Hipóteses do modelo básico**: linearidade em $\beta_0, \beta_1$ e $E(\varepsilon \mid X) = 0$;
- **Hipóteses para eficiência e inferência exata**: homoscedasticidade, ausência de correlação e normalidade.

Essa distinção é enfatizada na literatura de regressão aplicada, que separa claramente a estrutura do modelo da estrutura probabilística necessária para inferência (ver Weisberg (2005)).

Por fim, essa formulação pode ser interpretada sob duas perspectivas equivalentes. Na abordagem clássica, $X$ é tratado como fixo. Em contextos amostrais, pode-se admitir $X$ aleatório, desde que se mantenha a condição

$$E(\varepsilon_i \mid X_i) = 0,$$

que garante a validade das propriedades do modelo condicionalmente a $X$. Em ambas as leituras, permanece a essência: a regressão linear simples é um modelo para a média condicional de $Y$ dado $X$, e não para cada observação individual. (ver Gujarati (2006))

## 4.4 Representação Gráfica e Intuição Geométrica

A Figura a seguir ilustra o que significa afirmar que o **MRLS é um modelo para a média condicional**. Os pontos azuis representam as observações empíricas $(X_i, Y_i)$, que se espalham devido ao ruído aleatório $\varepsilon_i$. A reta vermelha mostra a estrutura determinística $\beta_0 + \beta_1 X$, isto é, o valor esperado de $Y$ para cada valor de $X$. Já os círculos pretos conectados indicam médias locais de $Y$ em diferentes intervalos de $X$, funcionando como uma aproximação empírica de

$E[Y \mid X]$. O alinhamento dessas médias com a reta reforça a ideia de que o modelo busca descrever a **tendência média** e não cada observação individual.

Este gráfico torna visível a condição fundamental $E[\varepsilon_i \mid X_i] = 0$. Embora cada ponto esteja sujeito a variações não explicadas, quando olhamos para a média em cada faixa de $X$, os erros se compensam e a estrutura linear emerge. Assim, é possível notar intuitivamente por que se fala em "média condicional" e compreendemos que a regressão não elimina o ruído, mas organiza o comportamento médio das observações em torno de uma reta. (ver Charnet et al. (2008))

Mais adiante, quando abordarmos o método dos mínimos quadrados ordinários, introduziremos outra visualização complementar, na qual os resíduos aparecem como segmentos verticais entre os pontos observados e a reta ajustada. Essas representações reforçam a interpretação fundamental: o MRLS não pretende capturar cada realização individual, mas descrever a tendência média de $Y$ em função de $X$, admitindo explicitamente a presença de ruído.

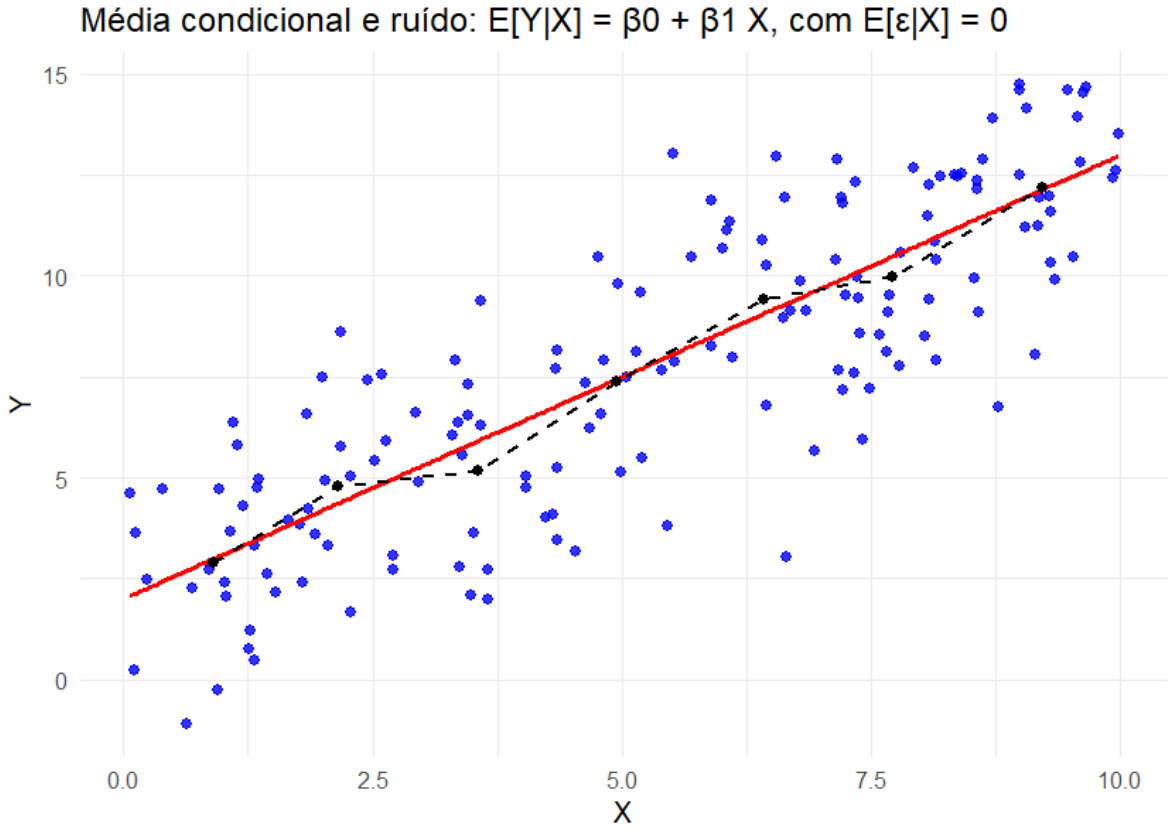



Figura 4.2: MRLS como média condicional: pontos observados (X,Y), reta verdadeira (média condicional) e médias locais de Y por faixas de X.

# 5 Estimação por Mínimos Quadrados no MRLS

## 5.1 Paradigmas de Estimação no MRLS

A formulação do **Modelo de Regressão Linear Simples (MRLS)**, discutida anteriormente, descreve a estrutura da média condicional de $Y$ em função de $X$, isto é,

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

O desafio agora é **estimar os parâmetros desconhecidos** $\beta_0$ e $\beta_1$ a partir de dados observados $\{(X_i, Y_i)\}_{i=1}^n$. Esse processo de estimação pode ser realizado via diferentes métodos, cada um apoiado em princípios e hipóteses próprias.

A estimação pode ser conduzida sob diferentes **paradigmas**, isto é, diferentes princípios fundamentais que definem o que significa "estimar bem" um parâmetro. Esses paradigmas não diferem apenas em técnica, mas em filosofia estatística e nas hipóteses assumidas sobre o modelo.

O **Método dos Mínimos Quadrados Ordinários (MQO)** é a abordagem clássica no contexto da regressão linear. Seu princípio é puramente geométrico e algébrico: escolher $\widehat{\beta}_0$ e $\widehat{\beta}_1$ de modo a minimizar a soma dos quadrados dos resíduos,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2.$$

Esse critério não exige, para a obtenção dos estimadores, a especificação de uma distribuição para os erros. A minimização conduz a um sistema de equações conhecido como **equações normais**, que caracteriza a solução de mínimos quadrados (ver Draper e Smith (1998); Montgomery, Peck, e Vining (2021)). A ausência de suposição distributiva mostra que o MQO é, antes de tudo, um procedimento de ajuste determinístico baseado na estrutura linear do modelo.

Do ponto de vista estatístico, sob as hipóteses já apresentadas, a saber, linearidade nos parâmetros, $E(\varepsilon_i \mid X_i) = 0$, homoscedasticidade e ausência de correlação entre erros os estimadores de MQO possuem propriedades fundamentais como não viés e variâncias com forma explícita. Essas propriedades não dependem da normalidade dos erros; a normalidade é necessária apenas quando se desejam distribuições exatas em amostras finitas para testes e intervalos de confiança (ver Kutner et al. (2005)). Assim, o MQO é um método de estimação

que se apoia primariamente na estrutura do modelo médio e nas condições de regularidade, e não em hipóteses distributivas fortes.

Outro caminho é o **Método da Máxima Verossimilhança (MV)**. Nesse paradigma, parte-se da especificação completa da distribuição condicional de $Y_i \mid X_i$, frequentemente assumindo

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

o que implica que $Y_i \mid X_i$ também segue distribuição normal com média $\mu_i = \beta_0 + \beta_1 X_i$ e variância $\sigma^2$. Os estimadores são então definidos como aqueles que maximizam a função de verossimilhança, isto é, a probabilidade conjunta dos dados observados vista como função dos parâmetros. Quando o modelo probabilístico está corretamente especificado, a MV produz estimadores consistentes, assintoticamente normais e eficientes sob condições regulares (ver Casella e Berger (2002)).

No caso particular do modelo linear com erros normais homoscedásticos e não correlacionados, os estimadores de máxima verossimilhança coincidem com os estimadores de mínimos quadrados. Essa coincidência não é acidental: a minimização da soma de quadrados é equivalente à maximização da verossimilhança normal. Contudo, conceitualmente, os dois métodos partem de princípios distintos, um geométrico/algebraico e outro probabilístico.

Uma terceira alternativa são os **métodos bayesianos**, nos quais os parâmetros $\beta_0$ e $\beta_1$ são tratados como variáveis aleatórias. Nesse caso, especifica-se uma distribuição a priori conjunta para $\beta_0$ e $\beta_1$ e combina-se essa informação com a verossimilhança dos dados por meio do Teorema de Bayes, obtendo-se a distribuição a posteriori

$$p(\beta_1, \beta_2 \mid y, X) \propto p(y, X \mid \beta_1, \beta_2)\, p(\beta_1, \beta_2).$$

A estimação passa então a ser baseada em características dessa distribuição a posteriori (como média, mediana ou moda). Esse paradigma explicita a incerteza sobre os parâmetros e permite incorporar informação prévia de forma formal (ver Casella e Berger (2002); Gelman et al. (2014)).

Portanto, a estimação no MRLS pode ser conduzida sob diferentes paradigmas: minimização de resíduos (MQO), maximização da verossimilhança (MV) ou atualização bayesiana de crenças. Cada abordagem parte de fundamentos conceituais distintos, tais quais, geométrico-algébrico, probabilístico ou epistemológico, e conduz a interpretações próprias dos parâmetros e da incerteza associada.

Neste livro, a **estimação por mínimos quadrados ordinários (MQO)** receberá tratamento mais detalhado e sistemático. A razão é dupla: em primeiro lugar, o MQO não exige a especificação de uma distribuição para os erros para a obtenção dos estimadores, apoiando-se apenas na estrutura do modelo médio e nas hipóteses clássicas de exogeneidade e regularidade;

em segundo lugar, ele constitui a base do Teorema de Gauss–Markov e de grande parte da teoria dos modelos lineares, servindo como alicerce conceitual para extensões posteriores.

A **máxima verossimilhança (MV)** também será contemplada, sobretudo quando discutirmos aspectos inferenciais e conexões entre estrutura probabilística e eficiência assintótica. No caso do modelo linear com erros normais, veremos inclusive a coincidência formal entre MQO e MV, o que reforça a unidade conceitual entre os métodos sob hipóteses adicionais.

Por outro lado, embora o paradigma **bayesiano** seja conceitualmente relevante e metodologicamente poderoso, sua abordagem completa exigiria o desenvolvimento de ferramentas próprias, como escolha de distribuições a priori, análise da posteriori e métodos computacionais, que extrapolam os objetivos centrais deste texto. Assim, ele será mencionado para fins de contextualização, mas não será desenvolvido formalmente neste livro.

## 5.2 O Critério dos Mínimos Quadrados Ordinários

O Método dos Mínimos Quadrados Ordinários (MQO) é a abordagem clássica para a estimação em regressão linear. Seu objetivo é encontrar a reta que melhor descreve a relação média entre a variável resposta $Y$ e a variável explicativa $X$. Essa "melhor" reta é definida como aquela que minimiza a soma dos quadrados dos **resíduos**, isto é, das diferenças entre os valores observados e os valores ajustados pelo modelo (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Se denotarmos por $\hat{Y}_i$ o valor ajustado para a observação $i$, temos:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

e o resíduo correspondente é

$$e_i = Y_i - \hat{Y}_i.$$

O critério de mínimos quadrados escolhe os parâmetros que minimizam a função de perda quadrática

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[ Y_i - (\beta_0 + \beta_1 X_i) \right]^2 .$$

Do ponto de vista matemático, trata-se de um problema clássico de otimização: encontrar $(\hat{\beta}_0, \hat{\beta}_1)$ que minimizam $S(\beta_0, \beta_1)$ sobre $\mathbb{R}^2$. A condição de primeira ordem leva a um sistema de duas equações lineares nas incógnitas $\beta_0$ e $\beta_1$, conhecido como **equações normais**. Essas equações caracterizam completamente a solução de mínimos quadrados no modelo linear simples (ver Draper e Smith (1998)).

Esse procedimento garante que, entre todas as retas possíveis, a escolhida é aquela que deixa os resíduos, em conjunto, "o mais curtos possível" no sentido quadrático. A escolha da penalização quadrática não é arbitrária: a função objetivo é uma função polinomial de segundo grau nos parâmetros, contínua e diferenciável, e admite solução única sempre que os valores de $X$ não forem todos iguais, isto é, sempre que houver variabilidade na variável explicativa. Essa condição assegura a existência e a unicidade da reta de mínimos quadrados.

Além disso, a penalização pelo quadrado dos desvios atribui maior peso a observações mais afastadas, o que explica tanto a eficiência do método sob hipóteses clássicas quanto sua sensibilidade a valores discrepantes (ver Weisberg (2005)).

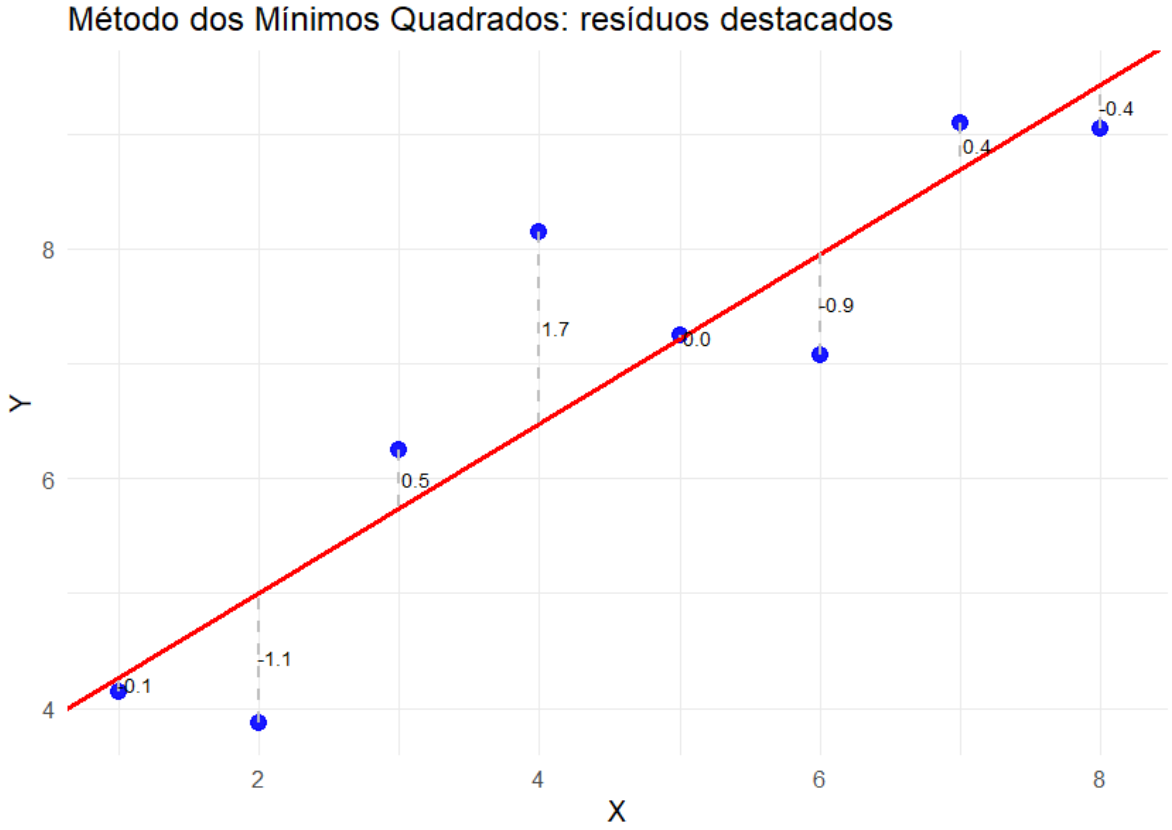



Figura 5.1: MQO: reta ajustada e resíduos destacados.

A figura acima ilustra essa lógica. Os pontos azuis representam as observações $(X_i, Y_i)$, a reta vermelha mostra a reta ajustada pelo MQO, e as linhas tracejadas cinzas indicam os resíduos associados a cada ponto. Visualmente, o MQO busca a reta que minimiza a soma dos quadrados dessas distâncias verticais. Essa interpretação geométrica ajuda a compreender que a regressão não elimina o erro, mas organiza o ruído de forma a recuperar a estrutura média do fenômeno.

Essa formulação admite duas interpretações complementares.

- **Geométrica**: o MQO pode ser visto como a projeção ortogonal do vetor de respostas $\mathbf{Y} = (Y_1, \dots, Y_n)'$ no subespaço gerado pelos vetores $\mathbf{1} = (1, \dots, 1)'$ e $\mathbf{X} = (X_1, \dots, X_n)'$. A condição de minimização implica que o vetor de resíduos $\hat{\varepsilon} = Y - \hat{Y}$ é ortogonal ao espaço gerado pelos regressores, isto é,

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \quad \text{e} \quad \sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0.$$

Essas duas condições são precisamente as equações normais no caso simples.

- **Estatística**: a ortogonalidade amostral dos resíduos aos regressores é o análogo empírico da hipótese populacional

$$E(\varepsilon_i \mid X_i) = 0.$$

Em outras palavras, após o ajuste, não resta componente linear em $X$ capaz de explicar sistematicamente os resíduos. A condição populacional de exogeneidade é refletida, no nível amostral, pela ortogonalidade dos resíduos estimados (ver Kutner et al. (2005)).

Um aspecto central é que o MQO não exige, para a obtenção dos estimadores, a especificação de uma distribuição para os erros. Sob as hipóteses de média condicional corretamente especificada, homoscedasticidade e ausência de correlação entre erros, os estimadores resultantes são não viesados e apresentam variâncias com forma explícita, propriedades que independem da normalidade (ver Montgomery, Peck, e Vining (2021)).

A normalidade é introduzida apenas quando se desejam distribuições exatas em amostras finitas para estatísticas de teste e construção de intervalos de confiança. Em contextos práticos com caudas pesadas ou observações discrepantes, podem ser considerados métodos robustos ou funções de perda alternativas. Essa generalidade explica por que o MQO constitui o ponto de partida natural e o método mais amplamente ensinado e utilizado na análise de regressão linear.

## 5.3 Solução Analítica: A Reta de Regressão por MQO

**!** Teorema — Reta de Regressão por MQO

No modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

os estimadores de mínimos quadrados ordinários (MQO) de $\beta_0$ e $\beta_1$ são obtidos como aqueles que **minimizam a soma dos quadrados dos resíduos**. A solução do problema de minimização leva às formas fechadas:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

onde

$$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2, \qquad S_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

A demonstração resulta da minimização da função

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[ Y_i - (\beta_0 + \beta_1 X_i) \right]^2,$$

por meio do cálculo das derivadas parciais em relação a $\beta_0$ e $\beta_1$ e da resolução do sistema de equações normais correspondente. A dedução algébrica completa pode ser consultada no Apêndice de Demonstrações {#demo}.

Esse resultado estabelece a **reta de regressão por MQO** como a linha que, ao mesmo tempo, minimiza a soma dos quadrados dos resíduos e traduz o padrão médio de associação entre $X$ e $Y$. A estrutura explícita das soluções mostra que a existência e a unicidade dependem apenas de $S_{xx} > 0$, isto é, da presença de variabilidade em $X$, condição necessária para que a informação sobre a inclinação seja identificável (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Do ponto de vista matemático, a minimização de $S(\beta_0, \beta_1)$ conduz a uma função quadrática estritamente convexa nos parâmetros quando $S_{xx} > 0$, assegurando que a solução encontrada pelas equações normais seja única. A demonstração detalhada dessa propriedade pode ser consultada no Apêndice de Demonstrações {#demo}.

No entanto, conhecer a forma explícita da reta ajustada é apenas o primeiro passo. A expressão fechada dos estimadores revela como eles dependem das quantidades amostrais, mas não informa, por si só, se tais estimadores são centrados nos verdadeiros parâmetros, quão precisos são ou como se comportam sob repetição amostral. Para que possamos confiar nesses estimadores e

utilizá-los em inferência estatística, precisamos examinar suas **propriedades probabilísticas**: não viés, variâncias, covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$ e qualidade das predições produzidas. É justamente esse o foco da próxima seção (ver Kutner et al. (2005)).

### 5.3.1 Interpretação dos Estimadores obtidos via MQO

- O estimador da inclinação pode ser reescrito como

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

o que evidencia que ele corresponde à **covariância amostral entre $X$ e $Y$ dividida pela variância amostral de $X$**. Essa forma deixa claro que $\hat{\beta}_1$ mede a variação média de $Y$ associada a um aumento unitário em $X$, sendo proporcional ao grau de associação linear entre as duas variáveis (ver Montgomery, Peck, e Vining (2021)).

- O estimador do intercepto,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

implica que a reta ajustada satisfaz

$$\hat{Y}(\bar{X}) = \bar{Y},$$

ou seja, a reta de regressão **passa necessariamente pelo ponto médio amostral** $(\bar{X}, \bar{Y})$. Essa propriedade decorre diretamente das equações normais e da ortogonalidade dos resíduos aos regressores (ver Kutner et al. (2005)).

## 5.4 Propriedades Probabilísticas dos Estimadores de MQO

Nesta seção reunimos as propriedades essenciais dos estimadores de mínimos quadrados no modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

sob as hipóteses usuais de exogeneidade fraca e regularidade:

$$E[\varepsilon_i \mid X_i] = 0, \qquad Var(\varepsilon_i \mid X_i) = \sigma^2, \qquad Cov(\varepsilon_i, \varepsilon_j \mid X_i, X_j) = 0 \ \forall (i \neq j),$$

com

$$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2 > 0.$$

Essas condições são suficientes para estabelecer as principais propriedades dos estimadores de MQO, sem necessidade de assumir normalidade dos erros. Trata-se exatamente do conjunto de hipóteses sob o qual se desenvolve a teoria clássica do modelo linear (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Os estimadores de mínimos quadrados ordinários (MQO)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

em que

$$S_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \qquad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \qquad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

possuem propriedades importantes, que garantem sua validade para inferência estatística.

### 5.4.1 Não viés

Ambos os estimadores são não viesados:

$$E[\hat{\beta}_0] = \beta_0 \qquad \text{e} \qquad E[\hat{\beta}_1] = \beta_1.$$

Um estimador é dito **não viesado** quando sua esperança coincide com o parâmetro verdadeiro. No presente caso, os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são **centrados** em $\beta_0$ e $\beta_1$, respectivamente. Em termos frequentistas, isso significa que, sob repetição hipotética do processo amostral nas mesmas condições, a média das estimativas convergiria para os valores verdadeiros.

A demonstração formal desse resultado baseia-se na linearidade do operador esperança e na hipótese de exogeneidade fraca $E[\varepsilon_i \mid X_i] = 0$, e pode ser consultada no Apêndice de Demonstrações {#demo}. Conceitualmente, o ponto central é que, ao condicionar em $X$, o erro não contém componente sistemática capaz de deslocar, em média, os estimadores.

### 5.4.2 Variâncias e covariância dos estimadores

As variâncias dos estimadores são dadas por

$$Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2, \qquad Var(\hat{\beta}_1) = \frac{1}{S_{xx}}\sigma^2,$$

e a covariância entre eles é

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}}{S_{xx}}\sigma^2.$$

Essas expressões decorrem diretamente da representação linear dos estimadores em função dos $Y_i$ e das hipóteses sobre a estrutura de variância-covariância dos erros. A demonstração detalhada também pode ser vista no Apêndice de Demonstrações {#demo} (ver Kutner et al. (2005)).

Algumas interpretações conceituais importantes emergem dessas fórmulas:

1) **Influência da dispersão de** $X$: quanto maior $S_{xx}$, menor $Var(\hat{\beta}_1)$. Portanto, amostras com maior variabilidade em $X$ contêm mais informação sobre a inclinação da reta. Se os valores de $X$ estiverem muito concentrados, a estimativa da inclinação torna-se imprecisa.

2) **Dependência do intercepto em relação à origem**: a variância de $\hat{\beta}_0$ depende de $\bar{X}$. Quanto mais distante a média de $X$ estiver da origem, maior será a variância do intercepto, refletindo o fato de que $\hat{\beta}_0$ é obtido por extrapolação da reta até $X = 0$.

3) **Covariância negativa**: quando $\bar{X} > 0$, a covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$ é negativa. Isso indica que uma estimativa maior da inclinação tende a ser compensada por uma estimativa menor do intercepto, preservando a propriedade geométrica de que a reta ajustada passa por $(\bar{X}, \bar{Y})$.

Do ponto de vista geométrico, essas propriedades decorrem da ortogonalidade dos resíduos aos regressores, isto é,

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \qquad \text{e} \qquad \sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0.$$

Essas condições são equivalentes às equações normais e garantem que a projeção de $Y$ sobre o subespaço gerado por 1 e $X$ seja ortogonal ao vetor de resíduos. A conexão entre ortogonalidade e estrutura de variâncias é discutida em textos clássicos de regressão linear (ver Montgomery, Peck, e Vining (2021)).

### 5.4.3 Estimativa de $\sigma^2$ (graus de liberdade e não viés)

Definindo a **soma dos quadrados dos resíduos** como

$$SQRes = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

temos que

$$E[SQRes] = (n-2)\sigma^2.$$

A demonstração desse resultado utiliza a decomposição ortogonal do vetor $Y$ em componente ajustada e componente residual, podendo ser consultada no Apêndice de Demonstrações {#demo}.

Assim, o estimador

$$s^2 = \frac{SQRes}{n-2}$$

é **não viesado** para a variância dos erros:

$$E[s^2] = \sigma^2.$$

Os dois graus de liberdade subtraídos refletem a estimação dos dois parâmetros do modelo $(\beta_0, \beta_1)$. Essa correção garante que a variabilidade residual não seja subestimada pelo fato de termos ajustado uma reta aos dados.

Na prática, substitui-se $\sigma^2$ por $s^2$ nas expressões de $Var(\hat{\beta}_0)$ e $Var(\hat{\beta}_1)$, obtendo-se estimativas dos erros-padrão. Observe que até aqui **não foi necessária a suposição de normalidade**: as propriedades de não viés e as fórmulas de variância decorrem apenas das hipóteses de média zero, homoscedasticidade e ausência de correlação entre erros (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Portanto, os estimadores de MQO no MRLS apresentam um conjunto de propriedades fundamentais: são **não viesados**, possuem **variâncias explicitamente caracterizadas**, exibem **covariância estrutural negativa** entre intercepto e inclinação e permitem a construção de um **estimador não viesado de** $\sigma^2$ a partir dos resíduos.

Essas características asseguram a solidez probabilística do método sob hipóteses relativamente gerais e preparam o terreno para a próxima questão natural: dentro da classe dos estimadores lineares não viesados, seria possível obter variâncias menores? O **Teorema de Gauss–Markov** responde negativamente a essa pergunta, estabelecendo a eficiência relativa do MQO.

## 5.5 Teorema de Gauss–Markov

**!** Teorema (Gauss–Markov)

No modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

sob as hipóteses

$$E[\varepsilon_i \mid X_i] = 0, \qquad Var(\varepsilon_i \mid X_i) = \sigma^2, \qquad Cov(\varepsilon_i, \varepsilon_j \mid X_i, X_j) = 0 \ (\forall i \neq j),$$

os estimadores de mínimos quadrados ordinários são **lineares em** $Y$, **não viesados** e possuem **variância mínima** dentro da classe de todos os estimadores lineares não viesados dos parâmetros $\beta_0$ e $\beta_1$.

Em outras palavras, se restringirmos nossa atenção a estimadores que sejam combinações lineares das observações $Y_i$ e que sejam não viesados para os parâmetros verdadeiros, então nenhum outro estimador dessa classe terá variância menor que a dos estimadores de MQO. Essa é a essência do qualificativo *best*: não significa "melhor entre todos os estimadores possíveis", mas "melhor dentro da classe dos estimadores lineares não viesados".

Este teorema organiza três ideias fundamentais:

1) **Linearidade do estimador**: o estimador pode ser escrito como combinação linear das respostas observadas.

2) **Não viés**: sua esperança coincide com o parâmetro verdadeiro.

3) **Eficiência relativa**: entre todos os estimadores que satisfazem (1) e (2), o MQO apresenta a menor variância.

O resultado não depende da normalidade dos erros. Essa é uma distinção crucial: a normalidade é necessária apenas quando se deseja obter distribuições exatas finitas para estatísticas. A propriedade BLUE decorre exclusivamente da estrutura de média e variância do modelo linear clássico (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Geometricamente, o teorema está intimamente ligado à interpretação do MQO como projeção ortogonal do vetor $Y$ no subespaço gerado pelos regressores. A projeção ortogonal é, por construção, o vetor ajustado que minimiza a distância quadrática a $Y$. A minimização da distância quadrática no espaço amostral se traduz, no plano probabilístico, em minimização da variância entre estimadores lineares não viesados. Essa ponte entre geometria e probabilidade é um dos aspectos mais profundos do modelo linear.

É importante enfatizar também o alcance do resultado. O teorema não afirma que o MQO é o estimador de menor variância entre todos os estimadores imagináveis. Métodos não lineares ou estimadores viesados podem, em certos contextos, apresentar menor erro quadrático médio. O que o Teorema de Gauss–Markov garante é a **otimalidade dentro da classe linear não viesada**, uma classe ampla e natural no contexto da regressão.

A demonstração formal do teorema, baseada em argumentos de decomposição de variância e ortogonalidade, pode ser consultada no Apêndice de Demonstrações {#demo}.

Em termos práticos, o teorema fornece a base teórica que sustenta o uso do MQO como método padrão de estimação em regressão linear. Ele mostra que, sob hipóteses relativamente fracas e sem necessidade de normalidade, o procedimento adotado é eficiente dentro de uma classe ampla de estimadores. Essa combinação de simplicidade algébrica, interpretação geométrica clara e fundamentação probabilística sólida explica por que o MQO ocupa posição central na estatística aplicada e na econometria.

# 6 Inferência no MRLS com erros normais

## 6.1 Por que assumir normalidade?

Até aqui, estudamos as propriedades dos estimadores de mínimos quadrados ordinários (MQO) no Modelo de Regressão Linear Simples (MRLS). Mostramos que $\hat{\beta}_0$ e $\hat{\beta}_1$ são **não viesados**, possuem **variâncias explícitas** e, pelo **Teorema de Gauss–Markov**, são os **melhores estimadores lineares não viesados (BLUE)** sob as hipóteses clássicas de exogeneidade, homoscedasticidade e independência dos erros (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

No entanto, até este ponto conhecemos apenas **momentos de primeira e segunda ordem** das distribuições amostrais dos estimadores, ou seja, suas esperanças e variâncias. Não conhecemos suas **distribuições exatas**. De inferência, já sabemos que um estimador ser não viesado e eficiente dentro de uma classe não é suficiente para construir intervalos de confiança exatos ou realizar testes de hipóteses com nível de significância controlado em amostras finitas.

Para superar essa limitação, acrescentamos uma hipótese mais forte e específica: a **normalidade dos erros**,

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, n, \quad \text{independentes.}$$

Ou seja, cada erro segue uma **distribuição normal** com média zero e variância constante $\sigma^2 > 0$, sendo ainda independentes entre si.

### 6.1.1 Estrutura probabilística do MRLS com erros normais

Com essa suposição adicional, a formulação probabilística do modelo passa a ser

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Logo, condicionalmente a $X_i$, temos

$$Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i,\, \sigma^2).$$

Isso significa que o modelo deixa de ser apenas um modelo para a média condicional e passa a especificar completamente a **distribuição condicional de $Y$ dado $X$**. Em outras palavras, a normalidade fornece não apenas a forma do valor esperado, mas também a forma funcional da incerteza em torno dessa média.

### 6.1.2 Consequências conceituais da normalidade

A introdução da hipótese de normalidade tem implicações precisas:

- O **MRLS continua sendo um modelo para a média condicional**, mas agora a variabilidade em torno dessa média é descrita por uma estrutura probabilística completamente especificada.
- As hipóteses de Gauss–Markov já asseguravam que os estimadores de MQO eram BLUE, mas **não determinavam suas distribuições exatas**. A normalidade preenche exatamente essa lacuna.
- Como combinações lineares de variáveis normais são normais, os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, que são combinações lineares dos $Y_i$, passam a ter distribuições normais exatas em amostras finitas (ver Montgomery, Peck, e Vining (2021)).
- A estatística baseada na soma dos quadrados dos resíduos passa a ter distribuição qui-quadrado, o que permite derivar distribuições $t$ e $F$ de forma exata (ver Kutner et al. (2005)).

Do ponto de vista metodológico, a normalidade não é necessária para a obtenção das propriedades de não viés ou eficiência relativa, mas é uma boa alternativa para a construção de **procedimentos inferenciais exatos em amostras finitas**. Essa hipótese não altera os estimadores de MQO, mas altera o que podemos afirmar sobre sua variabilidade e sobre a incerteza associada às estimativas.

Portanto, a introdução da normalidade transforma o MRLS de um modelo com propriedades ótimas em termos de média e variância em um modelo com **estrutura probabilística completa**, apto a sustentar intervalos de confiança e testes de hipóteses com boas propriedades.

## 6.2 Distribuições amostrais no MRLS com erros normais

Sob a hipótese adicional de normalidade dos erros no MRLS, podemos derivar as **distribuições amostrais exatas** dos principais estimadores do modelo. Esse é o ponto de transição entre propriedades puramente algébricas (não viés, variância mínima dentro de uma classe) e **inferência estatística formal**.

Recordemos que, sob normalidade,

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{independentes.}$$

Como os estimadores de MQO podem ser escritos como **combinações lineares dos** $Y_i$, e cada $Y_i$ é normal condicionalmente a $X_i$, segue que $\hat{\beta}_0$ e $\hat{\beta}_1$ são também normalmente distribuídos. Essa conclusão decorre do fato fundamental de que combinações lineares de variáveis normais independentes permanecem normais (ver Montgomery, Peck, e Vining (2021)).

### 6.2.1 Distribuição de $\hat{\beta}_0$, $\hat{\beta}_1$

Para a inclinação, obtemos:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right), \qquad S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Para o intercepto:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\right).$$

A demonstração dessas distribuições pode ser vista no Apêndice de Demonstrações {#demo}, onde se explora explicitamente a representação linear dos estimadores em função dos $Y_i$ e a estrutura de variância-covariância do vetor de respostas.

Além disso, para o estimador da variância residual,

$$s^2 = \frac{SQRes}{n-2}, \qquad SQRes = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

vale o resultado fundamental:

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Esse resultado decorre da decomposição ortogonal do vetor $Y$ em componente ajustada e componente residual, cuja demonstração também pode ser consultada no Apêndice {#demo} (ver Kutner et al. (2005)). A perda de dois graus de liberdade reflete a estimação dos dois parâmetros $\beta_0$ e $\beta_1$.

É importante destacar que a normalidade **não altera os estimadores de MQO**: as expressões de $\hat{\beta}_0$, $\hat{\beta}_1$ e $s^2$ permanecem as mesmas. O ganho está em outro ponto: ela fornece uma descrição

probabilística completa da variabilidade desses estimadores, algo que as hipóteses de Gauss–Markov não entregam por si só.

Em particular, o resultado

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$$

é a peça-chave que permite obter, de forma exata, as distribuições $t$ e $F$ usadas em intervalos de confiança e testes de hipóteses. Os $n-2$ graus de liberdade refletem a estimação de $(\beta_0, \beta_1)$ e garantem que $s^2$ seja não viesado para $\sigma^2$.

## 6.3 Predição pontual da média condicional

A predição pontual é o primeiro passo para inferir sobre a relação média entre $Y$ e $X$ no MRLS. Antes de introduzir intervalos, é útil explicitar que os valores ajustados são quantidades aleatórias (pois dependem da amostra) e, sob normalidade, possuem distribuição conhecida.

### 6.3.1 Distribuição dos valores ajustados

Para um valor genérico $X_0$, definimos o valor ajustado (ou média condicional estimada) como

$$\hat{\mu}(X_0) \;=\; \hat{\beta}_0 + \hat{\beta}_1 X_0.$$

Sob erros normais, $\hat{\beta}_0$ e $\hat{\beta}_1$ são combinações lineares dos $Y_i$ e, portanto, $\hat{\mu}(X_0)$ também é uma combinação linear de variáveis normais. Assim, $\hat{\mu}(X_0)$ é normalmente distribuído (ver Montgomery, Peck, e Vining (2021)).

Além disso, sua esperança é

$$E[\hat{\mu}(X_0)] = \beta_0 + \beta_1 X_0 = \mu(X_0),$$

isto é, $\hat{\mu}(X_0)$ é **não viesado** para a média condicional.

Sua variância é

$$Var(\hat{\mu}(X_0)) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right], \qquad S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

de modo que

$$\hat{\mu}(X_0) \sim N\Bigg(\mu(X_0),\ \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right]\Bigg).$$

Essa expressão evidencia um aspecto estrutural da regressão: a incerteza sobre a média ajustada é **menor perto de** $\bar{X}$ e aumenta à medida que $X_0$ se afasta do centro dos dados, refletindo a geometria do ajuste por mínimos quadrados (ver Kutner et al. (2005)).

### 6.3.2 Predição pontual

A predição pontual da resposta média em $X_0$ é, portanto,

$$\hat{Y}_0 = \hat{\mu}(X_0) = \hat{\beta}_0 + \hat{\beta}_1 X_0,$$

que corresponde à função de regressão estimada avaliada em $X_0$. É importante enfatizar que $\hat{Y}_0$ se refere à **média condicional** $E[Y \mid X_0]$, e não ao valor de uma nova observação individual.

### 6.3.3 Limitação da predição pontual

Embora $\hat{Y}_0$ forneça uma estimativa central, ela não quantifica a incerteza associada ao ajuste. Por isso, em aplicações, a predição pontual deve ser acompanhada de:

- um **intervalo de confiança** para $\mu(X_0)$, quando o interesse é a tendência média; ou
- um **intervalo de predição**, quando o objetivo é prever uma nova observação individual.

Essas duas construções serão desenvolvidas nas próximas subseções (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Considere o seguinte gráfico, onde o ponto destacado corresponde a $\hat{Y}_0$, isto é, à estimativa da média condicional em $X_0$. Observe que o valor pontual não informa, por si só, o grau de incerteza associado à estimativa, questão que será tratada na próxima subseção.

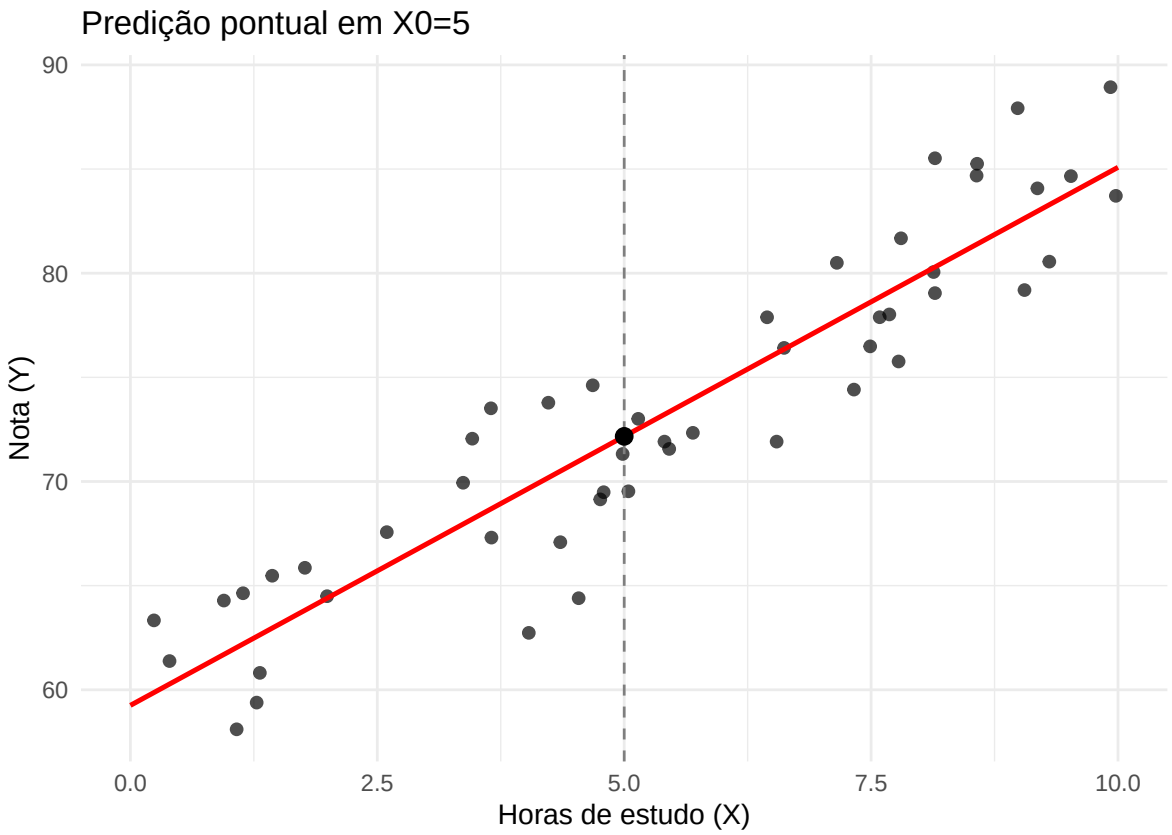



Figura 6.1: Predição pontual: reta ajustada e valor previsto para X0=5 horas.

## 6.4 Intervalos de confiança no MRLS

Na seção anterior vimos que, sob a hipótese de erros normais, os estimadores de MQO possuem distribuições normais quando a variância $\sigma^2$ é conhecida. No entanto, na prática, $\sigma^2$ é desconhecida e deve ser substituída por seu estimador não viesado

$$s^2 = \frac{SQRes}{n-2}, \qquad SQRes = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Essa substituição tem consequência direta na forma das distribuições amostrais: ao padronizarmos os estimadores utilizando $s$ em vez de $\sigma$, as estatísticas resultantes deixam de seguir a normal padrão e passam a seguir distribuições $t$ de Student com $n-2$ graus de liberdade (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Essa mudança decorre do fato de que

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2},$$

e de que $\hat{\beta}_0$ e $\hat{\beta}_1$ são independentes de $s^2$ sob normalidade dos erros, resultado cuja demonstração pode ser consultada no Apêndice de Demonstrações {#demo}.

### 6.4.1 Intervalos para $\beta_0$, $\beta_1$

Para a inclinação, a estatística

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}}, \qquad S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

segue distribuição t de Student com $n-2$ graus de liberdade

$$T_{\beta_1} \sim t_{n-2}.$$

Analogamente, para o intercepto, temos a mesma distribuição t de Student com $n-2$ graus de liberdade

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}}} \sim t_{n-2}.$$

A demonstração formal dessas distribuições padronizadas pode ser vista no Apêndice {#demo}, onde se utiliza a independência entre estimadores lineares normais e a soma de quadrados residual.

Com base nessas estatísticas, os intervalos de confiança de nível $(1-\alpha)\times 100\%$ são dados por

$$IC_{1-\alpha}(\beta_0) = \hat{\beta}_0 \pm t_{n-2;1-\alpha/2}\; s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}}$$

e

$$IC_{1-\alpha}(\beta_1) = \hat{\beta}_1 \pm t_{n-2;1-\alpha/2}\; \frac{s}{\sqrt{S_{xx}}}.$$

Aqui, $t_{n-2;1-\alpha/2}$ denota o quantil superior da distribuição $t$ com $n-2$ graus de liberdade.

### 6.4.2 Intervalo de confiança para a média condicional

Anteriormente vimos que a predição pontual da média condicional em $X_0$ é

$$\hat{\mu}(X_0) = \hat{\beta}_0 + \hat{\beta}_1 X_0,$$

estimativa natural de

$$\mu(X_0) = E[Y \mid X_0].$$

Sob a hipótese de erros normais, a combinação linear $\hat{\mu}(X_0)$ possui distribuição normal quando $\sigma^2$ é conhecido. Como, na prática, $\sigma^2$ é substituído por seu estimador não viesado $s^2 = SQ_{Res}/(n-2)$, a estatística padronizada

$$T = \frac{\hat{\mu}(X_0) - \mu(X_0)}{s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}}}$$

segue distribuição $t$ de Student com $n-2$ graus de liberdade (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)). A demonstração formal pode ser consultada no Apêndice de Demonstrações {#demo}.

Consequentemente, um intervalo de confiança de nível $(1-\alpha)\times 100\%$ para a média condicional é

$$IC_{1-\alpha}\left[\mu(X_0)\right] = \hat{\mu}(X_0) \pm t_{n-2;\,1-\alpha/2}\; s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}},$$

em que:

- $s^2 = SQ_{Res}/(n-2)$ é a variância residual estimada;
- $S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$;
- o termo dentro da raiz representa a **variabilidade da estimativa da média condicional**.

O fator

$$\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}$$

possui interpretação estrutural clara. Ele combina:

1) $\frac{1}{n}$: componente associado à incerteza na estimação do intercepto;

2) $\frac{(X_0-\bar{X})^2}{S_{xx}}$: componente associado à incerteza da inclinação e ao afastamento de $X_0$ em relação ao centro da amostra.

Essa decomposição revela que a precisão da estimativa depende da posição de $X_0$ no domínio observado. O intervalo é:

- **mais estreito** quando $X_0 = \bar{X}$;
- **mais largo** à medida que $X_0$ se afasta da média amostral.

Esse comportamento decorre diretamente da geometria da regressão linear e da estrutura de projeção ortogonal subjacente aos mínimos quadrados.

É fundamental enfatizar que esse intervalo refere-se à **média condicional**

$$\mu(X_0) = E[Y \mid X_0],$$

e não a uma nova observação individual. Ele quantifica a incerteza sobre a **tendência média da resposta** para a condição $X = X_0$.

**Dicas de uso**

- Utilize este intervalo quando o objetivo for inferir sobre a **tendência média** da resposta para um valor específico do regressor.
- Não o confunda com o intervalo de predição para um novo indivíduo, que incorpora variabilidade adicional do erro aleatório.

A figura a seguir ilustra um intervalo de confiança de 95% para a média condicional ao longo do domínio observado.

Observe qye banda em torno da reta representa a incerteza sobre $\mu(X)$ ao longo dos valores observados de $X$. Note que ela é mais estreita nas proximidades de $\bar{X}$ e se alarga progressivamente nos extremos do domínio amostral, refletindo o aumento da variância de $\hat{\mu}(X_0)$.

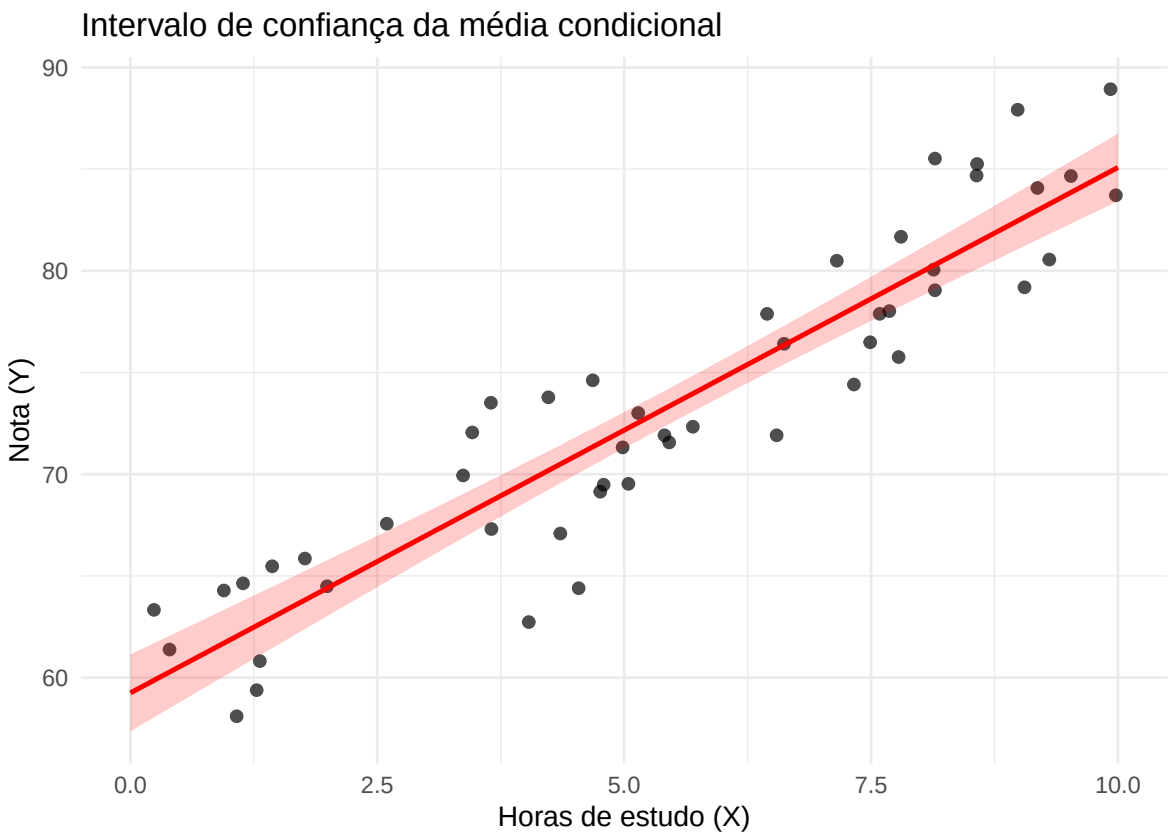



Figura 6.2: Intervalo de confiança (95%) para a média condicional.

### 6.4.3 Interpretação frequentista e significado dos intervalos

A interpretação frequentista de um intervalo de confiança de nível $1-\alpha$ é a seguinte: se o experimento fosse repetido um número grande de vezes e sob as mesmas condições, aproximadamente $(1-\alpha)\times 100\%$ dos intervalos construídos conteriam o verdadeiro parâmetro. O parâmetro é fixo; o que varia é o intervalo, pois ele depende da amostra observada.

A introdução da distribuição $t$ de Student desempenha papel essencial nesse contexto. Ao substituir $\sigma$ por seu estimador $s$, incorporamos a incerteza adicional decorrente da estimação da variância. Essa correção é particularmente relevante em amostras pequenas: quanto menor $n$, mais pesada é a cauda da distribuição $t_{n-2}$ e, consequentemente, mais largos são os intervalos. À medida que $n$ cresce, a distribuição $t_{n-2}$ converge para a normal padrão, e os intervalos passam a se aproximar daqueles que seriam obtidos se $\sigma^2$ fosse conhecido.

Os intervalos são centrados em $\hat{\beta}_0$ e $\hat{\beta}_1$ porque esses estimadores são não viesados. Em média, as retas ajustadas coincidem com a reta verdadeira; os intervalos quantificam precisamente a incerteza em torno dessa centralidade.

É fundamental distinguir os diferentes objetos inferenciais:

- O intervalo para $\beta_0$ e $\beta_1$ refere-se a **parâmetros estruturais do modelo**.
- O intervalo para $\mu(X_0) = E[Y \mid X_0]$ refere-se à **média condicional**.
- Nenhum desses intervalos corresponde à previsão de uma nova observação individual.

A distinção entre inferência sobre a média condicional e previsão individual é conceitualmente importante, pois a segunda incorpora não apenas a incerteza na estimação dos parâmetros, mas também a variabilidade intrínseca do processo aleatório. Essa diferença será aprofundada na subseção seguinte.

## 6.5 Intervalo de predição para nova observação

Até aqui construímos intervalos de confiança para a **média condicional** $\mu(X_0) = E[Y \mid X_0]$. No entanto, muitas aplicações exigem algo diferente: prever o valor de uma **nova observação individual** associada a $X_0$.

Se uma nova unidade experimental for observada no mesmo valor $X_0$, seu modelo é

$$Y_{\text{novo}}(X_0) = \beta_0 + \beta_1 X_0 + \varepsilon_{\text{novo}},$$

em que $\varepsilon_{\text{novo}} \sim N(0, \sigma^2)$ e é independente dos erros da amostra original.

A diferença conceitual é que agora não estamos estimando apenas a média condicional, mas prevendo uma realização específica que contém, além da incerteza na estimação dos parâmetros, a **variabilidade intrínseca do erro aleatório**.

Portanto, a quantidade relevante é

$$Y_{\text{novo}}(X_0) - \hat{\mu}(X_0),$$

cuja variância é

$$Var[Y_{\text{novo}}(X_0) - \hat{\mu}(X_0)] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right].$$

O termo adicional "+1" aparece porque a nova observação contém um erro próprio, independente daquele utilizado na estimação dos parâmetros. A demonstração formal dessa variância pode ser consultada no Apêndice de Demonstrações {#demo} (ver Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Padronizando essa quantidade por $s$, obtemos uma estatística que segue distribuição t de Student com $n-2$ graus de liberdade ($t_{n-2}$) sob normalidade dos erros. Assim, o intervalo de predição de nível $(1-\alpha)\times 100\%$ é

$$IC[Y_{\text{novo}}(X_0)] = \hat{\mu}(X_0) \pm t_{n-2;\,1-\alpha/2}\; s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{xx}}}.$$

Comparando com o intervalo para a média condicional, temos

$$\hat{\mu}(X_0) \pm t_{n-2;\,1-\alpha/2}\; s\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{xx}}}.$$

Daí, vemos claramente a presença do termo adicional 1 dentro da raiz. Esse termo representa a **variabilidade individual irreducível** do processo aleatório.

Consequentemente:

- O intervalo de predição é **sempre mais largo** que o intervalo de confiança da média.
- A diferença entre eles é tanto maior quanto maior for $\sigma^2$.
- Ambos se alargam quando $X_0$ se afasta de $\bar{X}$, refletindo a incerteza adicional associada à extrapolação.

O intervalo de predição fornece um conjunto de valores plausíveis para uma **nova observação individual**, e não para a média populacional. Em termos frequentistas, se o processo fosse repetido nas mesmas condições, aproximadamente $(1-\alpha)\times 100\%$ desses intervalos conteriam a nova observação gerada pelo modelo.

**Dica prática**

- Use **intervalo de confiança** quando o objetivo for inferir sobre a **tendência média**.
- Use **intervalo de predição** quando o interesse for antecipar o valor de um **novo indivíduo**.

A figura a seguir ilustra simultaneamente o intervalo de confiança para a média e o intervalo de predição para nova observação. Comparando as bandas, perceba que o intervalo de predição é sempre mais largo que o de confiança. Isso ocorre porque ele incorpora a variabilidade individual das novas observações, além da incerteza da média.

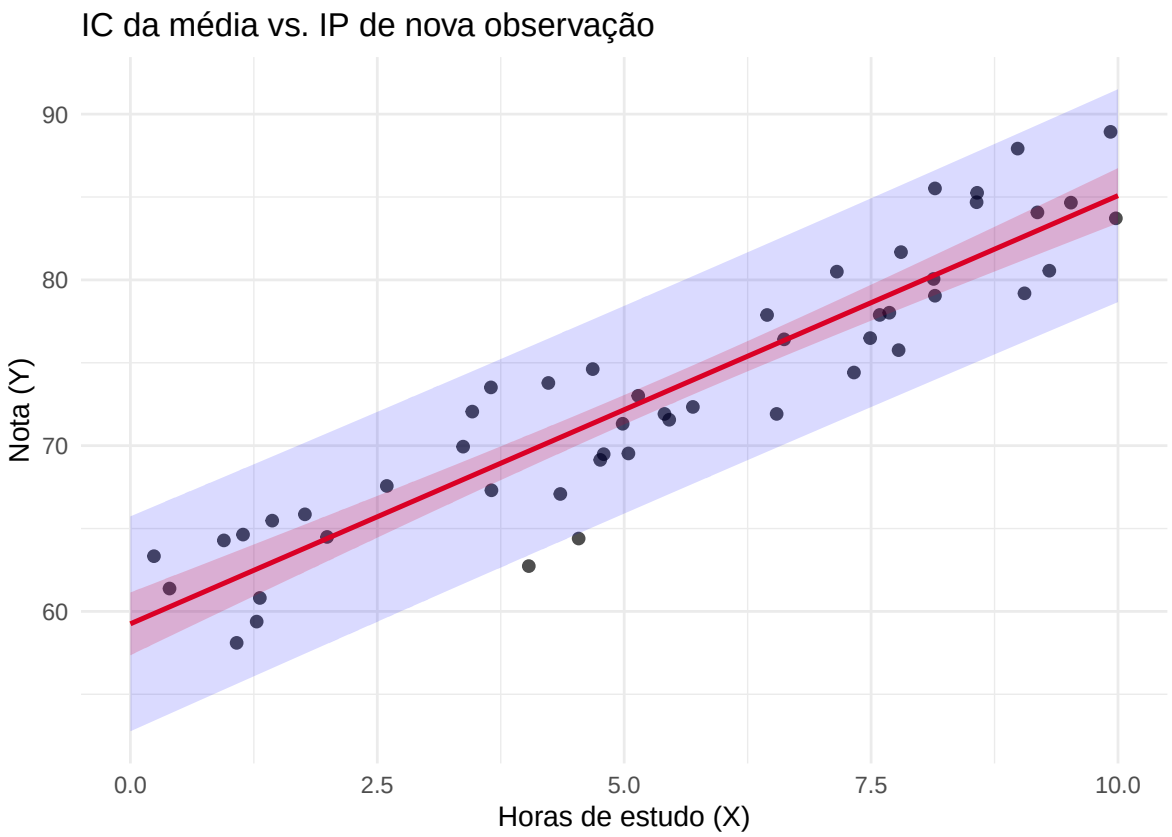



Figura 6.3: Intervalo de confiança vs. intervalo de predição (95%).

**Advertência sobre extrapolação**

As expressões obtidas para o intervalo de predição são válidas para qualquer valor numérico $X_0$. No entanto, é fundamental distinguir entre **interpolação** (quando $X_0$ pertence ao intervalo observado da amostra) e **extrapolação** (quando $X_0$ está fora do domínio observado de $X$).

Seja o intervalo amostral observado

$$X_{(min)} \leq X_i \leq X_{(max)}.$$

Quando $X_0 \in [X_{(min)}, X_{(max)}]$, o modelo está sendo utilizado em uma região sustentada pelos dados. Nesse caso, embora o intervalo possa se alargar à medida que $X_0$ se afasta de $\bar{X}$, a inferência permanece ancorada na informação empírica disponível.

Por outro lado, quando $X_0$ está fora desse intervalo, ocorre **extrapolação**. Nessa situação, a validade formal da expressão algébrica do intervalo permanece, mas sua confiabilidade prática pode ser comprometida, pois o modelo passa a depender fortemente da suposição de

linearidade fora da região observada. Pequenas violações da forma funcional podem produzir erros substanciais de predição.

Como destacam Kutner et al. (2005) e Montgomery, Peck, e Vining (2021), a regressão linear deve ser utilizada com cautela fora do domínio amostral, pois o comportamento da relação entre $X$ e $Y$ além dos dados observados não é garantido pelo modelo ajustado. Assim, intervalos de predição em extrapolação tendem a ser não apenas mais largos, mas também potencialmente menos representativos do processo real.

# 7 Testes de hipóteses e ANOVA

A construção de intervalos de confiança fornece uma forma de expressar a incerteza nos estimadores. Outra abordagem complementar é a dos **testes de hipóteses** e **análise de variância (ANOVA)**, que avaliam formalmente se os coeficientes do modelo diferem significativamente de determinados valores, em especial do zero. No Modelo de Regressão Linear Simples (MRLS) com erros normais, os testes se apoiam nas distribuições exatas dos estimadores discutidas anteriormente.

Sob as hipóteses clássicas do MRLS normal: (a) linearidade na forma funcional, (b) independência dos erros, (b) homocedasticidade e (d) normalidade, os estimadores de mínimos quadrados coincidem com os estimadores de máxima verossimilhança, e possuem distribuições amostrais exatas baseadas na distribuição normal e na distribuição $t$ de Student (Hoffmann (2016); Montgomery, Peck, e Vining (2021)). Em particular, condicionalmente aos valores observados de $X$, temos

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{S_{xx}}\right),$$

e, como $\sigma^2$ é desconhecida e estimada por $s^2 = SQ_{Res}/(n-2)$, a padronização conduz à distribuição $t_{n-2}$, resultado central para a inferência clássica em regressão linear simples (Kutner et al. (2005); Casella e Berger (2002)).

É importante enfatizar que os testes de hipóteses são construídos *dentro do modelo*. A validade exata da distribuição $t$ depende da normalidade dos erros; na ausência dessa hipótese, os resultados passam a ter caráter assintótico. Assim, significância estatística deve sempre ser interpretada à luz das suposições estruturais do modelo.

## 7.1 Testes marginais para $\beta_1$ e $\beta_0$

### 7.1.1 Teste para a $\beta_1$

No caso da inclinação, o teste é:

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_{1,0},$$

com estatística de teste dada por

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{s/\sqrt{S_{xx}}} \sim t_{n-2}.$$

Esse teste é especialmente relevante quando $\beta_{1,0} = 0$, situação em que verificamos se existe associação linear entre $X$ e $Y$. Em termos práticos, ele responde à pergunta: *vale a pena incluir* $X$ para explicar $Y$? Se $|T|$ ultrapassa o valor crítico da distribuição $t_{n-2}$, rejeitamos $H_0$ e concluímos que a inclinação é estatisticamente diferente de zero.

Do ponto de vista conceitual, testar $\beta_1 = 0$ equivale a testar se a melhor reta ajustada possui inclinação nula, isto é, se o modelo reduz-se a $Y_i = \beta_0 + \varepsilon_i$. Portanto, o teste compara dois modelos aninhados: o modelo completo (com inclinação livre) e o modelo restrito (com $\beta_1 = 0$). Essa interpretação como comparação entre modelos é fundamental para compreender a ligação posterior com a estatística $F$ (Draper e Smith (1998); Kutner et al. (2005)).

Além disso, há equivalência formal entre o teste $t$ bilateral ao nível $\alpha$ e o intervalo de confiança $(1-\alpha)$ para $\beta_1$: rejeitar $H_0$ é equivalente a verificar que $\beta_{1,0}$ não pertence ao intervalo de confiança correspondente Casella e Berger (2002). O $p$-valor, por sua vez, é definido como

$$p = P(|T| \geq |t_{obs}| \mid H_0),$$

e quantifica evidência contra $H_0$ dentro da estrutura probabilística assumida.

### 7.1.2 Teste para a $\beta_0$

De modo análogo, para o intercepto temos:

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{vs} \quad H_1 : \beta_0 \neq \beta_{0,0},$$

com estatística de teste

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}}} \sim t_{n-2}.$$

Esse teste é menos central do ponto de vista prático, mas pode ser importante quando se deseja avaliar se o valor médio de $Y$ para $X = 0$ coincide com alguma referência teórica ou prática.

Conceitualmente, $\beta_0 = E(Y \mid X = 0)$ dentro do modelo linear. Assim, sua interpretação depende criticamente de $X = 0$ ter significado no fenômeno estudado e estar dentro do intervalo de observação dos dados. Caso contrário, o intercepto pode representar apenas uma

extrapolação matemática da reta ajustada, ainda que perfeitamente bem definido do ponto de vista estatístico (Montgomery, Peck, e Vining (2021); Weisberg (2005)).

> **Apêndice de Demonstrações {#demo}:** as distribuições exatas das estatísticas $T$ decorrem da normalidade dos erros, da independência entre $\hat{\beta}_j$ e $SQ_{Res}$ e da relação entre variância residual e distribuição qui-quadrado, conforme desenvolvimento clássico da inferência em modelos lineares (Casella e Berger (2002); Kutner et al. (2005)).

## 7.2 Análise de Variância no MRLS

### 7.2.1 Decomposição da soma de quadrados

O teste $F$ pode ser entendido a partir da decomposição da variabilidade total em $Y$:

$$SQ_{Total} = SQ_{Reg} + SQ_{Res}.$$

Aqui, $SQ_{Total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ mede a variabilidade total das observações em torno da média. Essa variabilidade pode ser separada em duas partes:

- **variabilidade explicada pela regressão**

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- **Variabilidade não explicada**

$$SQ_{Res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

associada aos resíduos. Em termos geométricos, a $SQ_{Reg}$ corresponde à projeção de $Y$ no espaço gerado por $X$, enquanto $SQ_{Res}$ corresponde ao componente ortogonal (erro).

Formalmente, essa decomposição decorre da ortogonalidade entre resíduos e valores ajustados no método dos mínimos quadrados. No MRLS, tem-se

$$\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{Y}_i - \bar{Y}) = 0,$$

o que implica que a variabilidade total pode ser particionada sem termo de cruzamento. Em notação matricial, essa propriedade está associada ao fato de que o vetor de resíduos é ortogonal ao espaço coluna da matriz de projeto $\mathbf{X}$, isto é, $\mathbf{X}^\top\hat{\varepsilon} = \mathbf{0}$ (Harville (2000); Searle (2016)).

**Apêndice de Demonstrações {#demo}:** a identidade $SQ_{Total} = SQ_{Reg} + SQ_{Res}$ é obtida expandindo $\sum(Y_i - \bar{Y})^2$ em termos de $(\hat{Y}_i - \bar{Y})$ e $(Y_i - \hat{Y}_i)$ e utilizando a ortogonalidade dos resíduos.

Essa decomposição mostra que o ajuste por regressão não apenas fornece estimativas pontuais, mas também permite quantificar de forma clara quanto da variabilidade total de $Y$ é capturada pela relação linear com $X$. Quanto maior $SQ_{Reg}$ em relação a $SQ_{Total}$, maior o poder explicativo do modelo.

Do ponto de vista probabilístico, sob $H_0 : \beta_1 = 0$ e normalidade dos erros, as somas de quadrados associadas à regressão e aos resíduos, quando devidamente padronizadas por $\sigma^2$, seguem distribuições qui-quadrado independentes com 1 e $n-2$ graus de liberdade, respectivamente (Kutner et al. (2005)). Essa independência é a base formal da estatística $F$.

### 7.2.2 Quadrados Médios e Estatística F

Para formalizar o teste global, cada soma de quadrados é dividida pelos graus de liberdade correspondentes:

- Quadrado médio da regressão:

$$QM_{Reg} = \frac{SQ_{Reg}}{1}.$$

- Quadrado médio dos resíduos:

$$QM_{Res} = \frac{SQ_{Res}}{n-2}.$$

A razão entre eles define a estatística $F$:

$$F = \frac{QM_{Reg}}{QM_{Res}} \sim F_{1,n-2} \quad \text{sob } H_0 : \beta_1 = 0.$$

Esse teste avalia, portanto, se a proporção de variabilidade explicada pela regressão é grande o suficiente em comparação com a variabilidade residual, justificando o uso do modelo.

Interpretativamente, $QM_{Res}$ é um estimador não viesado de $\sigma^2$, enquanto $QM_{Reg}$ mede a variação explicada *por grau de liberdade associado ao efeito linear de X*. Assim, o teste $F$ compara um componente sistemático (sinal) com um componente aleatório (ruído).

Valores elevados de $F$ indicam que a redução em $SQ_{Res}$ ao incluir $X$ é grande demais para ser atribuída apenas ao acaso (Montgomery, Peck, e Vining (2021); Weisberg (2005)).

### 7.2.3 Tabela ANOVA do MRLS

| Fonte de variação | SQ | GL | QM | Estatística |
|---|---|---|---|---|
| Regressão | $SQ_{Reg}$ | 1 | $QM_{Reg}$ | $F = QM_{Reg}/QM_{Res}$ |
| Resíduo | $SQ_{Res}$ | $n-2$ | $QM_{Res}$ | |
| Total | $SQ_{Total}$ | $n-1$ | | |

A tabela resume de maneira padronizada a decomposição da variabilidade e fornece a base para a aplicação do teste $F$. Note que os graus de liberdade totais satisfazem

$$(n-1) = 1 + (n-2),$$

refletindo que dois parâmetros foram estimados no modelo (intercepto e inclinação).

### 7.2.4 Coeficiente de determinação $R^2$

Um desdobramento natural dessa análise é o **coeficiente de determinação**:

$$R^2 = \frac{SQ_{Reg}}{SQ_{Total}} = 1 - \frac{SQ_{Res}}{SQ_{Total}}.$$

Ele mede a proporção da variabilidade total de $Y$ explicada pela regressão e está limitado ao intervalo $0 \leq R^2 \leq 1$. Em termos práticos, $R^2 \times 100\%$ indica o percentual da variabilidade de $Y$ que é explicado linearmente por $X$. Quanto maior $R^2$, maior o poder explicativo do modelo.

É importante compreender que $R^2$ é uma medida descritiva da qualidade de ajuste dentro da amostra observada. Ele **não implica causalidade, nem garante desempenho preditivo**. Além disso, em modelos com múltiplos preditores, $R^2$ tende a aumentar com a inclusão de variáveis, mesmo que irrelevantes, motivo pelo qual se introduz posteriormente o $R^2$ ajustado (Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

No caso do MRLS, há uma relação direta com a estatística descritiva da correlação linear:

$$R^2 = r_{XY}^2, \quad r_{XY} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

em que $r_{XY}$ é o coeficiente de correlação amostral entre $X$ e $Y$. Essa igualdade decorre das expressões algébricas do estimador $\hat{\beta}_1$ e das somas de quadrados no caso univariado (Charnet et al. (2008); Hoffmann (2016)).

Assim, $R^2$ conecta três dimensões: é a proporção de variabilidade explicada (ANOVA), é equivalente ao quadrado da correlação linear (estatística descritiva) e fundamenta a estatística $F$ da ANOVA por meio da relação

$$F = \frac{QM_{Reg}}{QM_{Res}} = \frac{R^2}{1 - R^2}(n - 2).$$

> **Apêndice de Demonstrações {#demo}:** a relação entre $F$ e $R^2$ é obtida substituindo as definições de $SQ_{Reg}$ e $SQ_{Res}$ na razão $QM_{Reg}/QM_{Res}$ e utilizando a identidade $R^2 = SQ_{Reg}/SQ_{Total}$.

Em resumo, os testes de hipóteses no MRLS permitem verificar tanto a significância individual dos coeficientes quanto o poder explicativo global do modelo. A equivalência entre os testes $t$ e $F$, a decomposição da soma de quadrados e a interpretação do $R^2$ reforçam a visão integrada da regressão como técnica que conecta **estimação, inferência e análise da variabilidade** em um único arcabouço teórico (Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

### 7.2.5 Equivalência entre $T^2$ e $F$

Um resultado fundamental é que, no MRLS normal, o teste $t$ para a inclinação e o teste $F$ global da regressão são equivalentes. Quando a hipótese nula é $\beta_1 = 0$, temos:

$$F = T^2.$$

Isso ocorre porque o modelo possui apenas um preditor. Em modelos múltiplos, a situação muda: o teste $t$ continua avaliando parâmetros individuais, enquanto o teste $F$ passa a ter papel central ao considerar hipóteses conjuntas sobre vários coeficientes.

Para compreender essa equivalência com rigor, observe que, sob $H_0 : \beta_1 = 0$, a estatística $t$ pode ser escrita como

$$T = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}},$$

de modo que

$$T^2 = \frac{\hat{\beta}_1^2 S_{xx}}{s^2}.$$

Por outro lado, no MRLS, pode-se mostrar que

$$SQ_{Reg} = \hat{\beta}_1^2 S_{xx},$$

e que

$$QM_{Res} = s^2.$$

Logo,

$$F = \frac{QM_{Reg}}{QM_{Res}} = \frac{SQ_{Reg}/1}{s^2} = \frac{\hat{\beta}_1^2 S_{xx}}{s^2} = T^2.$$

Portanto, a estatística $F$ nada mais é do que o quadrado da estatística $t$ quando há apenas um parâmetro de inclinação sendo testado.

Do ponto de vista distribucional, se

$$T \sim t_{n-2},$$

então

$$T^2 \sim F_{1,n-2},$$

o que decorre da relação geral entre as distribuições $t$ e $F$ (Casella e Berger (2002)). Assim, a equivalência também se verifica ao nível das distribuições.

Essa identidade tem uma consequência didática importante: no MRLS, o teste global da regressão e o teste individual da inclinação são exatamente o mesmo teste, apenas expressos em escalas diferentes. Em outras palavras, testar a significância global do modelo é o mesmo que testar se a inclinação é nula.

> **Apêndice de Demonstrações {#demo}:** a identidade $SQ_{Reg} = \hat{\beta}_1^2 S_{xx}$ e a relação distribucional entre $t$ e $F$ podem ser demonstradas a partir das propriedades do MQO e da definição da distribuição $F$ como razão de qui-quadrados independentes Casella e Berger (2002); Harville (2000).

# 8 Diagnóstico e Avaliação no MRLS

## 8.1 Por que analisar resíduos?

Após o ajuste de um modelo de regressão, é essencial verificar se as **hipóteses do MRLS** do modelos para os erros aleatórios foram atendidas. Essa verificação se dá por diversos meios, sendo algumas dela via a análise dos **resíduos**.

Os resíduos mais intuitivos são definidos como:

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, 2, ..., n.$$

Estes resíduos representam a parte de $Y$ que **não foi explicada pelo modelo**. Enquanto os erros verdadeiros $\varepsilon_i$ são inobserváveis, os resíduos são acessíveis e servem como suas aproximações.

Um ponto conceitual importante é distinguir "hipóteses sobre os erros" de "propriedades dos resíduos". As hipóteses clássicas do MRLS são formuladas para os **erros aleatórios** $\varepsilon_i$ (componentes não observáveis do mecanismo gerador de dados). Já os resíduos $e_i$ são funções dos dados e dos estimadores, logo carregam restrições algébricas impostas pelo MQO. Assim, mesmo que o MRLS seja verdadeiro (isto é, as hipóteses sobre $\varepsilon_i$ sejam satisfeitas), os resíduos **não** se comportam como uma amostra i.i.d. de uma mesma distribuição; em particular, eles são correlacionados e apresentam variâncias diferentes ao longo de $i$ a dependendo da alavancagem (Searle (2016); Harville (2000)).

As principais hipóteses do modelo para os erros ($\varepsilon$) do MRLS são:

- Média zero ($E[\varepsilon_i] = 0$)
- Variância constante ($Var[\varepsilon_i] = \sigma^2$)
- Não correlação entre os erros ($cov[\varepsilon_i, \varepsilon_j] = 0, \forall i \neq j$)

Podem ser feitas hipóteses adicionais sobre a forma da distribuição dos erros, como assumir certa assimetria, curtose específica ou até uma distribuição conhecida.

A suposição (hipótese) distribuição mais considerada para a distribuição dos erros é:

- Normalidade ($\varepsilon_i \sim N(0, \sigma^2)$).

Um modelo só pode ser considerado adequado se os resíduos se comportarem como erros aleatórios: sem tendência sistemática, com variância aproximadamente constante, não correlacionaos e, em muitos contextos, aproximadamente normais. Em prática aplicada, é útil interpretar isso como: **(i)** a média condicional foi bem especificada (linearidade na forma funcional), **(ii)** a variância condicional não muda de forma sistemática (homocedasticidade) e **(iii)** não há estrutura temporal/espacial remanescente (independência), além de **(iv)** normalidade como hipótese adicional que viabiliza inferência exata e diagnósticos probabilísticos baseados em caudas (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

## 8.2 Tipos de resíduos e propriedades

### 8.2.1 Resíduos ordinários

O ponto de partida são os **resíduos ordinários**:

$$e_i = Y_i - \hat{Y}_i.$$

Eles indicam o desvio direto entre a observação e a reta ajustada. Por exemplo, $e_i > 0$ mostra que o modelo **subestimou** $Y_i$, enquanto $e_i < 0$ mostra que o modelo **superestimou**.

Do ponto de vista conceitual, o resíduo é uma *estimativa observável* do erro aleatório $\varepsilon_i$. Como $\varepsilon_i$ não é observável, toda a etapa de diagnóstico repousa sobre a análise do comportamento dos $e_i$. Entretanto, é fundamental compreender que resíduos **não são** os erros verdadeiros: eles dependem dos parâmetros estimados e, portanto, carregam estrutura imposta pelo método de mínimos quadrados Hoffmann (2016); Montgomery, Peck, e Vining (2021).

#### 8.2.1.1 Propriedades básicas dos resíduos ordinários

O método dos mínimos quadrados impõe três propriedades estruturais:

1. **Soma nula**

$$\sum_{i=1}^{n} e_i = 0.$$

   A reta ajustada sempre passa pelo ponto médio amostral $(\bar{X}, \bar{Y})$.

2. **Ortogonalidade com o preditor**

$$\sum_{i=1}^{n} e_i X_i = 0.$$

   Não há associação linear entre os resíduos e a variável explicativa. Caso existisse, o modelo poderia ser melhorado ajustando novamente a inclinação.

3. **Soma de quadrados dos resíduos**

$$\sum_{i=1}^{n} e_i^2 = SQ_{Res},$$

isto é, os resíduos concentram exatamente a variabilidade não explicada pelo modelo.

Essas propriedades decorrem diretamente das **equações normais do método dos mínimos quadrados** no caso univariado, obtidas pela minimização de $\sum(Y_i - \beta_0 - \beta_1 X_i)^2$ em relação a $\beta_0$ e $\beta_1$ (Charnet et al. (2008); Kutner et al. (2005)).

> **Apêndice de Demonstrações {#demo}:** as propriedades acima são obtidas substituindo $\widehat{\beta}_0$ e $\widehat{\beta}_1$ nas expressões dos resíduos e manipulando os somatórios resultantes das equações normais.

Essas três propriedades têm implicações importantes: mesmo que os erros verdadeiros sejam independentes e homocedásticos, os resíduos não são independentes entre si e tampouco possuem variância constante.

#### 8.2.1.2 Esperança, variância, covariância e distribuição dos resíduos ordinários

1. **Esperança**

$$E[e_i] = 0.$$

Sob as hipóteses do MRLS, cada resíduo tem média zero. Isso significa que, em termos probabilísticos, o modelo não superestima nem subestima sistematicamente a resposta.

2. **Variância**

$$Var(e_i) = \sigma^2(1 - h_{ii}),$$

em que

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{X})^2}{S_{xx}}.$$

A quantidade $h_{ii}$ é chamada de **alavancagem** da observação $i$. Ela mede o quanto o valor de $X_i$ influencia o próprio ajuste $\widehat{Y}_i$.

Observações com valores de $X_i$ muito afastados da média $\bar{X}$ apresentam maior alavancagem. Como consequência, possuem menor variância residual, pois "ancoram" a reta ajustada com maior intensidade (Belsley, Kuh, e Welsch (1980); Montgomery, Peck, e Vining (2021)).

3. **Covariância**

$$Cov(e_i, e_j) = -\sigma^2 h_{ij}, \quad i \neq j,$$

em que

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{X})(x_j - \bar{X})}{S_{xx}}.$$

Portanto, os resíduos são **correlacionados entre si**. Isso é consequência direta do fato de que todos os resíduos dependem dos mesmos estimadores $\widehat{\beta}_0$ e $\widehat{\beta}_1$ e, consequentemente, os resíduos não podem ser tratados como uma nova amostra independente de erros aleatórios (Kutner et al. (2005); Weisberg (2005)).

4. **Distribuição**

Se assumimos normalidade para os erros aleatórios,

$$\varepsilon_i \sim N(0, \sigma^2),$$

então os resíduos ordinários também seguem distribuição normal, pois são combinações lineares das variáveis $\varepsilon_i$:

$$e_i \sim N(0, \sigma^2(1 - h_{ii})).$$

Essa normalidade é exata sob a hipótese de erros normais. Caso a normalidade não seja assumida, a distribuição dos resíduos pode ser aproximada por resultados assintóticos.

> **Apêndice de Demonstrações {#demo}:** as expressões de variância e covariância dos resíduos são obtidas substituindo $e_i = Y_i - \widehat{Y}_i$ e utilizando as propriedades das variâncias de combinações lineares, juntamente com as expressões explícitas de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ Kutner et al. (2005); Montgomery, Peck, e Vining (2021).

#### 8.2.1.3 Implicações para diagnóstico

Esses resultados mostram que, mesmo quando as hipóteses usuais de média zero, variância constante, não correlação e normalidade para os erros aleatórios são satisfeitas, os resíduos ordinários apresentam:

- variância não constante (dependente de $h_{ii}$),
- correlação entre si,
- dependência dos parâmetros estimados.

Portanto, embora úteis para visualização inicial e interpretação direta do ajuste, os resíduos ordinários não são ideais para comparações diretas entre observações com diferentes níveis de alavancagem.

Essa limitação motiva a construção de resíduos transformados, como os **resíduos padronizados** e os **resíduos estudentizados**, que ajustam explicitamente a variabilidade individual e permitem diagnósticos mais adequados de pontos discrepantes e violações das hipóteses do modelo (Belsley, Kuh, e Welsch (1980); Weisberg (2005)).

### 8.2.2 Resíduos padronizados

Com o objetivo de tornar os resíduos **comparáveis entre si**, ajustando a diferença de variâncias individuais, definem-se os **resíduos padronizados** como

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \quad \text{com} \quad s^2 = \frac{SQ_{Res}}{n-2}.$$

Aqui, $e_i$ é o resíduo ordinário, $h_{ii}$ é a alavancagem da observação $i$ e $s^2$ é o estimador não viesado de $\sigma^2$. A ideia central é simples: como

$$Var(e_i) = \sigma^2(1-h_{ii}),$$

dividir $e_i$ por uma estimativa de seu desvio-padrão elimina a heterogeneidade de variâncias e produz uma quantidade adimensional.

Do ponto de vista conceitual, essa padronização desempenha papel análogo ao de uma estatística $z$: ela mede o "tamanho" do desvio em unidades de desvio-padrão estimado.

#### 8.2.2.1 Propriedades fundamentais

Sob as hipóteses do MRLS:

1. **Esperança aproximada**
$$E[r_i] \approx 0.$$

2. **Variância aproximada**
$$Var(r_i) \approx 1.$$

A aproximação decorre do fato de que $s^2$ é uma estimativa de $\sigma^2$. Se $\sigma^2$ fosse conhecido, teríamos exatamente

$$\frac{e_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1),$$

sob normalidade dos erros.

Entretanto, como $\sigma^2$ é substituído por $s^2$, a estatística passa a envolver uma razão entre variáveis aleatórias dependentes.

#### 8.2.2.2 Distribuição dos resíduos padronizados

Se os erros seguem

$$\varepsilon_i \sim N(0, \sigma^2),$$

então, para amostras moderadas ou grandes, vale a aproximação:

$$r_i \approx t_{n-2}.$$

A aproximação não é exata porque $e_i$ e $s^2$ não são independentes: ambos dependem das mesmas observações e dos mesmos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ (Weisberg (2005)).

Em amostras grandes, pela consistência de $s^2$ para $\sigma^2$, a distribuição de $r_i$ aproxima-se da normal padrão:

$$r_i \overset{aprox}{\sim} N(0,1).$$

> **Apêndice de Demonstrações {#demo}:** a aproximação $r_i \approx t_{n-2}$ decorre da substituição de $\sigma^2$ por $s^2$ na padronização e do fato de que $(n-2)s^2/\sigma^2 \sim \chi^2_{n-2}$ sob normalidade dos erros.

#### 8.2.2.3 Interpretação prática

Os resíduos padronizados permitem comparar observações com diferentes alavancagens. Um mesmo valor absoluto de resíduo ordinário pode ser pequeno ou grande dependendo de $h_{ii}$. A padronização corrige esse efeito.

Uma regra prática frequentemente utilizada é:

- $|r_i| > 2 \rightarrow$ possível observação discrepante.
- $|r_i| > 3 \rightarrow$ forte indício de discrepância.

Esses limiares baseiam-se na probabilidade de observar valores extremos sob uma distribuição aproximadamente normal ou $t$. Por exemplo, sob normalidade, a probabilidade de $|Z| > 2$ é aproximadamente 5%.

Contudo, essa interpretação deve ser feita com cautela:

- Em amostras grandes, é esperado que alguns valores ultrapassem 2 apenas por variabilidade natural.
- Em amostras pequenas, a aproximação pode ser imprecisa.
- A presença de múltiplos testes simultâneos pode inflar a taxa de falsos positivos.

#### 8.2.2.4 Limitações conceituais

Apesar de mais informativos que os resíduos ordinários, os resíduos padronizados ainda apresentam uma limitação importante: o denominador $s$ é calculado utilizando **todas as observações**, inclusive a própria observação $i$.

Assim, um ponto extremo pode inflar $s$, reduzindo artificialmente seu próprio resíduo padronizado, fenômeno conhecido como *masking* (mascaramento) (Belsley, Kuh, e Welsch (1980)).

Essa limitação motiva a definição dos **resíduos estudentizados externos**, nos quais a variância é estimada excluindo-se a própria observação sob análise.

Em síntese:

- **Resíduos ordinários** medem o erro bruto.
- **Resíduos padronizados** tornam os erros comparáveis.
- A padronização é essencial para diagnóstico formal de outliers e para construção de gráficos de resíduos mais informativos.

Nos próximos tópicos, veremos como a estudentização externa corrige a dependência entre numerador e denominador e fornece uma estatística com distribuição $t$ exata sob as hipóteses do modelo.

### 8.2.3 Resíduos estudentizados (externos)

Os **resíduos estudentizados externos** (também chamados de *externally studentized residuals* ou *deleted residuals*) foram propostos no contexto de diagnóstico de regressão para contornar a dependência entre numerador e denominador presente nos resíduos padronizados (Belsley, Kuh, e Welsch (1980); Weisberg (2005)).

Eles são definidos por

$$t_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}},$$

em que:

- $e_i$ é o resíduo ordinário da observação $i$;
- $h_{ii}$ é a alavancagem da observação $i$;
- $s^2_{(i)}$ é o estimador da variância do erro calculado **excluindo a i-ésima observação**.

Isto é, $s^2_{(i)}$ é obtido ajustando o modelo com $n-1$ observações, removendo o ponto $i$. Assim, o denominador não sofre influência direta da própria observação cujo resíduo está sendo avaliado.

#### 8.2.3.1 Motivação conceitual

Nos resíduos padronizados,

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}},$$

Como apresentado anteriormente, o estimador $s^2$ é calculado usando todas as observações. Se a observação $i$ for discrepante, ela pode inflar $s^2$, reduzindo artificialmente $|r_i|$ e dificultando sua própria detecção.

Ao substituir $s$ por $s_{(i)}$, eliminamos essa retroalimentação. O resíduo passa a ser avaliado em relação a um modelo que não foi influenciado por ele mesmo.

#### 8.2.3.2 Distribuição exata

Sob as hipóteses do MRLS com erros normais,

$$\varepsilon_i \sim N(0, \sigma^2),$$

temos que

$$t_i^* \sim t_{n-3}.$$

A perda de um grau de liberdade adicional (em comparação com $t_{n-2}$) decorre do fato de que a variância foi estimada com $n-3$ graus de liberdade no modelo ajustado sem a observação $i$ (Kutner et al. (2005); Montgomery, Peck, e Vining (2021)).

Essa é uma propriedade importante: diferentemente dos resíduos padronizados, aqui a distribuição $t$ é **exata** sob normalidade dos erros.

> **Apêndice de Demonstrações {#demo}:** a distribuição exata de $t_i^*$ é obtida mostrando que, sob $H_0$, o numerador é normal e independente do estimador $s^2_{(i)}$, o qual é proporcional a uma variável qui-quadrado com $n-3$ graus de liberdade.

#### 8.2.3.3 Interpretação prática

Como $t_i^*$ segue exatamente uma distribuição $t$, podemos utilizar pontos críticos formais para avaliar discrepância individual:

- $|t_i^*| > t_{1-\alpha/2,\, n-3} \rightarrow$ evidência de que a observação $i$ é discrepante ao nível $\alpha$.

Na prática:

- $|t_i^*| > 2$ sugere possível discrepância;
- $|t_i^*| > 3$ indica forte indício de outlier, especialmente em amostras moderadas.

Elevando ao quadrado:

$$t_i^{*2} \sim F_{1,n-3},$$

pois o quadrado de uma variável com distribuição $t_k$ segue distribuição $F_{1,k}$ (Casella e Berger (2002)). Essa relação conecta o diagnóstico individual de observações com a lógica dos testes $F$ discutidos anteriormente.

## 8.3 Influência, alavancagem e leitura conjunta dos resíduos

A etapa mais importante do diagnóstico no MRLS consiste em integrar três dimensões distintas, mas complementares:

- discrepância na variável resposta ($Y$);
- posição extrema na variável explicativa ($X$);
- impacto global sobre os estimadores do modelo.

Essa integração é fundamental para evitar conclusões equivocadas baseadas apenas no tamanho do resíduo.

### 8.3.1 Relação entre discrepância, alavancagem e influência

É importante distinguir conceitualmente:

- **Possível outlier em** $Y$: grande $|t_i^*|$;
- **Alta alavancagem:** grande $h_{ii}$;
- **Observação influente:** combinação de grande $|t_i^*|$ e grande $h_{ii}$.

Um ponto pode apresentar alto resíduo, mas baixa alavancagem, afetando pouco a inclinação da reta. Nesse caso, ele é discrepante na resposta, mas não necessariamente influente.

Por outro lado, uma observação pode ter pequena discrepância em $Y$, mas alta alavancagem em $X$, alterando significativamente a inclinação estimada $\widehat{\beta}_1$. Nesse caso, mesmo com resíduo pequeno, o ponto pode ser estruturalmente influente.

### 8.3.2 Alavancagem no MRLS

A **alavancagem** da observação $i$ é dada por

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}, \quad \text{com} \quad S_{xx} = \sum_{j=1}^{n} (X_j - \bar{X})^2.$$

Ela mede o quanto o valor de $X_i$ influencia o próprio ajuste $\widehat{Y}_i$.

Propriedades importantes no MRLS:

- $0 < h_{ii} < 1$;
- $$\sum_{i=1}^{n} h_{ii} = 2,$$

  pois dois parâmetros são estimados ($\beta_0$ e $\beta_1$);

- a média das alavancagens é $2/n$.

Observações com $X_i$ muito afastado da média $\bar{X}$ possuem maior alavancagem e exercem maior influência geométrica sobre a reta ajustada (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

Uma regra prática comum é considerar como potencialmente alta alavancagem valores tais que

$$h_{ii} > \frac{2p}{n},$$

em que $p$ é o número de parâmetros do modelo (no MRLS, $p = 2$). Assim, valores acima de $4/n$ merecem atenção especial (Belsley, Kuh, e Welsch (1980)).

### 8.3.3 Conexão entre alavancagem e variância residual

Recordando que

$$Var(e_i) = \sigma^2(1 - h_{ii}),$$

vemos que observações com maior alavancagem apresentam menor variância residual. Isso ocorre porque esses pontos "puxam" a reta para mais perto de si.

Portanto, um ponto com alto $h_{ii}$ pode ter resíduo pequeno não porque esteja bem ajustado, mas porque influenciou fortemente o ajuste.

Essa distinção é conceitualmente importante:

- **Resíduo** mede discrepância vertical.
- **Alavancagem** mede posição extrema em $X$.
- **Influência** mede alteração no modelo quando a observação é removida.

### 8.3.4 Síntese diagnóstica

A leitura conjunta pode ser organizada da seguinte forma:

- **Resíduos grandes + baixa alavancagem**
  → outliers na resposta ($Y$), com impacto limitado na inclinação.
- **Resíduos pequenos + alta alavancagem**
  → observações potencialmente influentes, mesmo sem grande discrepância aparente.

- **Resíduos grandes + alta alavancagem**
  → casos críticos, com forte potencial de distorcer significativamente o ajuste.

Medidas integradas, como a distância de Cook,

$$D_i = \frac{t_i^{*2}}{2} \cdot \frac{h_{ii}}{1 - h_{ii}},$$

quantificam diretamente o quanto os estimadores $(\hat{\beta}_0, \hat{\beta}_1)$ se alterariam caso a observação $i$ fosse removida (Belsley, Kuh, e Welsch (1980); Weisberg (2005)).

### 8.3.5 Resumo comparativo dos resíduos

| Tipo de resíduo | Fórmula | $E(.)$ | $Var(.)$ | Distribuição |
|---|---|---|---|---|
| Ordinário $e_i$ | $Y_i - \hat{Y}_i$ | 0 | $\sigma^2(1 - h_{ii})$ | $N(0, \sigma^2(1 - h_{ii}))$ |
| Padronizado $r_i$ | $\frac{e_i}{s\sqrt{1 - h_{ii}}}$ | $\approx 0$ | $\approx 1$ | Aprox. $t_{n-2}$ |
| Estudentizado $t_i^*$ | $\frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}$ | 0 | 1 | $t_{n-3}$ |

Em síntese:

- **Resíduos ordinários** fornecem a discrepância bruta.
- **Resíduos padronizados** tornam as observações comparáveis.
- **Resíduos estudentizados externos** permitem inferência formal com distribuição $t$ exata sob normalidade.
- **Alavancagem** identifica observações estruturalmente extremas.
- **Medidas de influência** integram discrepância e posição.

Somente essa leitura integrada permite avaliar adequadamente a robustez do ajuste no MRLS e identificar observações com potencial de comprometer a inferência estatística.

## 8.4 Testes formais dos resíduos

Antes da inspeção gráfica, é possível realizar **testes estatísticos formais** aplicados aos resíduos do MRLS. Esses testes não substituem a análise gráfica, mas fornecem evidência quantitativa sobre possíveis violações das hipóteses clássicas, especialmente **normalidade** e **independência** dos erros.

É fundamental compreender que tais testes avaliam hipóteses específicas do modelo (por exemplo, normalidade dos erros), e não a "qualidade geral" da regressão. A interpretação correta exige articulação entre teoria, estatística e contexto (Kutner et al. (2005); Montgomery, Peck, e Vining (2021); Weisberg (2005)).

### 8.4.1 Teste para Assimetria (Skewness)

A estatística de assimetria é definida por

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2\right)^{3/2}},$$

cujo valores de referência são:

- $S = 0 \rightarrow$ simetria;
- $S > 0 \rightarrow$ cauda longa à direita;
- $S < 0 \rightarrow$ cauda longa à esquerda.

Sob a hipótese de normalidade dos resíduos, podemos formular as seguintes hipóteses:

$$H_0 : \text{Distribuição simétrica } (S = 0)$$

$$H_1 : \text{Distribuição assimétrica } (S \neq 0)$$

Para amostras grandes, vale a aproximação assintótica:

$$Z_S = \sqrt{\frac{n}{6}}\, S \sim N(0, 1).$$

Esse teste verifica se há evidência estatística de assimetria na distribuição residual. Valores positivos indicam cauda longa à direita; valores negativos indicam cauda longa à esquerda.

Assimetria residual pode indicar: - variável resposta naturalmente assimétrica (ex.: tempos, rendas); - necessidade de transformação; - presença de outliers em apenas um lado da distribuição.

A assimetria detectada estatisticamente pode ser irrelevante do ponto de vista prático se o impacto sobre estimativas e previsões for pequeno. Por isso, a análise gráfica (histograma e QQ-plot) é complementar e essencial (Weisberg (2005); Montgomery, Peck, e Vining (2021)).

### 8.4.2 Teste para Curtose (Kurtosis)

A curtose é definida por

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2\right)^2},$$

com os seguinte valores de referência:

- $K = 3 \rightarrow$ normal (mesocúrtica);
- $K > 3 \rightarrow$ caudas pesadas (leptocúrtica);
- $K < 3 \rightarrow$ caudas leves (platicúrtica).

Sob a hipótese de normalidade dos resíduos, podemos formular as seguintes hipóteses:

$$H_0 : K = 3$$

$$H_1 : K \neq 3$$

Para amostras grandes:

$$Z_K = \sqrt{\frac{n}{24}}(K - 3) \sim N(0, 1).$$

Curtose elevada frequentemente sinaliza presença de outliers ou heterogeneidade de variância. Caudas pesadas significam maior probabilidade de valores extremos, o que pode afetar inferência e previsão.

Assim como a assimetria, a curtose deve ser interpretada junto com resíduos estudentizados e medidas de influência. Muitas vezes, poucos pontos extremos explicam grande parte da rejeição da normalidade (Belsley, Kuh, e Welsch (1980); Weisberg (2005)).

### 8.4.3 3. Omnibus Test (D'Agostino–Pearson)

O teste Omnibus combina os dois testes anteriores (assimetria e curtose) em uma única estatística.

Sejam:

$$Z_1 = Z_S \quad \text{e} \quad Z_2 = Z_K.$$

A estatística do teste é:

$$OM = Z_1^2 + Z_2^2.$$

Ou seja, $Z_1$ é a estatística padronizada da assimetria e $Z_2$ a da curtose. Sob $H_0$ (normalidade), vale assintoticamente:

Para o teste Omnibus, formulmos as seguintes hipóteses:

$$H_0 : \text{Resíduos seguem distribuição normal}$$

$$H_1 : \text{Resíduos não seguem distribuição normal}$$

Sob $H_0$,

$$OM \sim \chi^2_{(2)}.$$

O Omnibus é um teste conjunto: ele detecta qualquer violação que afete simetria ou curtose. Em vez de avaliar dois testes separados, consolida evidência em uma única estatística. Adicionalmente, como a distribuição é assintótica, sua confiabilidade aumenta com o tamanho amostral. Em amostras pequenas, o teste pode apresentar distorções no nível de significância.

### 8.4.4 4. Jarque–Bera (JB)

O teste de Jarque–Bera também combina assimetria e curtose, mas diretamente em termos de seus estimadores:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right).$$

Hipóteses do teste são:

$$H_0 : \text{Resíduos seguem distribuição normal}$$

$$H_1 : \text{Resíduos não seguem distribuição normal}$$

Sob $H_0$,

$$JB \sim \chi^2_{(2)}.$$

Observe que o JB é equivalente, do ponto de vista assintótico, à soma dos quadrados das versões padronizadas de $S$ e $K - 3$.

O JB mede a distância conjunta entre a distribuição empírica dos resíduos e a normal, considerando forma (assimetria) e peso de caudas (curtose).

Note que rejeitar normalidade não implica que o modelo linear esteja incorreto, isso pode indicar apenas que os erros não são gaussianos. A relevância prática depende do objetivo (estimação, teste, previsão) e do tamanho da amostra (Casella e Berger (2002); Kutner et al. (2005)). ### 5. Durbin–Watson (DW)

O teste de Durbin–Watson verifica autocorrelação serial:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

As hipóteses clássicas são:

$$H_0 : \rho = 0 \quad \text{(ausência de autocorrelação)}$$

$$H_1 : \rho \neq 0 \quad \text{(autocorrelação)}$$

A interpretação usual é:

- $DW \approx 2 \rightarrow$ ausência de autocorrelação;
- $DW < 2 \rightarrow$ autocorrelação positiva;
- $DW > 2 \rightarrow$ autocorrelação negativa.

O DW mede o quanto os resíduos consecutivos diferem entre si. Se $e_t$ e $e_{t-1}$ forem semelhantes (dependência positiva), o numerador será pequeno e *DW* ficará abaixo de 2.

Este teste é especialmente relevante em dados ordenados temporalmente (econometria, séries temporais). Em dados sem ordem natural, sua aplicação é menos informativa (Gujarati (2006)).

Autocorrelação residual pode indicar: - tendência não modelada; - variáveis omitidas; - estrutura dinâmica inerente ao fenômeno.

Detectar autocorrelação é apenas o início do diagnóstico.

## 8.5 Diagnóstico gráfico do MRLS

A análise gráfica dos resíduos é uma das etapas mais importantes na verificação das hipóteses do MRLS. Os gráficos funcionam como ferramentas de diagnóstico visual, permitindo identificar padrões que revelem problemas estruturais no modelo Montgomery, Peck, e Vining (2021); Kutner et al. (2005); Weisberg (2005).

Em um **modelo bem especificado**, os resíduos devem se comportar como **ruído puro**: dispersão aleatória em torno de zero, variância aproximadamente constante e sem estrutura aparente. Em termos práticos, isso significa que, **condicionado aos valores de** $X$, não deve existir informação sistemática remanescente nos resíduos que pudesse ser capturada por uma reespecificação simples do modelo (por exemplo, inclusão de termos não lineares ou transformação da resposta) Montgomery, Peck, e Vining (2021); Weisberg (2005).

A seguir, são descritos os principais gráficos e o que se esperar de cada um.

### 8.5.1 Resíduos vs ajustados (linearidade e homoscedasticidade)

Este é o gráfico diagnóstico mais usado na prática, pois confronta diretamente o "erro" estimado (resíduo) com o nível de resposta previsto pelo modelo.

- **O que se espera**:
  - pontos dispersos aleatoriamente em torno da linha horizontal 0, sem padrão definido.
  - amplitude (dispersão vertical) aproximadamente constante ao longo de toda a faixa de $\hat{Y}_i$.
  - poucos pontos ultrapassando as faixas de referência usuais (por exemplo, $|r_i| \approx 2$ ou $|t_i^*| \approx 2$, dependendo do resíduo adotado).
- **O que indica problema**:

- **padrão em curva** → sugere que a relação média $E(Y \mid X)$ não está bem representada por uma função linear; pode indicar necessidade de termos como $X^2$ ou outra reespecificação funcional.

- **forma de funil** (variância aumenta ou diminui com $\hat{Y}_i$) → indício de heteroscedasticidade (variância não constante).

- **concentração de resíduos positivos (negativos)** em certas regiões → modelo subestima (superestima) sistematicamente nessas regiões, sugerindo viés local de especificação.

- **pontos isolados muito afastados** do conjunto principal → possível outlier/influência; a confirmação deve ser feita em leitura conjunta com resíduos estudentizados $t_i^*$, alavancagem $h_{ii}$ e medidas de influência como a distância de Cook Belsley, Kuh, e Welsch (1980); Weisberg (2005).

Para diagnóstico visual, é recomendável utilizar resíduos que sejam comparáveis entre observações. Assim, em muitos contextos prefere-se plotar resíduos padronizados ($r_i$) ou estudentizados externos ($t_i^*$), em vez de resíduos ordinários ($e_i$), pois estes últimos têm variância dependente da alavancagem ($1 - h_{ii}$) (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

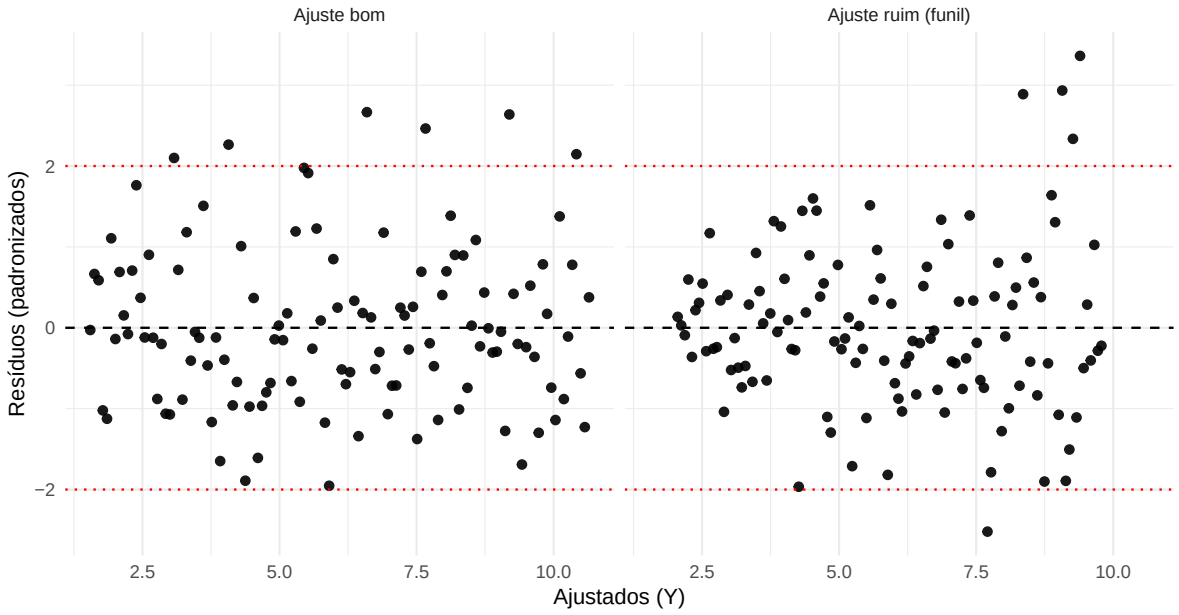



Figura 8.1: Resíduos vs Ajustados — Esquerda: Ajuste bom (homocedástico); Direita: Ajuste ruim (funil/heterocedasticidade). Linhas em 0 e ±2.

### 8.5.2 Resíduos vs $X$ (forma funcional)

Este gráfico é conceitualmente muito próximo ao anterior, mas desloca o foco: em vez de relacionar os resíduos com os valores ajustados $\hat{Y}_i$, relaciona-os diretamente com a variável explicativa $X_i$.

- **O que se espera**:
  - aleatoriedade semelhante ao gráfico anterior, mas agora em função de $X$.
  - dispersão aproximadamente constante ao longo de toda a faixa de $X$.
  - ausência de estruturas sistemáticas associadas a regiões específicas de $X$.
- **O que indica problema**:
  - **estruturas em forma de arco ou curva** $\rightarrow$ o efeito de $X$ pode ser não linear; o modelo linear $E(Y \mid X) = \beta_0 + \beta_1 X$ pode estar omitindo termos relevantes (por exemplo, $X^2$ ou outra transformação).
  - **padrões em "S" ou mudança de inclinação** $\rightarrow$ possível quebra de regime ou efeito estrutural não capturado.
  - **faixas onde a dispersão muda** $\rightarrow$ variação da variância conforme $X$, sugerindo heteroscedasticidade.
  - **concentração de pontos extremos em regiões específicas de** $X \rightarrow$ possível influência associada a valores extremos da variável explicativa.

Enquanto o gráfico resíduos vs ajustados enfatiza o comportamento do erro em relação à resposta prevista, o gráfico resíduos vs $X$ enfatiza a adequação da **forma funcional** da regressão. Ele permite avaliar diretamente se a hipótese de linearidade entre $X$ e a média condicional de $Y$ é plausível (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

Este gráfico é especialmente informativo quando $X$ possui interpretação física, econômica ou temporal clara. Nesses casos, padrões sistemáticos ao longo de $X$ podem revelar efeitos omitidos, mudanças estruturais ou fenômenos não lineares que não são imediatamente visíveis no gráfico resíduos vs ajustados.

Assim como no gráfico anterior, recomenda-se utilizar resíduos padronizados ou estudentizados para tornar a escala comparável entre observações, principalmente quando há variação relevante na alavancagem $h_{ii}$.

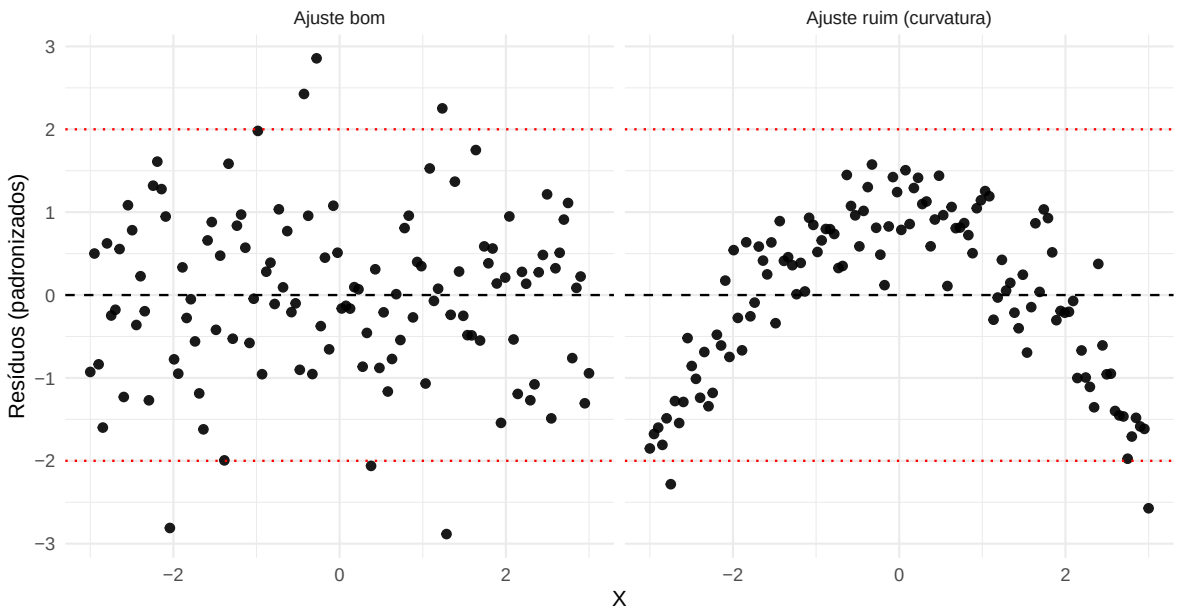



Figura 8.2: Resíduos vs X — Esquerda: Ajuste bom (linear); Direita: Ajuste ruim (não linearidade em arco). Linhas em 0 e ±2.

### 8.5.3 Resíduos estudentizados vs valores ajustados (outliers + estrutura)

Este gráfico é uma versão refinada do gráfico resíduos vs ajustados, utilizando os **resíduos estudentizados externos** $t_i^*$. Ele combina duas dimensões do diagnóstico: discrepância individual e possível estrutura sistemática.

- **Por que usar**:
  - tornam resíduos comparáveis, pois ajustam pela variância individual de cada ponto, incorporando o fator $(1 - h_{ii})$ associado à alavancagem.
  - utilizam uma estimativa da variância $\sigma^2$ calculada sem a observação $i$ $(s_{(i)})$, reduzindo o efeito de mascaramento que pode ocorrer quando um ponto extremo influencia a própria estimativa de variância.
  - possuem, sob normalidade dos erros, distribuição exata $t_{n-3}$, permitindo interpretação inferencial mais precisa (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).
- **O que se espera**:
  - aleatoriedade em torno da linha horizontal 0.

  - a maioria dos pontos entre $-2$ e $+2$, sendo raros valores com $|t_i^*| > 3$ em amostras moderadas.

  - ausência de padrão sistemático ao longo da faixa de valores ajustados.

- **O que indica problema**:

  - **pontos fora do intervalo** $[-2, 2]$ $\rightarrow$ observações potencialmente discrepantes; valores acima de $|t_i^*| > 3$ são frequentemente considerados fortemente suspeitos.

  - **estruturas visíveis (curvas, funis)** $\rightarrow$ possíveis violações de linearidade ou homocedasticidade, agora avaliadas com resíduos que já consideram diferenças de variância individual.

  - **concentração de valores extremos em regiões de alta alavancagem** $\rightarrow$ possível influência desproporcional sobre os estimadores.

Os resíduos estudentizados externos medem o quanto cada observação se afasta do modelo ajustado, levando em conta tanto a variabilidade residual quanto sua própria posição geométrica no conjunto de dados. Assim, eles são especialmente adequados para identificar **outliers reais**, isto é, observações cuja discrepância não pode ser explicada apenas por sua alavancagem.

Um ponto com resíduo ordinário grande pode deixar de parecer extremo após a estudentização se sua variância condicional for naturalmente maior. Por outro lado, um ponto que permanece extremo mesmo após a correção por $(1 - h_{ii})$ e por $s_{(i)}$ merece investigação cuidadosa — seja por erro de registro, seja por representar um fenômeno estrutural distinto (Belsley, Kuh, e Welsch (1980); Weisberg (2005)).

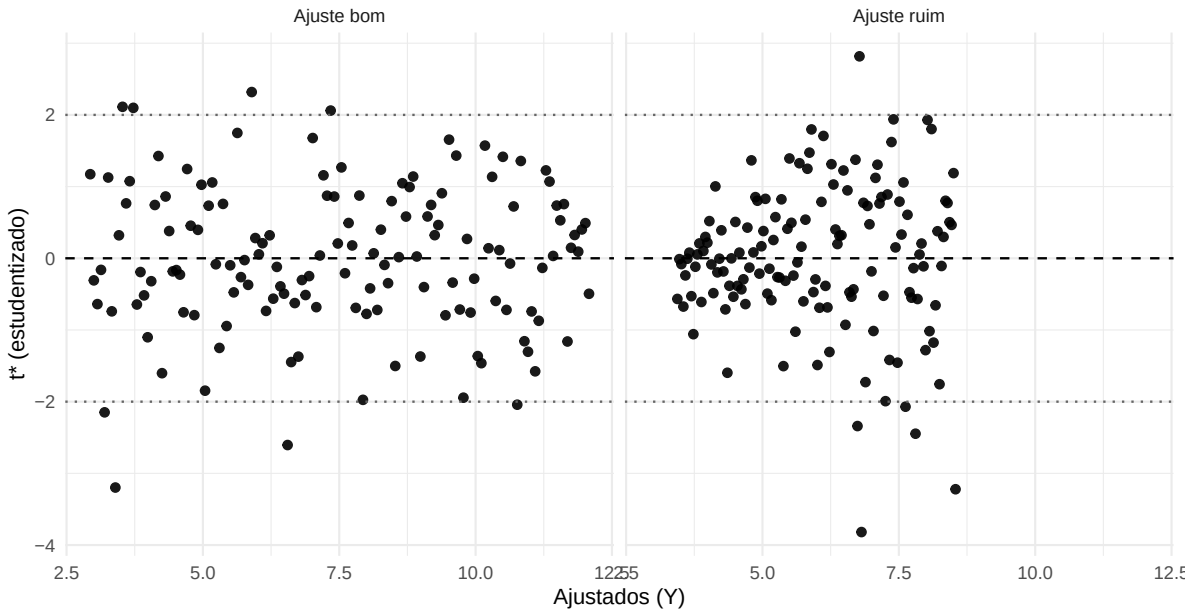



Figura 8.3: Resíduos estudentizados vs Ajustados — Esquerda: Ajuste bom; Direita: Ajuste ruim (curvatura + funil). Linhas em ±2.

### 8.5.4 QQ-plot (normalidade)

O gráfico QQ-plot (quantile–quantile) compara os quantis empíricos dos resíduos com os quantis teóricos de uma distribuição normal padrão. Ele é uma das ferramentas mais informativas para avaliar a hipótese de normalidade dos erros no MRLS (Montgomery, Peck, e Vining (2021); Kutner et al. (2005); Weisberg (2005)).

- **O que se espera**:
  - pontos aproximadamente alinhados em torno da reta de 45°, indicando que os resíduos seguem aproximadamente uma distribuição normal.
  - pequenas flutuações aleatórias ao redor da reta, especialmente no centro da distribuição.
  - ausência de desvios sistemáticos nas caudas.
- **O que indica problema**:
  - **desvios sistemáticos nas extremidades** → caudas mais pesadas (pontos afastados da reta nas pontas) ou mais leves que a normal.

- **desvios em formato de "S"** → indício de assimetria dos resíduos.

- **afastamentos persistentes ao longo de toda a reta** → possível inadequação global da suposição de normalidade.

- **pontos isolados muito distantes nas pontas** → presença de outliers, que podem ser responsáveis por grande parte da violação observada.

O QQ-plot compara toda a **forma da distribuição**. Se os resíduos forem normais, seus quantis empíricos devem crescer linearmente com os quantis teóricos da normal. Desvios sistemáticos dessa linearidade indicam diferenças estruturais entre as distribuições.

É fundamental interpretar o QQ-plot em conjunto com resíduos estudentizados e medidas de influência. Muitas vezes, poucos pontos extremos explicam a maior parte do desvio observado nas caudas. Além disso, pequenas curvaturas no centro do gráfico, especialmente em amostras grandes, podem não ter relevância prática para a inferência, sobretudo quando o objetivo principal é previsão e não testes exatos em pequenas amostras (Casella e Berger (2002); Kutner et al. (2005)).

O QQ-plot, portanto, oferece uma visão global da normalidade e complementa tanto os testes formais (como Jarque–Bera) quanto os gráficos de histograma.

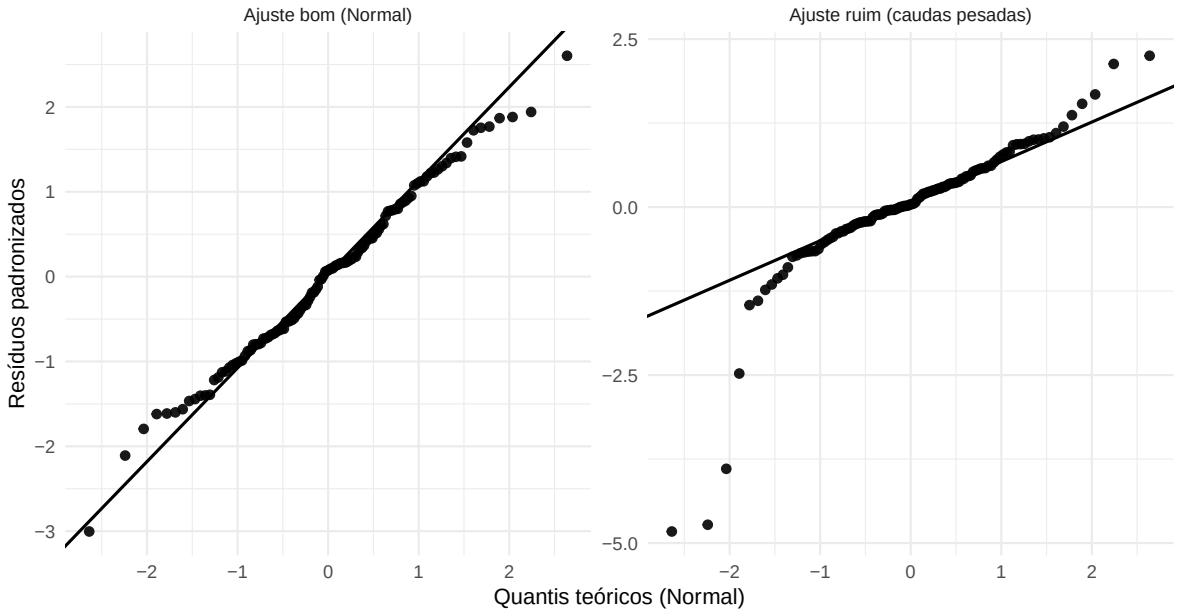



Figura 8.4: QQ-plot dos resíduos — Esquerda: Ajuste bom (erros ~ Normal); Direita: Ajuste ruim (erros ~ t com caudas pesadas).

### 8.5.5 Histograma (assimetria e caudas)

O histograma dos resíduos é uma ferramenta complementar ao QQ-plot. Enquanto o QQ-plot enfatiza o alinhamento com a normal teórica por meio de quantis, o histograma permite visualizar diretamente a **forma empírica** da distribuição residual (Montgomery, Peck, e Vining (2021); Kutner et al. (2005); Weisberg (2005)).

- **O que se espera**:
  - distribuição aproximadamente simétrica em torno de zero.
  - formato aproximadamente em sino (curva unimodal e suave).
  - maior concentração de valores próximos de 0, com frequência decrescente nas extremidades.
- **O que indica problema**:
  - **assimetria** → possível necessidade de transformação na resposta ($Y$), como $\log(Y)$ ou $\sqrt{Y}$, especialmente quando a assimetria é estrutural e não causada por poucos pontos extremos.
  - **caudas longas** → presença de outliers ou distribuição com maior probabilidade de valores extremos do que a normal.
  - **bimodalidade ou múltiplos picos** → possível mistura de grupos ou estrutura omitida no modelo (por exemplo, variável categórica não incluída).
  - **concentração excessiva no centro com poucas observações nas extremidades** → caudas leves (platicurtose), também incompatíveis com normalidade.

O histograma fornece uma visão direta da densidade empírica dos resíduos. Em um modelo com erros normais, espera-se que a forma geral seja compatível com a curva Normal$(0, \sigma^2)$. Desvios sistemáticos dessa forma indicam diferenças estruturais na distribuição do erro.

Adicionamente, o histograma é sensível à escolha do número de classes (bins). Diferentes escolhas podem alterar a percepção visual da forma. Por isso, recomenda-se utilizá-lo em conjunto com o QQ-plot e com medidas numéricas de assimetria e curtose.

Além disso, é importante lembrar que pequenas assimetrias visuais, especialmente em amostras grandes, podem não comprometer de forma relevante a inferência baseada em MQO, cuja robustez assintótica é discutida em (Casella e Berger (2002); Kutner et al. (2005)).

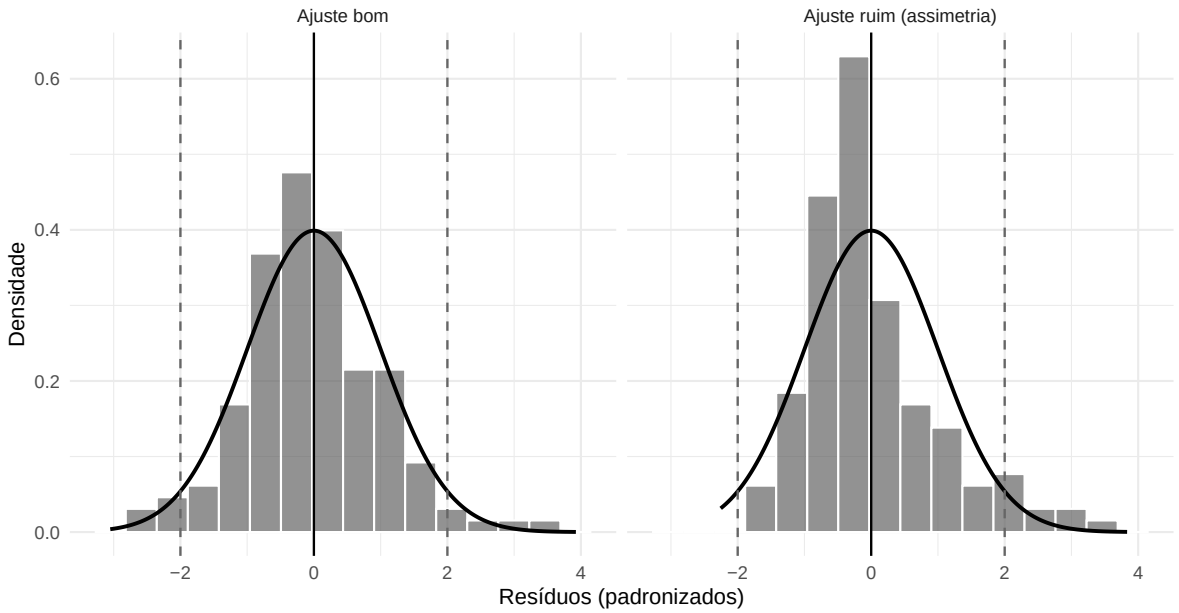



Figura 8.5: Histograma dos resíduos — Esquerda: Ajuste bom (simétrico); Direita: Ajuste ruim (assimetria à direita). Curva Normal(0,1) de referência.

### 8.5.6 Resíduos estudentizados vs índice (pontos atípicos)

Este gráfico apresenta os resíduos estudentizados externos $t_i^*$ em função do índice da observação $i$. Ele é particularmente útil para identificar **observações discrepantes individuais**, destacando sua posição relativa no conjunto de dados.

- **Por que usar**:
  - são melhores na detecção de outliers, pois corrigem a influência da própria observação ao utilizar a estimativa de variância $s_{(i)}$, calculada sem o ponto $i$.
  - possuem distribuição $t_{n-3}$ sob normalidade dos erros, permitindo interpretação inferencial direta.
  - facilitam a visualização de padrões associados à ordem natural dos dados (por exemplo, tempo ou sequência experimental) Montgomery, Peck, e Vining (2021); Kutner et al. (2005).
- **O que se espera**:
  - quase todos os pontos entre $-2$ e $+2$.

  - raros pontos ultrapassando $|t_i^*| > 3$, especialmente em amostras moderadas.

  - ausência de padrões sistemáticos ao longo do índice.

- **O que indica problema**:

  - **valores extremos em** $|t_i^*| \rightarrow$ sugerem observações potencialmente discrepantes; quanto maior o valor absoluto, maior a evidência de que o ponto não é compatível com a variabilidade esperada sob o modelo.

  - **agrupamento de valores extremos em determinadas regiões do índice** $\rightarrow$ pode indicar mudança estrutural, dependência temporal ou erro sistemático de medição.

  - **padrões alternados (positivo–negativo–positivo)** $\rightarrow$ possível autocorrelação residual, especialmente quando os dados possuem ordem temporal.

Este gráfico não apenas identifica outliers, mas também permite verificar se tais observações estão distribuídas aleatoriamente ao longo do conjunto de dados ou se seguem algum padrão estrutural.

O significado do eixo "índice" depende do contexto. Se os dados tiverem uma ordem natural (tempo, experimento sequencial, posição espacial), padrões nesse gráfico podem indicar violação da hipótese de independência dos erros. Se não houver ordem natural, o gráfico atua principalmente como ferramenta de localização de observações discrepantes.

Além disso, um valor extremo em $t_i^*$ não implica automaticamente exclusão da observação. Deve-se verificar conjuntamente a alavancagem $h_{ii}$ e medidas de influência (como a distância de Cook) antes de qualquer decisão sobre reespecificação ou remoção de dados (Belsley, Kuh, e Welsch (1980); Weisberg (2005)).

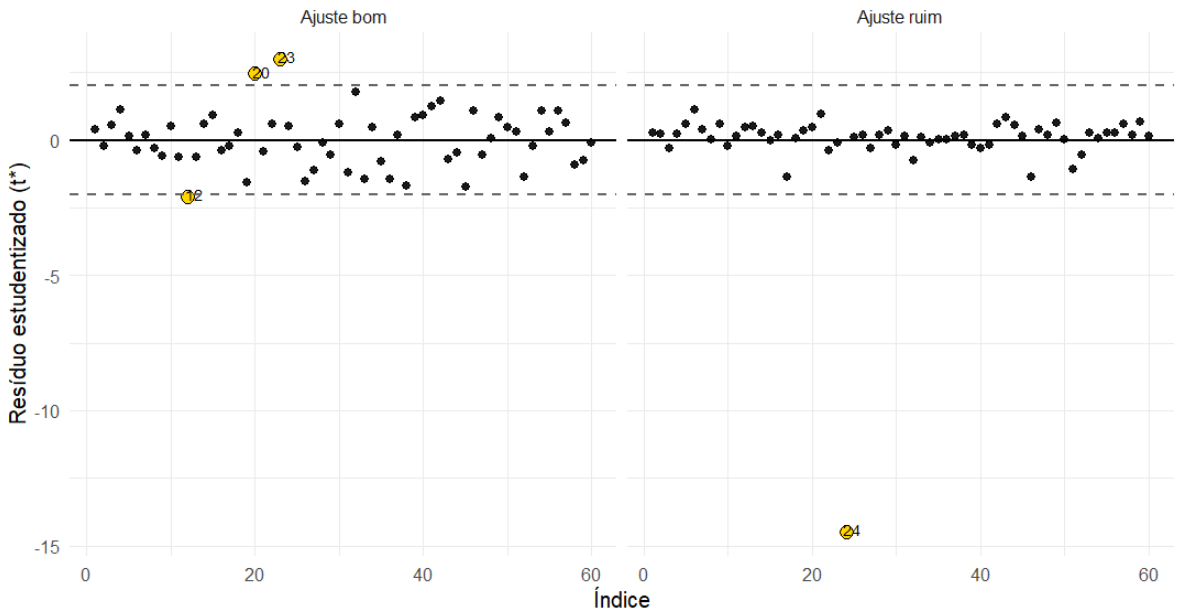



Figura 8.6: Resíduos estudentizados vs índice — Esquerda: ajuste bom (Normal, sem outliers); Direita: ajuste ruim (caudas pesadas + outliers). Linhas em 0 e ±2; pontos extremos destacados.

### 8.5.7 Resíduos estudentizados ao quadrado vs valores ajustados (heteroscedasticidade / influência)

Este gráfico utiliza $t_i^{*2}$ (resíduos estudentizados externos ao quadrado) no eixo vertical e os valores ajustados $\hat{Y}_i$ no eixo horizontal. Ao elevar ao quadrado, eliminamos o sinal e focamos exclusivamente na **magnitude da discrepância**, o que é particularmente útil para investigar padrões de variância.

- **O que se espera**:
  - dispersão aproximadamente uniforme ao longo da faixa de $\hat{Y}_i$.
  - ausência de tendência sistemática crescente ou decrescente.
  - pontos distribuídos sem estrutura definida ao redor de um nível aproximadamente constante.
- **O que indica problema**:
  - **crescimento ou redução sistemática** de $t_i^{*2}$ conforme $\hat{Y}_i$ aumenta → indício de heteroscedasticidade (variância não constante).

– **estrutura em arco** → possível não linearidade na função média.

– **pontos isolados com valores muito elevados de** $t_i^{*2}$ → observações potencialmente influentes ou discrepantes.

Se o modelo satisfaz a hipótese de homocedasticidade, então $Var(e_i)$ deve ser constante. Como $t_i^*$ já corrige por $(1 - h_{ii})$ e por $s_{(i)}$, padrões sistemáticos em $t_i^{*2}$ sugerem que a variância condicional de $Y$ depende do nível da resposta, ou seja, $Var(Y \mid X)$ não é constante.

Ao trabalhar com o quadrado do resíduo, pequenas diferenças tornam-se mais visíveis. Por isso, esse gráfico frequentemente revela tendências de variância que não são tão evidentes no gráfico simples resíduos vs ajustados.

A inclusão de uma curva suave (por exemplo, LOESS) auxilia na visualização de tendências médias na magnitude dos resíduos. Se essa curva apresentar inclinação clara ou formato sistemático, há evidência visual de heteroscedasticidade (Montgomery, Peck, e Vining (2021); Kutner et al. (2005); Weisberg (2005)).

Este gráfico, portanto, complementa o diagnóstico tradicional, oferecendo uma perspectiva focada especificamente na estrutura da variância.

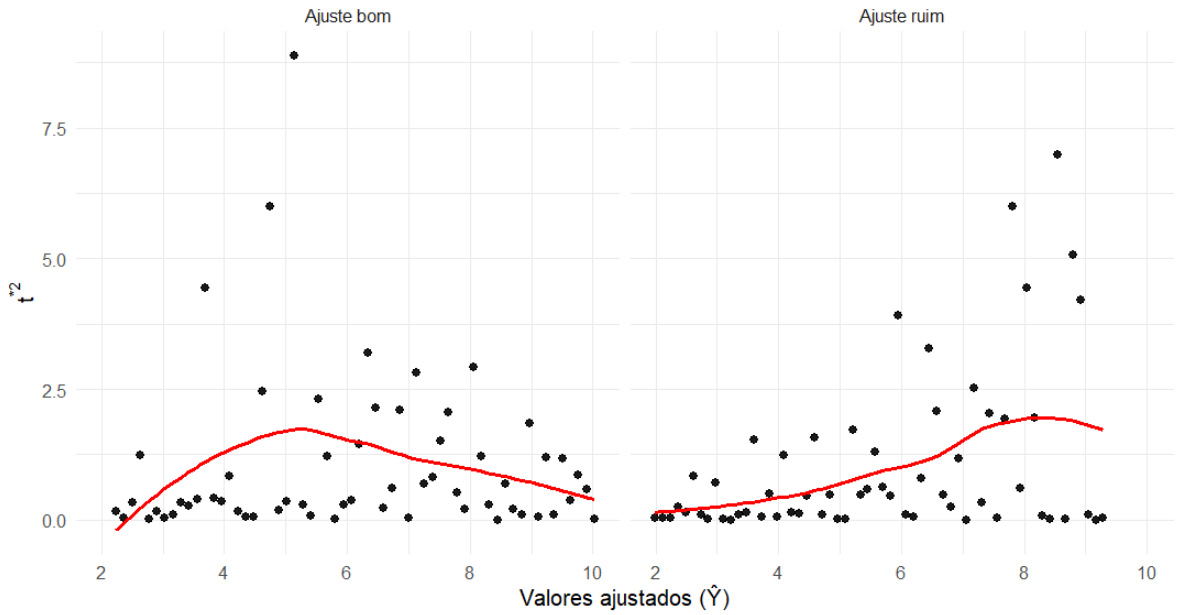



Figura 8.7: Valores ajustados vs resíduos estudentizados² — Esquerda: ajuste bom (dispersão uniforme); Direita: ajuste ruim (heteroscedasticidade/funil). LOWESS em vermelho.

### 8.5.8 Alavancagem vs resíduos estudentizados (influência / Cook)

Este gráfico combina duas dimensões centrais do diagnóstico no MRLS:
- **alavancagem** ($h_{ii}$), que mede o quão extremo é o valor de $X_i$;
- **resíduo estudentizado externo** ($t_i^*$), que mede a discrepância vertical ajustada pela variância condicional.

Ao analisar ambos simultaneamente, obtemos uma visão direta da **influência potencial** de cada observação sobre os estimadores do modelo (Belsley, Kuh, e Welsch (1980); Weisberg (2005); Montgomery, Peck, e Vining (2021)).

- **Objetivo**:
  - identificar observações influentes, isto é, aquelas que combinam alto resíduo e alta alavancagem.
  - distinguir entre pontos apenas discrepantes (grande $|t_i^*|$) e pontos estruturalmente extremos (grande $h_{ii}$).

- **O que se espera**:
  - maioria dos pontos dentro da "nuvem central", isto é, com valores moderados de $h_{ii}$ e $|t_i^*| \leq 2$.
  - poucos pontos próximos do limite usual de alavancagem (por exemplo, $h_{ii} > 2p/n$).
  - ausência de observações simultaneamente extremas em ambas as direções.

- **O que indica problema**:
  - **pontos afastados horizontalmente** (alto $h_{ii}$) $\rightarrow$ observações com grande potencial de influenciar a inclinação da reta.
  - **pontos afastados verticalmente** (alto $|t_i^*|$) $\rightarrow$ observações discrepantes na resposta.
  - **pontos afastados horizontal e verticalmente** $\rightarrow$ fortes candidatos a observações influentes, com impacto potencialmente desproporcional sobre os estimadores.

Influência não é sinônimo de discrepância. Um ponto pode ter grande resíduo, mas baixa alavancagem, afetando pouco os coeficientes. Da mesma forma, um ponto pode ter alta alavancagem e resíduo pequeno, mas ainda assim influenciar a inclinação da reta por estar em região extrema de $X$.

Observações influentes não devem ser automaticamente removidas. Elas podem representar fenômenos legítimos do processo gerador dos dados. O papel do diagnóstico é identificar e compreender tais pontos, não eliminá-los mecanicamente.

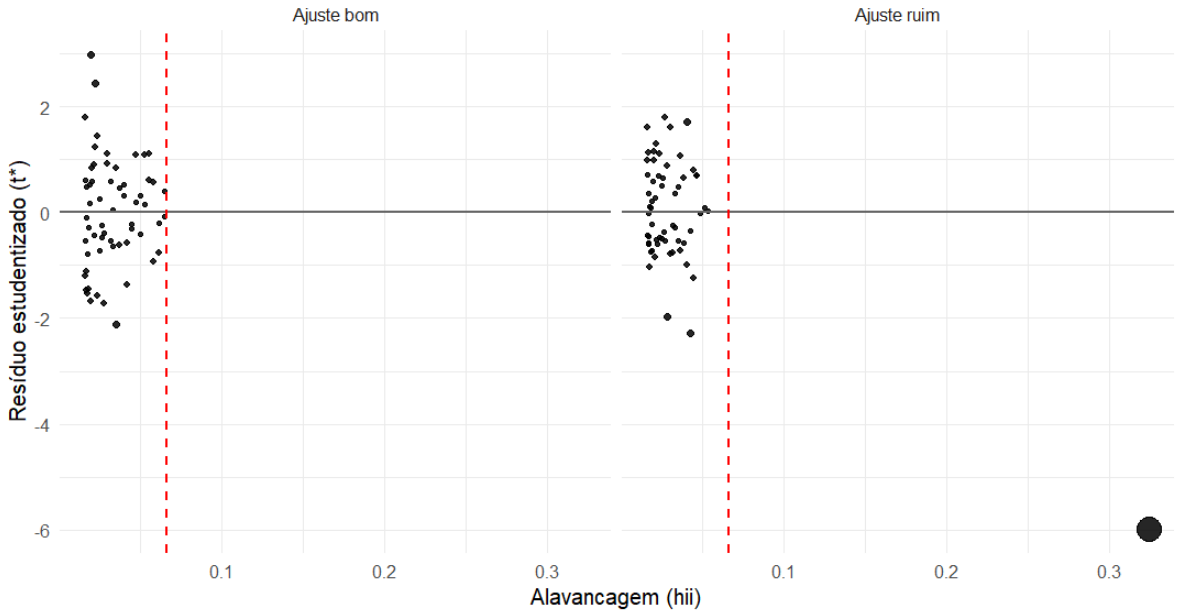



Figura 8.8: Alavancagem (h) vs resíduos estudentizados — Esquerda: ajuste bom; Direita: ajuste ruim com ponto de alta alavancagem e discrepância. Área  distância de Cook. Linha vertical: h > 2p/n.

Os gráficos de resíduos são **mapas visuais do ajuste**. Eles sintetizam, de forma intuitiva, as hipóteses estruturais do MRLS e permitem avaliar se o modelo capturou adequadamente a relação entre $X$ e $Y$ (Montgomery, Peck, e Vining (2021); Kutner et al. (2005); Weisberg (2005)).

Quando o modelo está bem especificado, espera-se que:

- os resíduos flutuem aleatoriamente em torno de zero;
- a variância seja aproximadamente constante ao longo de toda a faixa de $X$ ou $\hat{Y}$;
- a distribuição seja aproximadamente normal (quando a inferência exata via $t$ e $F$ é relevante);
- não existam observações com influência desproporcional sobre os estimadores.

Essas características indicam que, condicionalmente a $X$, o modelo não deixou estrutura sistemática não explicada.

Quando há padrões, eles indicam possíveis caminhos de correção:

- **Curvaturas** nos gráficos → possível inadequação da forma funcional; pode ser necessária a inclusão de termos como $X^2$, transformações em $X$ (por exemplo, $\log(X)$) ou outra reespecificação da função média.

- **Variância crescente ou decrescente** → indício de heteroscedasticidade; transformações em $Y$ (como $\log(Y)$ ou $\sqrt{Y}$) ou métodos que acomodem variância não constante podem ser considerados.

- **Assimetria na distribuição dos resíduos** → possível necessidade de transformação na resposta ou presença de outliers que devem ser investigados.

- **Observações influentes** → revisão individual do ponto, verificação de erros de registro ou análise substantiva do fenômeno representado.

Cada padrão visual corresponde a uma hipótese específica do modelo. Assim, o diagnóstico gráfico não é apenas uma etapa técnica, mas uma verificação das suposições matemáticas que fundamentam a inferência no MRLS.

Transformações não devem ser aplicadas de forma automática ou mecânica. Elas devem ser justificadas teoricamente, interpretadas no contexto do problema e validadas por novo ciclo de diagnóstico após o reajuste do modelo. O processo é iterativo: ajustar → diagnosticar → reespecificar → diagnosticar novamente.

## 8.6 Aspectos computacionais para resíduos no R

Para a análise gráfica de resíduos no **MRLS**, podemos usar principalmente os pacotes:

- `stats` → para ajustar o modelo com `lm()` e extrair resíduos e ajustados (`residuals()`, `fitted()`).
- `ggplot2` → para construir gráficos com `geom_point()`, `geom_hline()`, `geom_vline()`, `facet_wrap()`.
- `broom` → para organizar saídas do modelo em data frames (`augment()`), incluindo resíduos e ajustados.
- `car` (opcional) → para alguns diagnósticos e gráficos prontos (ex.: `residualPlots()`).
- `ggfortify` (opcional) → para gráficos diagnósticos automáticos a partir de objetos `lm` (`autoplot()`).
- `stats` + `ggplot2` → para QQ-plot (via `qqnorm()/qqline()` ou construção manual para `ggplot2`).

- **MASS** (opcional) → para simulações com distribuições alternativas quando necessário.
- **lmtest** / **sandwich** (opcional) → para testes formais (Breusch–Pagan etc.) e erros-padrão robustos.

A seguir, quais funções e pacotes usar em cada gráfico:

**1. Resíduos vs. Valores Ajustados**

- Objetivo: verificar aleatoriedade e homocedasticidade.
- Funções/objetos:
  - `modelo <- lm(Y ~ X, data=df)`
  - `fitted <- fitted(modelo)`
  - `resid  <- resid(modelo)`
- Em ggplot2: `geom_point()` + `geom_hline(yintercept=0, ...)`.

**2. Resíduos vs. Variável Explicativa ($X$)**

- Objetivo: avaliar linearidade da relação entre $Y$ e $X$.
- Funções/objetos:
  - `resid <- resid(modelo)`
- Em ggplot2: `geom_point()` com `aes(x=X, y=resid)` + `geom_hline(yintercept=0, ...)`.

**3. Resíduos Estudentizados vs. Valores Ajustados**

- Objetivo: detectar outliers e padrões (considerando alavancagem).
- Funções/objetos:
  - `stud <- rstudent(modelo)` (resíduos estudentizados externos)
  - `fitted <- fitted(modelo)`
- Em ggplot2: `geom_point()` + `geom_hline(yintercept=c(-2,2), ...)`.

**4. QQ-Plot dos Resíduos**

- Objetivo: verificar normalidade.
- Funções/objetos:
  - `resid <- resid(modelo)`
  - Base R: `qqnorm(resid); qqline(resid)`
  - Para ggplot2: `qq <- qqnorm(resid, plot.it=FALSE)` e então `geom_point()` + `geom_abline()`.

**5. Histograma dos Resíduos**

- Objetivo: verificar forma aproximada da distribuição.
- Funções/objetos:
  - `resid_pad <- scale(resid(modelo))` (opcional)
- Em `ggplot2`:
  - `geom_histogram(aes(y=after_stat(density)))`
  - Curva Normal de referência: `dnorm(x, 0, 1)` via `geom_line()` com uma grade `x`.

**6. Resíduos Estudentizados vs. Índice da Observação**

- Objetivo: identificar observações discrepantes.
- Funções/objetos:
  - `stud <- rstudent(modelo)`
  - `idx <- seq_along(stud)`
- Em `ggplot2`: `geom_point()` + `geom_hline(yintercept=c(-2,2), ...)`.

**7. Valores Ajustados vs. Resíduos Estudentizados²**

- Objetivo: investigar heterocedasticidade.
- Funções/objetos:
  - `fitted <- fitted(modelo)`
  - `t2 <- rstudent(modelo)^2`
- Suavização:
  - `geom_smooth(method="loess", se=FALSE, ...)` (substitui o LOWESS do Python de forma direta).

**8. Alavancagem vs. Resíduos Estudentizados (Influence Plot)**

- Objetivo: detectar observações influentes.
- Funções/objetos:
  - `h <- hatvalues(modelo)` (alavancagem)
  - `stud <- rstudent(modelo)` (resíduo estudentizado)
  - `ck <- cooks.distance(modelo)` (distância de Cook)
- Em `ggplot2`: `geom_point(aes(size=ck))` para tornar a área proporcional a Cook + linhas de referência (ex.: `geom_vline()` com regra 2(p+1)/n).

**Resumo didático:**

1. Ajuste o modelo com `lm()`:
   - `modelo <- lm(Y ~ X, data = df)`

2. Obtenha resíduos e ajustados:
    - `resid  <- resid(modelo)`
    - `fitted <- fitted(modelo)`
3. Obtenha diagnósticos extras:
    - `stud <- rstudent(modelo)`
    - `h    <- hatvalues(modelo)`
    - `ck   <- cooks.distance(modelo)`
4. (Opcional) Organize tudo em um único data frame (facilita gráficos):
    - `broom::augment(modelo)` (traz `.fitted`, `.resid`, `.std.resid`, `.hat`, `.cooksd`)

Com esses objetos, é possível montar todos os oito gráficos de resíduos apresentados neste capítulo.

# 9 Transformações nas Variáveis

## 9.1 Motivação

Quando os **gráficos de resíduos** revelam padrões sistemáticos, heteroscedasticidade ou violações de normalidade, é sinal de que a especificação do modelo pode estar inadequada. Em termos formais, isso significa que ao menos uma das hipóteses centrais do MRLS — linearidade da função média, homoscedasticidade ou normalidade dos erros — pode não estar sendo satisfeita (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

A regressão linear simples assume que

$$E(Y \mid X) = \beta_0 + \beta_1 X \quad \text{e} \quad Var(Y \mid X) = \sigma^2,$$

ou seja, uma **relação linear na média** e **variância constante** condicionalmente a $X$. Quando os resíduos exibem padrões como funil (variância crescente), curvaturas sistemáticas ou forte assimetria, o modelo ajustado não está capturando adequadamente a estrutura do processo gerador dos dados.

Uma das abordagens mais tradicionais para lidar com tais situações é aplicar **transformações matemáticas** às variáveis, de modo a:

- estabilizar a variância;
- aproximar a normalidade dos erros;
- linearizar a relação entre as variáveis;
- tornar a interpretação estatística mais coerente com a estrutura do fenômeno estudado.

Há duas formas de realizar tais transformações:
(a) na variável resposta ($Y$) e
(b) na variável explicativa ($X$).

Note que, ao aplicarmos uma transformação de variáveis, estamos implicitamente redefinindo o modelo estatístico. Por exemplo, ao transformar $Y$ em $\log(Y)$, deixamos de modelar diretamente $E(Y \mid X)$ e passamos a modelar $E(\log Y \mid X)$. Isso altera a interpretação dos coeficientes, a distribuição assumida para os erros e, em certos casos, a própria classe de modelos implícita (Weisberg (2005); Kutner et al. (2005)).

Portanto, transformações devem ser motivadas por evidência empírica (diagnóstico de resíduos) e, sempre que possível, por fundamentação teórica do fenômeno estudado. Aplicá-las de forma automática pode melhorar métricas numéricas de ajuste, mas comprometer a interpretação do modelo.

Além disso, a aplicação de uma transformação implica um novo ciclo completo de análise:

1. Ajustar o modelo transformado.
2. Reavaliar os resíduos.
3. Verificar se as hipóteses agora são plausíveis.

Esse processo é inerentemente iterativo.

Quando a transformação envolve apenas mudanças algébricas simples (como logaritmo ou raiz quadrada), as justificativas formais podem ser entendidas via propriedades de variância e distribuições conhecidas; demonstrações mais técnicas dessas propriedades podem ser consultadas no Apêndice de Demonstrações {#demo}.

Nas seções seguintes, analisaremos separadamente as transformações aplicadas à variável resposta e à variável explicativa, enfatizando sua motivação estatística e suas implicações interpretativas.

## 9.2 Transformações na variável resposta ($Y$)

As transformações na variável resposta têm como objetivo principal modificar a **estrutura probabilística condicional de $Y$ dado $X$**. Em muitos contextos aplicados, a violação das hipóteses do MRLS decorre não da forma funcional da média, mas do comportamento da variância ou da distribuição dos erros.

Recordando que o MRLS assume $Var(Y \mid X) = \sigma^2$,qualquer evidência de que

$$Var(Y \mid X) \neq \text{constante}$$

ou que os erros apresentam forte assimetria ou caudas pesadas pode motivar uma transformação em $Y$ (Montgomery, Peck, e Vining (2021); Kutner et al. (2005); Weisberg (2005)).

Para exemplificar, suponha que a variância condicional dependa da média:

$$Var(Y \mid X) = g(E(Y \mid X)),$$

para alguma função $g(\cdot)$ crescente. Isso ocorre, por exemplo:

- quando a variância é proporcional à média (dados de contagem);
- quando a variância cresce aproximadamente com o quadrado da média (dados positivos e assimétricos);
- quando há padrão de funil nos resíduos vs. ajustados.

Nesse cenário, podemos buscar uma transformação $Y^* = h(Y)$ tal que

$$Var(Y^* \mid X) \approx \text{constante}.$$

Esse princípio é conhecido como **estabilização da variância**, e sua justificativa formal pode ser obtida por expansão de Taylor de primeira ordem (ver Apêndice de Demonstrações {#demo}).

É importante destacar que transformar $Y$ altera simultaneamente: - a escala da resposta; - a forma da distribuição condicional; - a interpretação dos coeficientes.

Portanto, não estamos apenas "melhorando resíduos", mas redefinindo o modelo estatístico.

### 9.2.1 Transformação e distribuição implícita

Se ajustamos o modelo

$$Y_i^* = h(Y_i) = \beta_0^* + \beta_1^* X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

então estamos assumindo normalidade **na escala transformada**.

Isso implica que, na escala original, $Y$ possui distribuição derivada da transformação inversa. Por exemplo:

- se $h(Y) = \log Y$, então $Y$ é log-normal condicionalmente a $X$;
- se $h(Y) = \sqrt{Y}$, a distribuição resultante não é normal na escala original, mas a variância pode se tornar aproximadamente constante;
- se $h(Y) = 1/Y$, a estrutura média passa a ser modelada em termos do inverso da resposta.

Essa mudança tem implicações importantes para:

- interpretação de coeficientes;
- construção de intervalos de confiança;
- previsão na escala original.

Ao transformar $Y$, a hipótese de normalidade passa a ser avaliada na nova escala. Isso significa que testes $t$, testes $F$ e intervalos de confiança são válidos sob a suposição de normalidade dos erros na escala transformada, não necessariamente na escala original (Kutner et al. (2005)).

Nas subseções seguintes, discutiremos transformações específicas, a saber, logaritmo, raiz quadrada, inverso e potência.

### 9.2.2 Logaritmo ($Y^* = \log Y$)

A transformação logarítmica é uma das mais utilizadas em regressão, especialmente quando $Y$ é **positiva** e apresenta **assimetria à direita** ou variância crescente com o nível médio.

Se ajustamos o modelo

$$Y_i^* = \log(Y_i) = \beta_0^* + \beta_1^* X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

estamos assumindo normalidade dos erros **na escala logarítmica**. Isso implica que, condicionalmente a $X_i$, a variável original $Y_i$ segue uma **distribuição log-normal** (Kutner et al. (2005); Weisberg (2005)).

Mais precisamente, se

$$\log(Y_i) \mid X_i \sim N(\mu_i, \sigma^2),$$

então

$$Y_i \mid X_i \sim \text{LogNormal}(\mu_i, \sigma^2).$$

A transformação logarítmica converte relações multiplicativas em aditivas. Se, na escala original,

$$Y = \alpha \exp(\beta X) \cdot U,$$

onde $U$ é um termo multiplicativo de erro, então

$$\log Y = \log \alpha + \beta X + \log U,$$

ou seja, o modelo passa a ter estrutura linear com erro aditivo na escala transformada.

Isso significa que o logaritmo é particularmente adequado quando:

- a variabilidade é proporcional ao nível médio;

- o crescimento é aproximadamente exponencial;
- o erro atua de forma **multiplicativa** na escala original.

No modelo

$$E[\log(Y) \mid X] = \beta_0^* + \beta_1^* X,$$

o coeficiente $\beta_1^*$ representa a variação **aditiva no logaritmo da resposta** para cada unidade adicional em $X$.

Na escala original,

$$E(Y \mid X) = \exp(\beta_0^*) \cdot \exp(\beta_1^* X).$$

Assim, um aumento de 1 unidade em $X$ multiplica o valor esperado de $Y$ por $\exp(\beta_1^*)$.

Para valores pequenos de $\beta_1^*$, podemos usar a aproximação

$$\exp(\beta_1^*) \approx 1 + \beta_1^*,$$

o que leva à interpretação aproximada de que $\beta_1^* \times 100\%$ representa uma variação percentual em $Y$ para cada unidade de $X$.

Essa interpretação percentual é uma aproximação válida apenas quando $|\beta_1^*|$ é pequeno. O efeito exato é multiplicativo e deve ser interpretado via $\exp(\beta_1^*)$.

#### 9.2.2.1 Variância e estabilização

Se a variância cresce aproximadamente de forma proporcional ao quadrado da média,

$$Var(Y \mid X) \propto [E(Y \mid X)]^2,$$

então a transformação logarítmica tende a produzir variância aproximadamente constante na escala transformada.

**Advertência sobre retransformação**

Ao obter previsões na escala original, não é correto simplesmente calcular

$$\hat{Y} = \exp(\widehat{\log Y}),$$

pois, para variável log-normal,

$$E(Y \mid X) = \exp(\mu + \tfrac{1}{2}\sigma^2),$$

e não apenas $\exp(\mu)$ (Casella e Berger (2002)). Ignorar esse termo pode introduzir **viés de retransformação**.

**Quando usar**

- Resíduos vs. ajustados exibem padrão de funil.
- Histograma dos resíduos apresenta assimetria à direita.
- QQ-plot mostra caudas pesadas superiores.
- Relação parece exponencial na escala original.

**Gráficos sugeridos para verificação:**

- Resíduos vs. $\widehat{Y}$ (o que verificar: formato de funil; resultado esperado após transformação: dispersão homogênea).
- QQ-plot dos resíduos (o que verificar: caudas pesadas; resultado esperado após transformação: alinhamento mais próximo à reta).
- Histograma dos resíduos (o que verificar: assimetria; aresultado esperado após transformação: formato mais próximo da normal).

### 9.2.3 Raiz quadrada ($Y^* = \sqrt{Y}$)

A transformação pela raiz quadrada é frequentemente utilizada quando $Y$ representa **contagens** ou frequências.

Por exemplo, se $Y_i \sim \text{Poisson}(\mu_i)$, então

$$E(Y_i) = \mu_i, \quad Var(Y_i) = \mu_i,$$

ou seja, a variância é proporcional à média.

Aplicando a transformação

$$Y_i^* = \sqrt{Y_i},$$

e utilizando aproximação por expansão de Taylor, obtém-se

$$Var(Y_i^*) \approx \frac{1}{4},$$

que é aproximadamente constante e independe de $\mu_i$ (ver Apêndice de Demonstrações {#demo}).

Note que a raiz quadrada reduz a assimetria típica de distribuições de contagem e estabiliza a variância quando esta é proporcional à média.

O modelo passa então a explicar a variação na **raiz da resposta**, não na resposta original. Isso altera a escala e deve ser explicitado ao interpretar coeficientes.

**Quando usar**

- Dados de contagem.
- Variância aproximadamente proporcional à média.
- Histograma com cauda longa à direita.

**Gráficos sugeridos para verificação:**

- Histograma dos resíduos (o que verificar: cauda longa; resultado esperado após transformação: simetria).
- Resíduos vs. $\hat{Y}$ (o que verificar: variância crescente; resultado esperado após transformação: dispersão estável).

### 9.2.4 Inverso ($Y^* = 1/Y$)

A transformação inversa é útil quando a relação entre $Y$ e $X$ apresenta comportamento do tipo **decaimento rápido em direção a um limite**.

Se a relação original for não linear, por exemplo,

$$Y = \frac{1}{\alpha + \beta X},$$

então

$$\frac{1}{Y} = \alpha + \beta X,$$

que é linear.

Essa transformação é comum em fenômenos físicos e biológicos nos quais a resposta diminui rapidamente e depois se estabiliza.

O modelo ajustado explica o comportamento do **inverso da resposta**. A interpretação substantiva deve ser feita com cuidado, pois aumentos em $X$ passam a produzir efeitos lineares sobre $1/Y$, e não diretamente sobre $Y$.

**Quando usar**

Modelos onde a resposta diminui de forma não linear (ex.: tempo de reação decrescendo com aumento de dose).

**Gráficos sugeridos para verificação:**

- Resíduos vs. $X$ (o que verificar: padrão curvilíneo; resultado esperado após transformação: dispersão aleatória).
- QQ-plot dos resíduos (o que verificar: forte desvio; resultado esperado após transformação: mais alinhado).

### 9.2.5 Quadrado ($Y^* = Y^2$)

A transformação quadrática pode ser útil quando a variabilidade é maior em valores pequenos de $Y$ ou quando a relação é convexa na escala original.

Ao elevar $Y$ ao quadrado, ampliamos diferenças em níveis mais altos da resposta, o que pode reduzir padrões estruturais nos resíduos.

Essa transformação modifica significativamente a escala e deve ser utilizada apenas quando há justificativa empírica clara nos gráficos de resíduos.

**Gráficos sugeridos para verificação:**

- Resíduos vs. $\hat{Y}$ (o que verificar: padrão em arco; resultado esperado após transformação: dispersão homogênea).
- Histograma dos resíduos (o que verificar: concentração em torno de 0; resultado esperado após transformação: mais espalhado).

## 9.3 Transformações na variável explicativa ($X$)

As transformações na variável explicativa têm como objetivo principal **linearizar a relação entre $X$ e a média condicional de $Y$**, preservando, quando possível, a estrutura de variância constante dos erros.

Recordando que o MRLS assume $E(Y \mid X) = \beta_0 + \beta_1 X$, portanto, qualquer evidência de que a relação média entre $Y$ e $X$ não é linear, por exemplo, presença de curvatura sistemática no gráfico de resíduos vs. $X$, sugere que a especificação funcional pode estar inadequada (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

Nesse caso, buscamos uma transformação $X^* = h(X)$ tal que

$$E(Y \mid X^*) = \beta_0 + \beta_1 X^*$$

represente melhor a estrutura média do fenômeno.

Transformar $X$ significa alterar a **forma funcional da regressão** em relação à variável $X$, mas não a natureza probabilística da variável resposta. Diferentemente das transformações em $Y$, aqui a distribuição condicional de $Y$ não é redefinida; apenas modificamos a forma como a média depende da variável explicativa.

Portanto:

- Transformações em $Y \rightarrow$ alteram a escala da resposta e a estrutura probabilística.
- Transformações em $X \rightarrow$ alteram a forma funcional da média em relação à variável $X$.

### 9.3.1 Motivação estatística

Se o gráfico de resíduos vs. $X$ exibe:

- padrão em arco (concavidade ou convexidade),
- forma em S,
- tendência sistemática crescente ou decrescente,

então o modelo linear $\beta_0 + \beta_1 X$ não está capturando adequadamente a relação entre as variáveis. A solução é buscar uma transformação $h(X)$ tal que a nova relação seja aproximadamente linear.

Essa estratégia é coerente com o princípio geral de modelagem estatística: **especificar corretamente a função média antes de avaliar a variância dos erros** (Weisberg (2005)).

### 9.3.2 Logaritmo ($X^* = \log X$)

A transformação logarítmica em $X$ é apropriada quando a relação entre $Y$ e $X$ apresenta comportamento de crescimento proporcional ou lei de potência.

Suponha que a relação verdadeira seja

$$Y = \alpha X^{\beta} + \varepsilon.$$

Tomando logaritmo em ambos os lados (desconsiderando momentaneamente o erro),

$$\log Y = \log \alpha + \beta \log X.$$

Se também transformarmos $Y$, obtemos o chamado modelo **log-log**, amplamente utilizado em econometria.

No entanto, mesmo sem transformar $Y$, pode ser apropriado modelar

$$Y = \beta_0 + \beta_1 \log X + \varepsilon,$$

quando o efeito de $X$ diminui à medida que $X$ cresce.

Observe que no modelo

$$Y = \beta_0 + \beta_1 \log X,$$

o coeficiente $\beta_1$ representa a variação média em $Y$ associada a uma variação proporcional em $X$. Mais especificamente, um aumento percentual em $X$ está associado a uma variação aproximadamente linear em $Y$.

Portanto, a transformação logarítmica em $X$ não implica mudança na distribuição de $Y$.

**Gráfico sugerido:**

- Resíduos vs. $X$ (o que verificar: curva sistemática; resultado esperado após transformação: dispersão aleatória).

### 9.3.3 Inverso ($X^* = 1/X$)

A transformação inversa é adequada quando o efeito de $X$ é muito forte para valores pequenos e diminui rapidamente conforme $X$ aumenta.

Exemplo típico:

- fenômenos de aprendizado;
- processos de saturação;
- respostas que se estabilizam para grandes valores de $X$.

Se a relação verdadeira for aproximadamente

$$Y = \alpha + \frac{\beta}{X},$$

então

$$Y = \alpha + \beta X^*$$

com $X^* = 1/X$.

A transformação inversa captura comportamentos de **retorno decrescente** ou aproximação assintótica.

**Gráfico sugerido:**

- Resíduos vs. $X$ (o que verificar: curva decrescente; resultado esperado após transformação: sem padrão).

### 9.3.4 Potências e termos polinomiais ($X^2$, $X^3$, ...)

Quando a relação apresenta curvatura suave, uma alternativa é expandir o modelo com termos polinomiais:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \varepsilon.$$

Esse modelo continua sendo linear nos parâmetros (condição essencial do MRLS), embora não seja linear em $X$.

Temos então que:

- $\beta_2 > 0 \rightarrow$ concavidade para cima (convexidade).
- $\beta_2 < 0 \rightarrow$ concavidade para baixo.

Apesar de introduzir não linearidade em $X$, o modelo permanece linear em relação aos parâmetros $\beta_j$. Isso garante que:

- os estimadores de MQO continuam válidos;
- as propriedades inferenciais permanecem as mesmas.

Essa distinção entre **linearidade nos parâmetros** e **linearidade na variável explicativa** é conceitualmente central em regressão linear.

**Gráfico sugerido:**

- Resíduos vs. $X$ (o que verificar: arco ou S-curva; resultado esperado após transformação: resíduos aleatórios).

### 9.3.5 Transformar $X$ ou adicionar polinômios?

Há duas estratégias principais para lidar com curvaturas:

1. Transformar $X$ (por exemplo, $\log X$, $1/X$).
2. Incluir termos polinomiais ($X^2$, $X^3$).

A escolha depende de:

- coerência teórica;
- interpretação desejada;
- estabilidade numérica do ajuste.

Transformações simples costumam ter interpretação mais direta; polinômios oferecem maior flexibilidade, mas podem introduzir instabilidade em extrapolações.

Logo, as transformações na variável explicativa:

- buscam linearizar a função média;
- não alteram a estrutura probabilística da resposta;
- preservam a estrutura básica do MRLS;
- exigem reavaliação completa dos resíduos após o ajuste.

## 9.4 Guia prático: escolhendo a transformação

Após compreender as motivações teóricas para transformar $Y$ ou $X$, surge a questão prática: **como decidir qual transformação aplicar?**

A escolha não deve ser arbitrária nem baseada apenas em melhora numérica de $R^2$. O critério central é a adequação às hipóteses do modelo e à estrutura substantiva do fenômeno (Montgomery, Peck, e Vining (2021); Kutner et al. (2005)).

**Etapa 1 — Diagnóstico inicial**

Antes de qualquer transformação, ajusta-se o modelo na escala original:

$$Y_i = \beta_0 + \beta_1 X_i + e_i.$$

Em seguida, analisam-se os resíduos:

- Resíduos vs. ajustados $\rightarrow$ verificar homoscedasticidade.
- Resíduos vs. $X \rightarrow$ verificar linearidade.
- QQ-plot $\rightarrow$ verificar normalidade.

- Histograma → avaliar assimetria e caudas.

A transformação deve responder a um padrão empírico específico observado nos resíduos. Se não há padrão sistemático, não há justificativa para transformar.

**Etapa 2 — Identificação do padrão**

Alguns padrões típicos e suas possíveis soluções:

| Padrão observado | Possível causa | Transformação sugerida |
|---|---|---|
| Funil (variância cresce com média) | $Var(Y \mid X)$ proporcional à média ou ao quadrado da média | $\log(Y)$ ou $\sqrt{Y}$ |
| Curvatura em arco | Relação não linear entre $Y$ e $X$ | $X^2$ ou $\log(X)$ |
| Assimetria à direita | Distribuição assimétrica positiva | $\log(Y)$ |
| Efeito forte em valores pequenos de $X$ | Retorno decrescente | $1/X$ |

Cada transformação responde a um mecanismo estrutural distinto. Não se trata de "corrigir o gráfico", mas de modelar melhor o processo subjacente.

**Etapa 3 — Ajuste do modelo transformado**

Após escolher a transformação, ajusta-se o novo modelo, por exemplo:

$$\log(Y_i) = \beta_0^* + \beta_1^* X_i + \varepsilon_i,$$

ou

$$Y_i = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i.$$

Nesse momento, todo o ciclo de verificação deve ser repetido:

- análise gráfica dos resíduos;
- testes formais (quando apropriado);
- avaliação da coerência interpretativa.

**Etapa 4 — Avaliação comparativa**

A comparação entre modelos deve considerar:

1. Qualidade dos resíduos (homoscedasticidade, normalidade).

2. Estabilidade dos coeficientes.
3. Interpretação substantiva.
4. Intervalos de confiança e precisão inferencial.

Um modelo com menor $R^2$ pode ser preferível se satisfaz melhor as hipóteses do MRLS e apresenta resíduos estruturalmente adequados.

## 9.5 Exemplo ilustrativo: antes e depois da transformação

Para consolidar as ideias discutidas até aqui, apresentamos um exemplo ilustrativo que mostra, de forma integrada, a **necessidade estatística** e a **eficácia inferencial** de uma transformação na variável resposta.

### 9.5.1 Cenário e especificação dos modelos

Suponha que estejamos interessados em modelar a relação entre **horas de prática de estudo** ($X$) e a **nota obtida em uma prova** ($Y$). Intuitivamente, espera-se que estudantes que dedicam mais horas de estudo tendam a alcançar notas maiores. Contudo, é plausível que:

- o ganho médio de desempenho seja aproximadamente exponencial;
- a variabilidade das notas aumente à medida que $X$ aumenta.

Essa estrutura implica que a variância condicional pode depender do nível médio da resposta, violando a hipótese de homoscedasticidade do MRLS.

Consideremos dois modelos concorrentes.

**Modelo A (sem transformação):**

$$Y_i = \beta_0 + \beta_1 X_i + e_i.$$

Esse modelo assume:

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i, \qquad Var(Y_i \mid X_i) = \sigma^2.$$

**Modelo B (com transformação logarítmica):**

$$Y_i^* = \log(Y_i) = \beta_0^* + \beta_1^* X_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Nesse caso, assumimos normalidade e homoscedasticidade **na escala logarítmica**. Implicitamente, isso significa que, na escala original,

$$Y_i \mid X_i \sim \text{LogNormal}(\mu_i, \sigma^2),$$

o que altera a estrutura probabilística do modelo.

### 9.5.2 Geração do conjunto de dados e análise descritiva

Para ilustrar essa situação, considere que a relação verdadeira é linear na escala logarítmica:

$$\log(Y_i) = 1.0 + 0.20X_i + u_i, \quad u_i \sim N(0, \sigma^2).$$

A partir dessa equação, obtemos

$$Y_i = \exp(1.0 + 0.20X_i + u_i),$$

o que gera:

- crescimento exponencial na média;
- variância crescente com $X$;
- assimetria à direita na escala original.

```
library(dplyr)
set.seed(2025)

n <- 120
X <- seq(0, 10, length.out = n)

# Relação linear na escala log(Y)
sigma <- 0.6
logY  <- 1.0 + 0.20*X + rnorm(n, 0, sigma)
Y     <- exp(logY)

df <- tibble(
  X = X,
  Y = Y,
  logY = log(Y)
)
```

#### Estatísticas descritivas

Devem ser apresentadas:

- primeiras linhas da base;
- estatísticas resumidas para $X$, $Y$ e $\log(Y)$.

```
head(df)
```

```
# A tibble: 6 x 3
       X     Y  logY
   <dbl> <dbl> <dbl>
1 0       3.95 1.37
2 0.0840  2.82 1.04
3 0.168   4.47 1.50
4 0.252   6.13 1.81
5 0.336   3.63 1.29
6 0.420   2.68 0.986
```

```
summary(df)
```

```
       X               Y               logY       
 Min.   : 0.0   Min.   : 1.141   Min.   :0.1318  
 1st Qu.: 2.5   1st Qu.: 4.384   1st Qu.:1.4779  
 Median : 5.0   Median : 6.833   Median :1.9217  
 Mean   : 5.0   Mean   : 9.724   Mean   :1.9833  
 3rd Qu.: 7.5   3rd Qu.:12.803   3rd Qu.:2.5497  
 Max.   :10.0   Max.   :40.396   Max.   :3.6987  
```

### Ajuste dos modelos

Ajustamos ambos os modelos via MQO:

- Modelo A: $Y \sim X$
- Modelo B: $\log(Y) \sim X$

Comparar:

- estimativas dos coeficientes;
- erros-padrão;
- $R^2$;
- estatísticas $t$ e $F$.

Note que os $R^2$ não são diretamente comparáveis entre escalas distintas, pois medem proporções de variabilidade em espaços diferentes.

```r
mA <- lm(Y ~ X, data = df)
mB <- lm(logY ~ X, data = df)

cat("=== MODELO A (Y ~ X) ===\n")
```

```
=== MODELO A (Y ~ X) ===
```

```r
print(summary(mA))
```

```
Call:
lm(formula = Y ~ X, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8050  -4.1989  -0.8159   1.8826  24.5691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0801     1.2119   1.716   0.0887 .
X             1.5288     0.2095   7.298 3.68e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.68 on 118 degrees of freedom
Multiple R-squared:  0.311, Adjusted R-squared:  0.3052
F-statistic: 53.27 on 1 and 118 DF,  p-value: 3.677e-11
```

```r
cat("\n=== MODELO B (logY ~ X) ===\n")
```

```

=== MODELO B (logY ~ X) ===
```

```r
print(summary(mB))
```

```
Call:
lm(formula = logY ~ X, data = df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.43792 -0.41121 -0.01437  0.36590  1.62093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15379    0.10872  10.612  < 2e-16 ***
X            0.16590    0.01879   8.828 1.16e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5992 on 118 degrees of freedom
Multiple R-squared:  0.3978,    Adjusted R-squared:  0.3927
F-statistic: 77.94 on 1 and 118 DF,  p-value: 1.164e-14
```

### ANOVA

Apresentar as tabelas ANOVA para ambos os modelos.

```r
cat("\n=== ANOVA - MODELO A ===\n")
```

```
=== ANOVA - MODELO A ===
```

```r
print(anova(mA))
```

```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X           1 2376.6 2376.59  53.266 3.677e-11 ***
Residuals 118 5264.9   44.62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n=== ANOVA - MODELO B ===\n")
```

```
=== ANOVA - MODELO B ===
```

```r
print(anova(mB))
```

```
Analysis of Variance Table

Response: logY
           Df Sum Sq Mean Sq F value    Pr(>F)
X           1 27.984 27.9837  77.938 1.164e-14 ***
Residuals 118 42.368  0.3591
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 9.5.3 Diagnóstico gráfico: antes e depois

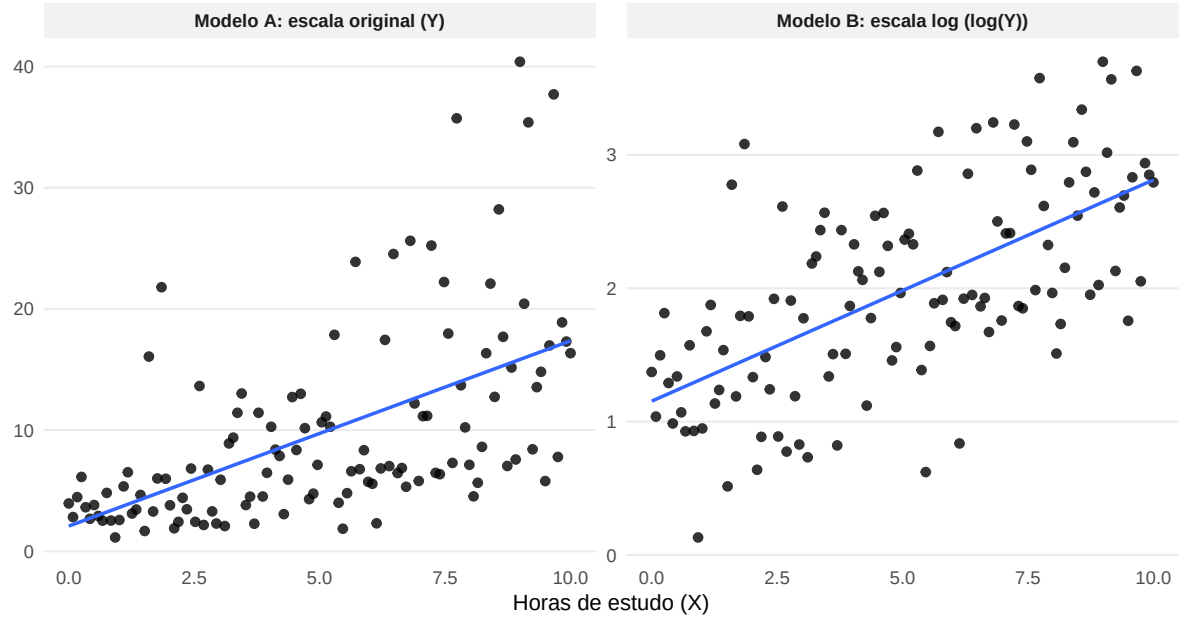
O contraste entre os modelos deve ser avaliado por meio dos gráficos diagnósticos.

```r
# Padronização visual para TODOS os gráficos do exemplo
library(ggplot2)
library(dplyr)
library(tidyr)

base_size <- 12

theme_diag <- theme_minimal(base_size = base_size) +
  theme(
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank(),
    strip.text = element_text(face = "bold"),
    strip.background = element_rect(fill = "grey95", color = NA),
    plot.title.position = "plot",
    plot.caption.position = "plot"
  )
```

#### 9.5.3.1 Dispersão da resposta vs. explicativa

- Modelo A: espera-se tendência crescente com dispersão aumentando (funil).
- Modelo B: dispersão aproximadamente constante e relação linear clara na escala transformada.

```r
df_plot <- df %>%
  select(X, Y, logY) %>%
  pivot_longer(cols = c(Y, logY), names_to = "resp", values_to = "valor") %>%
  mutate(
    resp = ifelse(resp == "Y",
                  "Modelo A: escala original (Y)",
                  "Modelo B: escala log (log(Y))")
  )

ggplot(df_plot, aes(x = X, y = valor)) +
  geom_point(alpha = 0.8, size = 2) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 0.9) +
  facet_wrap(~resp, ncol = 2, scales = "free_y") +
  labs(
    x = "Horas de estudo (X)",
    y = NULL
  ) +
  theme_diag
```



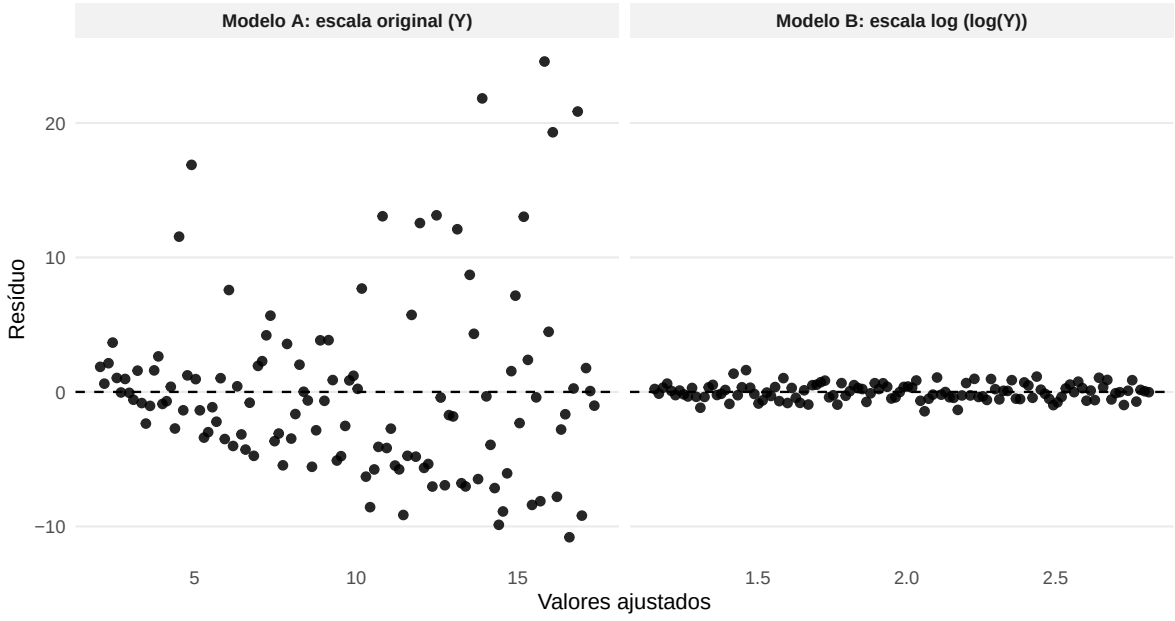

#### 9.5.3.2 Resíduos vs. valores ajustados

- Modelo A: padrão de funil e possível curvatura.
- Modelo B: resíduos aleatórios em torno de zero.

A ausência de estrutura no segundo caso indica que

$$Var(\varepsilon_i \mid X_i) \approx \sigma^2.$$

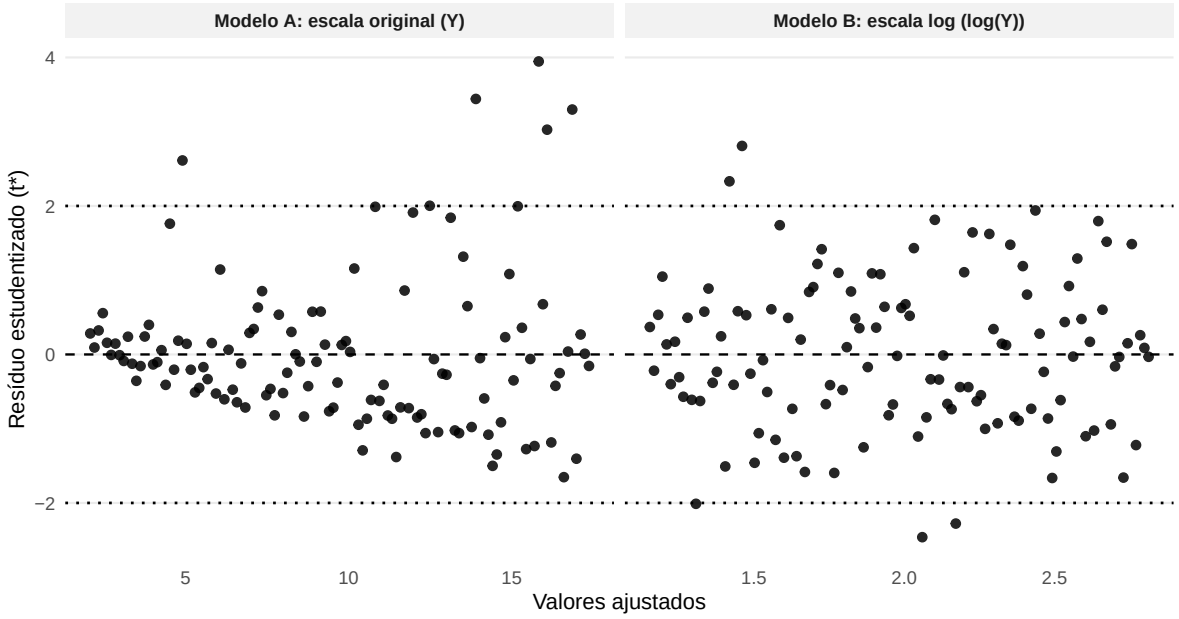
```
resA <- resid(mA); fitA <- fitted(mA)
resB <- resid(mB); fitB <- fitted(mB)

dados <- bind_rows(
  tibble(fit = fitA, res = resA, painel = "Modelo A: escala original (Y)"),
  tibble(fit = fitB, res = resB, painel = "Modelo B: escala log (log(Y))")
)

ggplot(dados, aes(x = fit, y = res)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~painel, ncol = 2, scales = "free_x") +
  labs(
    x = "Valores ajustados",
    y = "Resíduo"
  ) +
  theme_diag
```



#### 9.5.3.3 Resíduos estudentizados vs. ajustados

- Modelo A: maior número de pontos fora do intervalo $[-2, 2]$.
- Modelo B: poucos pontos extremos, ausência de padrão.

```
studA <- rstudent(mA)
studB <- rstudent(mB)

dados <- bind_rows(
  tibble(fit = fitA, stud = studA, painel = "Modelo A: escala original (Y)"),
  tibble(fit = fitB, stud = studB, painel = "Modelo B: escala log (log(Y))")
)

ggplot(dados, aes(x = fit, y = stud)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  geom_hline(yintercept = c(-2, 2), linetype = "dotted", linewidth = 0.7) +
  facet_wrap(~painel, ncol = 2, scales = "free_x") +
  labs(
    x = "Valores ajustados",
    y = "Resíduo estudentizado (t*)"
  ) +
  theme_diag
```



#### 9.5.3.4 QQ-plot

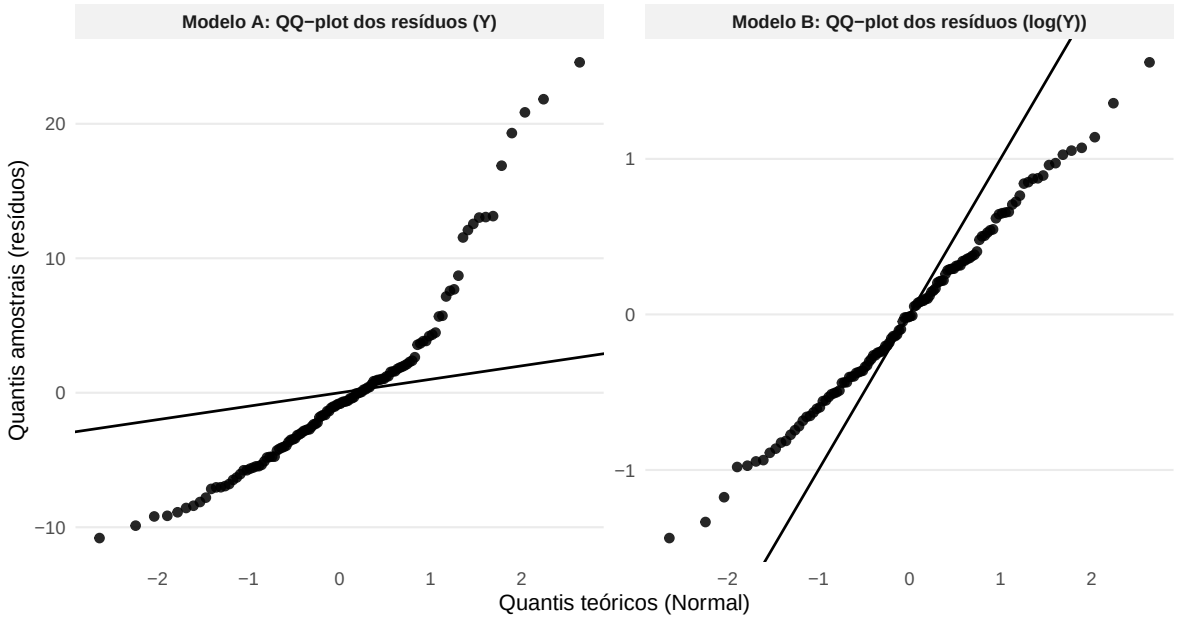
- Modelo A: desvios sistemáticos nas caudas.
- Modelo B: alinhamento próximo à reta de 45°.

Isso sugere que

$$\varepsilon_i \sim N(0, \sigma^2)$$

é plausível apenas na escala logarítmica.

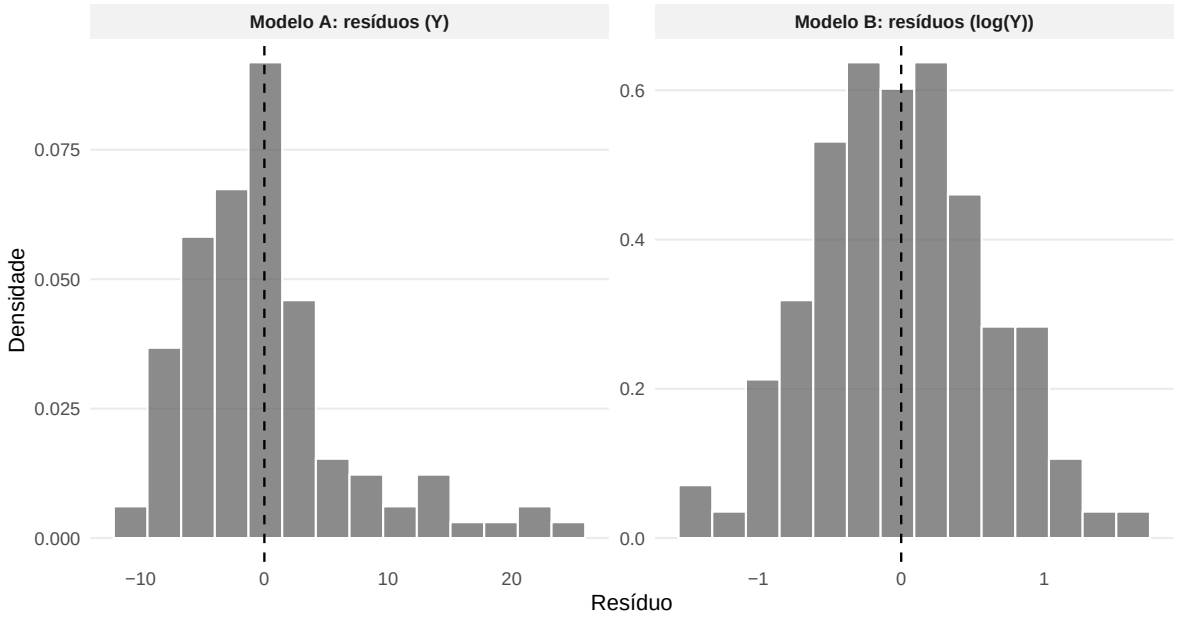
```
qq_df <- function(r, painel){
  q <- qqnorm(r, plot.it = FALSE)
  tibble(theo = q$x, samp = q$y, painel = painel)
}

dados <- bind_rows(
  qq_df(resA, "Modelo A: QQ-plot dos resíduos (Y)"),
  qq_df(resB, "Modelo B: QQ-plot dos resíduos (log(Y))")
)

ggplot(dados, aes(x = theo, y = samp)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_abline(intercept = 0, slope = 1, linewidth = 0.7) +
  facet_wrap(~painel, ncol = 2, scales = "free") +
  labs(
    x = "Quantis teóricos (Normal)",
    y = "Quantis amostrais (resíduos)"
  ) +
  theme_diag
```




#### 9.5.3.5 Histograma

- Modelo A: assimetria à direita e caudas pesadas.
- Modelo B: simetria aproximadamente normal.

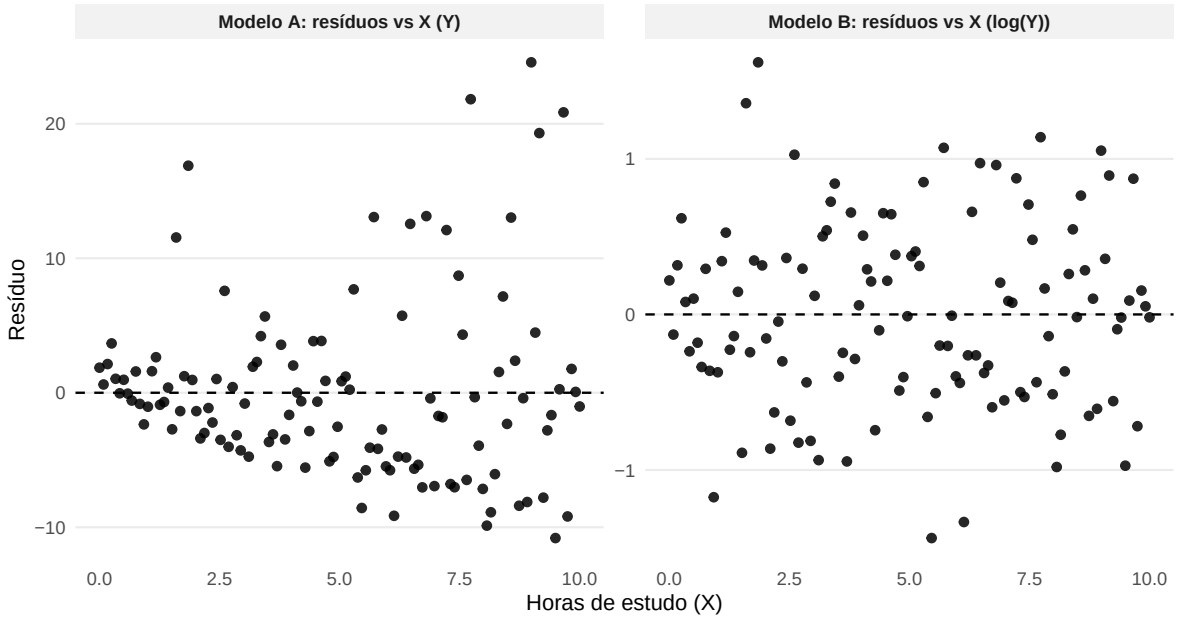
```
dados <- bind_rows(
  tibble(r = resA, painel = "Modelo A: resíduos (Y)"),
  tibble(r = resB, painel = "Modelo B: resíduos (log(Y))")
)

ggplot(dados, aes(x = r)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 14,
                 alpha = 0.7,
                 color = "white") +
  geom_vline(xintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~painel, ncol = 2, scales = "free") +
  labs(
    x = "Resíduo",
    y = "Densidade"
  ) +
  theme_diag
```




#### 9.5.3.6 Resíduos vs. $X$

- Modelo A: indícios de curvatura e variância crescente.
- Modelo B: dispersão uniforme.

```
dados <- bind_rows(
  tibble(X = df$X, res = resA, painel = "Modelo A: resíduos vs X (Y)"),
  tibble(X = df$X, res = resB, painel = "Modelo B: resíduos vs X (log(Y))")
)

ggplot(dados, aes(x = X, y = res)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~painel, ncol = 2, scales = "free_y") +
  labs(
    x = "Horas de estudo (X)",
    y = "Resíduo"
  ) +
  theme_diag
```




### 9.5.4 Comparação numérica e conclusão

**Modelo A (escala original)**

- $R^2$ aparentemente satisfatório.
- Violação da homoscedasticidade.
- Desvios de normalidade.
- Inferência potencialmente comprometida.

**Modelo B (escala logarítmica)**

- Melhor adequação estrutural dos resíduos.
- Testes $t$ e $F$ mais confiáveis.
- Intervalos de confiança coerentes com as hipóteses do modelo.

Um modelo pode apresentar bom $R^2$ e, ainda assim, violar hipóteses fundamentais. A validade da inferência depende da adequação dos resíduos, não apenas do poder explicativo.

#### 9.5.4.1 Interpretação

No Modelo B,

$$\log(Y) = \beta_0^* + \beta_1^* X.$$

Logo,

$$Y = \exp(\beta_0^*) \exp(\beta_1^* X).$$

O efeito de $X$ é **multiplicativo**:

$$\exp(\beta_1^*)$$

representa o fator pelo qual $Y$ é multiplicado para cada unidade adicional em $X$.

Entretanto, para previsões na escala original,

$$E(Y \mid X) = \exp\left(\mu + \tfrac{1}{2}\sigma^2\right),$$

e não simplesmente $\exp(\mu)$ (Casella e Berger (2002)). Ignorar esse termo introduz **viés de retransformação**

Este exemplo evidencia que:

- a análise de resíduos é central na validação do modelo;
- transformações podem corrigir violações estruturais;
- a interpretação deve sempre respeitar a escala do modelo ajustado;
- a escolha da transformação deve ser guiada por diagnóstico e fundamentação teórica.

Em regressão, a forma funcional adequada é aquela que torna os resíduos compatíveis com as hipóteses do modelo — não necessariamente aquela que produz o maior $R^2$.

# 10 Comparação de Modelos

## 10.1 Motivação e princípios de comparação

Após discutir a análise de resíduos, surge uma questão natural: **como escolher entre diferentes modelos candidatos?** Nem sempre um único modelo é suficiente; muitas vezes ajustamos **várias alternativas** e precisamos compará-las.

A comparação de modelos, porém, não é um "campeonato de métricas". Ela deve ser entendida como uma decisão estatística guiada por três ideias centrais: **adequação**, **parcimônia** e **finalidade** (explicar vs prever). Em particular, comparar modelos significa avaliar **trocas (trade-offs)**: um modelo pode ajustar melhor os dados, mas à custa de maior complexidade e menor estabilidade, especialmente em amostras moderadas/pequenas.

A comparação de modelos envolve duas dimensões principais:

1. **Qualidade estatística do ajuste** – medida por critérios formais (ex.: $R^2$, teste $F$, AIC, BIC).

2. **Pertinência substantiva** – se o modelo faz sentido em relação ao fenômeno estudado, é parcimonioso e interpretável.

Em regressão, "modelo melhor" **não significa** "modelo com o maior ajuste numérico". O ponto é: um modelo é um **compromisso** entre (i) representar a estrutura sistemática de $Y$ explicada por $X$ e (ii) não incorporar estrutura espúria (ruído) como se fosse sinal. Critérios como AIC e BIC foram propostos justamente para explicitar essa troca: melhorar o ajuste (via log-verossimilhança) **custa** complexidade (número de parâmetros), e essa penalização é uma maneira de desencorajar sobreajuste (Akaike (1974); Schwarz (1978); Burnham e Anderson (2002)).

Um ponto central é que **não adianta comparar dois modelos se ambos têm resíduos inadequados**. O primeiro filtro deve ser sempre o diagnóstico dos resíduos. Uma vez que os modelos estejam pelo menos aproximadamente bem especificados, podemos partir para critérios comparativos.

Essa ordem ("diagnóstico $\rightarrow$ comparação") é importante porque muitos critérios formais assumem que o modelo está **minimamente compatível** com as hipóteses estruturais do MRLS (por exemplo: relação média aproximadamente linear na escala adotada, variância

aproximadamente constante e ausência de padrões sistemáticos evidentes nos resíduos). Se essas condições falham, é comum observar situações enganosas como:

- um modelo com $R^2$ alto, mas com **padrão em funil** (heteroscedasticidade), o que compromete inferência usual e previsões com incerteza mal calibrada;
- um modelo com ligeira melhora em AIC/BIC após adicionar termos, mas com **curvatura residual persistente**, sinalizando que a forma funcional ainda está incorreta;
- um modelo "mais complexo" que parece melhor no ajuste dentro da amostra, mas piora a capacidade de generalização (sobreajuste).

### 10.1.1 Finalidade da comparação: explicar ou prever

Antes de escolher um critério, é útil explicitar o objetivo:

- **Explicação/interpretação:** prioriza parâmetros estáveis e interpretáveis e tende a valorizar parcimônia (frequentemente BIC é usado como referência em seleção mais "conservadora").

- **Predição:** prioriza desempenho fora da amostra; critérios baseados em verossimilhança com penalização moderada (como AIC) são frequentemente usados como aproximações de desempenho preditivo, mas idealmente devem ser complementados por validação (ex.: validação cruzada) quando isso fizer parte do desenho analítico (Burnham e Anderson (2002)).

Nesta seção, organizamos os principais critérios formais e práticos de comparação, seguidos de exemplos ilustrativos.

## 10.2 Critérios clássicos de comparação

### 10.2.1 Coeficiente de determinação $R^2$

O $R^2$ mede a proporção da variabilidade de $Y$ explicada pelo modelo:

$$R^2 = 1 - \frac{SQ_{Res}}{SQ_{Tot}},$$

em que $SQ_{Res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ é a soma de quadrados dos resíduos e $SQ_{Tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ é a soma de quadrados total.

O $R^2$ responde à pergunta: "**quanto da variabilidade total observada em $Y$ foi capturada pelo componente sistemático do modelo?**". Ele é, portanto, uma medida **descritiva de ajuste**: compara o erro do modelo com o erro de um "modelo nulo" que sempre prediz $\bar{Y}$.

Como a decomposição

$$SQ_{Tot} = SQ_{Reg} + SQ_{Res}$$

vale no MRLS com intercepto, também podemos escrever

$$R^2 = \frac{SQ_{Reg}}{SQ_{Tot}},$$

isto é, a fração da variabilidade total explicada pela regressão.

#### 10.2.1.1 Propriedades e limitações

É tentador ler $R^2$ como "o modelo é bom/ruim", mas isso é perigoso por três razões:

1. $R^2$ **não diagnostica adequação das hipóteses**: um modelo pode ter $R^2$ alto e ainda assim apresentar resíduos com curvatura, heteroscedasticidade ou dependência. Por isso, o diagnóstico de resíduos vem antes.
2. $R^2$ **não mede capacidade preditiva fora da amostra**: ele é calculado na amostra usada no ajuste.
3. $R^2$ **depende da variabilidade de** $Y$: em bases com pouca variação em $Y$, mesmo modelos úteis podem ter $R^2$ modestos; e em bases com grande variação em $Y$, $R^2$ pode ser alto sem que o modelo seja substantivamente satisfatório.

**Vantagens e limitaçoes**

- **Vantagem:** fornece uma medida intuitiva de ajuste.

- **Limitação 1 (monotonicidade):** em modelos de MQO com intercepto, $R^2$ **nunca diminui** quando adicionamos regressores ao conjunto de candidatos. Isso ocorre porque, ao adicionar parâmetros, o MQO minimiza $SQ_{Res}$ em um espaço maior, logo $SQ_{Res}$ só pode diminuir (ou permanecer igual).

- **Limitação 2 (comparabilidade):** $R^2$ **não é comparável** entre modelos com e sem intercepto, pois a decomposição de somas de quadrados muda. Em particular, quando o intercepto é omitido, o "modelo nulo" implícito deixa de ser $Y = \bar{Y}$, e interpretações usuais de $R^2$ podem se tornar enganosas.
- **Limitação 3 (escala da resposta):** se você transforma $Y$ (por exemplo, $\log(Y)$), o $R^2$ passa a descrever ajuste **na escala transformada**, não na escala original.

  **Observação:** no MRLS (com intercepto), vale a relação $R^2 = r^2_{XY}$, conectando a medida de ajuste com a intensidade de associação linear entre $X$ e $Y$.

### 10.2.2 Coeficiente de determinação ajustado $R^2_{aj}$

Para penalizar modelos excessivamente complexos, define-se:

$$R^2_{aj} = 1 - \frac{SQ_{Res}/(n-p)}{SQ_{Tot}/(n-1)},$$

em que $n$ é o tamanho da amostra e $p$ é o número de parâmetros (incluindo o intercepto).

O $R^2_{aj}$ substitui "proporção de variância explicada" por uma comparação entre **variâncias estimadas**:

- $SQ_{Res}/(n-p)$ é o estimador de $\sigma^2$ (variância do erro) sob o modelo candidato;
- $SQ_{Tot}/(n-1)$ é a variância amostral de $Y$.

Assim, $R^2_{aj}$ pergunta: "**o quanto a variância residual estimada caiu em relação à variância total de $Y$, levando em conta quantos parâmetros eu usei para isso?**".

Ao contrário do $R^2$, o $R^2_{aj}$ **pode diminuir** quando adicionamos regressoras. Isso é desejável: se um novo termo reduz pouco o $SQ_{Res}$, a penalização por perder graus de liberdade pode dominar, sinalizando que o ganho de ajuste não compensa a complexidade.

#### 10.2.2.1 Vantagens e limitações

- **Vantagem:** permite comparar modelos com diferentes números de parâmetros, desde que estejam na mesma escala da resposta e com intercepto.
- **Limitação:** em amostras pequenas, ainda pode favorecer modelos com leve sobreajuste, pois sua penalização é relativamente moderada.
- **Uso prático:** quando a comparação envolve modelos com diferentes estruturas, $R^2_{aj}$ é preferível ao $R^2$ simples.

### 10.2.3 Teste $F$ para modelos aninhados

Quando um modelo é **caso particular de outro** (modelo restrito vs. modelo completo), podemos usar:

$$F = \frac{(SQ_{Res,\,restrito} - SQ_{Res,\,completo})/q}{SQ_{Res,\,completo}/(n-p)},$$

em que:

- $q$ é o número de restrições impostas (equivalentemente, o número de parâmetros "removidos" quando passamos do completo para o restrito);
- $p$ é o número de parâmetros no modelo completo (incluindo intercepto).

O teste $F$ para modelos aninhados avalia se a redução em $SQ_{Res}$ ao passarmos do modelo restrito para o completo é **grande o suficiente** para justificar os parâmetros adicionais.

- No numerador está o "ganho médio" de ajuste por restrição relaxada: $(SQ_{Res,restrito} - SQ_{Res,completo})/q$.
- No denominador está uma estimativa de $\sigma^2$ no modelo completo: $SQ_{Res,completo}/(n-p)$.

Se o ganho médio (numerador) é grande comparado ao ruído residual esperado (denominador), a estatística $F$ fica grande e rejeitamos o modelo restrito.

#### 10.2.3.1 Hipóteses e decisão

- **Hipóteses:**
  - $H_0$: as restrições são válidas (o modelo mais simples é suficiente).
  - $H_1$: pelo menos uma restrição é falsa (o modelo completo melhora o ajuste de forma relevante).
- **Regra:** rejeitar $H_0$ para valores grandes de $F$ (ou valor-p pequeno).
- **Uso:** útil para verificar se a inclusão de um termo (ou intercepto) melhora de forma estatisticamente significativa o ajuste.
- **Limitação:** só se aplica a modelos **aninhados** (isto é, quando o modelo restrito pode ser obtido impondo restrições lineares nos parâmetros do modelo completo).

Adicionalmente, mesmo quando o teste $F$ rejeita $H_0$, ainda é necessário verificar:

- se o termo adicional tem interpretação substantiva coerente;
- se a inclusão introduz instabilidade (por exemplo, poucos dados sustentando um termo);
- se o ganho é relevante na prática (e não apenas detectável por tamanho amostral).

## 10.3 Critérios de informação

Além dos critérios clássicos, que se baseiam em soma de quadrados e variâncias, temos medidas que incorporam explicitamente a ideia de **verossimilhança e parcimônia**. Esses critérios nascem de uma perspectiva mais geral de modelagem estatística, na qual o modelo é visto como uma aproximação para a distribuição geradora dos dados.

No contexto do MRLS com erros normais,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

a estimação por MQO coincide com a estimação por **máxima verossimilhança**. Assim, podemos usar a log-verossimilhança avaliada nos estimadores para comparar modelos.

### 10.3.1 Critério de Informação de Akaike (AIC)

$$AIC = -2\ell(\hat{\theta}) + 2p,$$

em que:

- $\ell(\hat{\theta})$ é a log-verossimilhança no ponto estimado;
- $p$ é o número de parâmetros estimados (incluindo $\sigma^2$, quando apropriado).

O AIC pode ser interpretado como uma aproximação para a **distância de Kullback–Leibler** entre o modelo candidato e o verdadeiro mecanismo gerador dos dados (Akaike (1974); Burnham e Anderson (2002)). Ele busca selecionar o modelo que, entre os candidatos, minimiza a perda de informação esperada.

A estrutura do critério explicita o compromisso:

- $-2\ell(\hat{\theta})$ recompensa melhor ajuste (maior verossimilhança);
- $2p$ penaliza complexidade.

Assim, modelos com mais parâmetros precisam "ganhar" log-verossimilhança suficiente para compensar a penalização.

**Interpretação prática**

- **Regra:** entre modelos comparáveis, preferir o de **menor AIC**.
- O AIC não tem interpretação absoluta; só faz sentido **comparativamente**.
- Diferenças pequenas (por exemplo, menores que 2 unidades) geralmente indicam que os modelos têm suporte semelhante nos dados (Burnham e Anderson (2002)).

O AIC não testa hipóteses do tipo $H_0$ vs. $H_1$. Ele não produz valor-p nem decisão "rejeita/não rejeita". É um critério de **seleção por informação**, não um teste clássico de significância.

Além disso, o AIC tende a favorecer modelos levemente mais complexos, especialmente em amostras pequenas. Em contextos de amostras reduzidas, versões corrigidas (como AICc) podem ser mais adequadas (ver Burnham e Anderson (2002)).

### 10.3.2 Critério Bayesiano de Schwarz (BIC)

$$BIC = -2\ell(\hat{\theta}) + p\log(n).$$

O BIC possui forma semelhante ao AIC, mas a penalização cresce com $\log(n)$. Assim:

- Para amostras grandes, $\log(n)$ pode ser bem maior que 2;
- A penalização por parâmetro torna-se mais severa à medida que $n$ aumenta.

O BIC pode ser interpretado como uma aproximação ao **log do fator de Bayes** sob certas condições assintóticas (Schwarz (1978)). Por isso, ele é frequentemente associado a uma perspectiva de **seleção do modelo "verdadeiro"** dentro do conjunto candidato.

**Interpretação prática**

- **Regra:** entre modelos comparáveis, preferir o de **menor BIC**.
- O BIC tende a ser mais conservador que o AIC.
- Em amostras grandes, pode penalizar fortemente modelos com muitos parâmetros.

### 10.3.3 Boas práticas e comparabilidade entre escalas

**AIC vs. BIC: qual usar?**

- **AIC:** mais orientado à predição e desempenho fora da amostra.

- **BIC:** mais orientado à identificação de uma estrutura "mais plausível" dentro do conjunto de candidatos.
- Compare apenas modelos ajustados na **mesma escala** da variável resposta.

A escolha depende do objetivo da análise. Em contextos aplicados, é comum reportar ambos e discutir a coerência entre eles.

> **Nota importante sobre AIC e BIC:**
> Embora possamos calcular AIC e BIC para qualquer modelo ajustado, **não é correto comparar diretamente** os valores de um modelo ajustado em $Y$ com outro ajustado em $\log(Y)$, por exemplo.
>
> Isso acontece porque a transformação da resposta altera a escala e a própria função de verossimilhança usada no cálculo dos critérios. Assim, os valores de AIC/BIC ficam em "bases diferentes".
>
> **Como proceder então?**
> - Compare AIC/BIC apenas entre modelos ajustados **na mesma escala da resposta**.
> - Use análise de resíduos e diagnóstico gráfico para avaliar se a transformação foi

benéfica.
- Para fins de **predição**, utilize métricas de erro calculadas na escala original de $Y$ (ex.: RMSE, MAE) para decidir qual modelo é mais adequado.

Em síntese: AIC e BIC são ferramentas valiosas, mas devem ser usados com critério. Transformações em $Y$ exigem avaliação adicional baseada em resíduos e desempenho preditivo.

### 10.3.4 Estratégia prática de comparação

Ao comparar modelos alternativos:

1. **Primeiro passo:** verifique os **resíduos** (condição mínima).
   - Se um modelo viola homocedasticidade ou apresenta estrutura sistemática, mesmo que tenha $R^2$ alto, não deve ser preferido.
2. **Segundo passo:** use os **critérios formais**.
   - Teste $F$ quando os modelos são aninhados.
   - $R^2_{aj}$ quando a comparação é dentro da mesma escala.
   - AIC/BIC para comparações amplas (mesma resposta, diferentes ajustes).
3. **Terceiro passo:** considere a **interpretação substantiva**.
   - Um modelo matematicamente melhor pode ser substantivamente inadequado.

## 10.4 Exemplos ilustrativos de comparação de modelos

Para consolidar as ideias, apresentamos dois exemplos didáticos que demonstram como comparar modelos no contexto do **MRLS**. Cada exemplo explora um tipo de decisão prática enfrentada por analistas de dados.

### 10.4.1 Exemplo 1 — Intercepto e transformação na resposta

**Simulação do Exemplo 1:** gerar `df1` com as variáveis `X` e `Y` (com $Y > 0$, para permitir $\log(Y)$).

```r
library(dplyr)

set.seed(2025)

# Simulação: consumo (Y) vs tempo de funcionamento (X)
n <- 50
X <- seq(0, 10, length.out = n)

# Consumo com intercepto > 0 (stand-by) e heterocedasticidade moderada
mu <- 5 + 1.2*X
sigma <- 0.6 + 0.05*X
Y <- mu + rnorm(n, 0, sigma)
Y <- pmax(Y, 0.1)  # garantir Y>0 (para log)

df1 <- tibble(X = X, Y = Y)
```

(i) **Cenário e modelos candidatos:** consumo de energia ($Y$) em função do tempo de funcionamento ($X$), com $n = 50$ observações.

**Modelos candidatos:**

1. Com intercepto

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

   Representa consumo mínimo (stand-by). Esperado em contextos onde $Y > 0$ mesmo quando $X = 0$.

2. Sem intercepto

$$Y = \beta_1 X + \varepsilon$$

   Força a reta a passar pela origem. Só faz sentido se sabemos que $Y = 0$ quando $X = 0$.

3. Transformado (log da resposta)

$$\log Y = \beta_0 + \beta_1 X + \varepsilon$$

Nesta configuração com variável transformada, o coeficiente $\beta_1$ indica a mudança **aditiva em** $\log(Y)$ a cada unidade adicional em $X$. Na escala original de $Y$, isso equivale a dizer que um aumento de 1 unidade em $X$ está associado a uma multiplicação esperada de $Y$ por $\exp(\beta_1)$.

Aqui estamos comparando três "ideias de modelo" que, apesar de parecidas na forma, respondem a perguntas ligeiramente diferentes:

- **Modelo A (com intercepto)** permite que exista um nível médio de consumo quando $X = 0$ (consumo residual/stand-by).

- **Modelo A0 (sem intercepto)** declara, como hipótese estrutural, que "se $X = 0$, então $Y = 0$". Essa é uma afirmação forte: se ela for falsa, o ajuste pode "compensar" deslocando a inclinação e gerando resíduos estruturados.

- **Modelo B (com $\log Y$)** muda a escala da resposta e, portanto, muda o tipo de erro "natural" no modelo (efeitos multiplicativos no $Y$ original tendem a virar efeitos aditivos em $\log(Y)$).

**Visualização da primeiras linhas da Base de dados**

```
head(df1)
```

```
# A tibble: 6 x 2
      X     Y
  <dbl> <dbl>
1 0      5.37
2 0.204  5.27
3 0.408  5.97
4 0.612  6.54
5 0.816  6.22
6 1.02   6.12
```

(i) **Estimação e resumos dos modelos**

**Ajustes e tabela de coeficientes:** ajustar `mod_A`, `mod_A0` e `mod_log`.

```
# Ajustes
mod_A   <- lm(Y ~ X, data = df1)       # com intercepto
mod_A0  <- lm(Y ~ 0 + X, data = df1)   # sem intercepto
mod_log <- lm(log(Y) ~ X, data = df1)  # resposta em log

cat("Modelo com intercepto (A):\n")
```

```
Modelo com intercepto (A):
```

```
print(coef(summary(mod_A)))
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 5.137603 0.24760614 20.74909 1.297563e-25
X           1.196581 0.04266949 28.04301 2.032694e-31
```

```
cat("\nModelo sem intercepto (A0):\n")
```

```
Modelo sem intercepto (A0):
```

```
print(coef(summary(mod_A0)))
```

```
  Estimate Std. Error t value     Pr(>|t|)
X 1.959437 0.06767437 28.9539 1.770922e-32
```

```
cat("\nModelo com log(Y) (B):\n")
```

```
Modelo com log(Y) (B):
```

```
print(coef(summary(mod_log)))
```

```
             Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 1.7791870 0.025029186 71.08450 2.673492e-50
X           0.1143677 0.004313232 26.51554 2.548074e-30
```

**Comentários:**
- O **Modelo A (com intercepto)** mostra um termo constante significativo, representando consumo em stand-by ($Y > 0$ em $x \approx 0$).
- O **Modelo A0 (sem intercepto)** ignora o consumo em stand-by e força a reta a passar pela origem.
- O **Modelo B (log da resposta)** reduz heteroscedasticidade: os coeficientes mantêm significância e a interpretação passa a ser na escala da variável transformada.

**Teste de resíduos (omnibus, Jarque-Bera, skew, kurtosis, Durbin-Watson, etc.)**

```
# Testes/medidas adicionais dos resíduos
  library(lmtest)   # dwtest
  library(tseries)  # jarque.bera.test
  library(moments)  # skewness, kurtosis

res_tests <- function(mod, nome){
  e <- resid(mod)
```

```r
  jb <- jarque.bera.test(e)
  sk <- skewness(e)
  ku <- kurtosis(e)  # kurtosis "crua" (Normal ~ 3)

  # Durbin-Watson (lmtest)
  dw <- dwtest(mod)

  # Omnibus D'Agostino-Pearson (quando disponível via fBasics::dagoTest)
  omni_out <- NULL
  if (requireNamespace("fBasics", quietly = TRUE)) {
    omni_out <- fBasics::dagoTest(e)
  }

  cat("\n==============================\n")
  cat("Testes dos resíduos -", nome, "\n")
  cat("==============================\n")

  cat("Jarque-Bera:\n")
  print(jb)

  cat("\nAssimetria (skewness): ", round(sk, 4), "\n", sep = "")
  cat("Curtose (kurtosis):   ", round(ku, 4), "  (Normal ~ 3)\n", sep = "")

  cat("\nDurbin-Watson (lmtest::dwtest):\n")
  print(dw)

  if (!is.null(omni_out)) {
    cat("\nOmnibus (D'Agostino-Pearson) - fBasics::dagoTest:\n")
    print(omni_out)
  } else {
    cat("\nOmnibus (D'Agostino-Pearson):\n")
    cat("Pacote 'fBasics' não disponível no ambiente. (Opcional: instalar para executar o om
  }
}

res_tests(mod_A,   "Modelo A (Y ~ X, com intercepto)")
```

```
==============================
Testes dos resíduos - Modelo A (Y ~ X, com intercepto)
==============================
Jarque-Bera:
```

```
    Jarque Bera Test

data:  e
X-squared = 0.39891, df = 2, p-value = 0.8192


Assimetria (skewness): 0.1998
Curtose (kurtosis):   2.8216  (Normal ~ 3)

Durbin-Watson (lmtest::dwtest):

    Durbin-Watson test

data:  mod
DW = 2.165, p-value = 0.6695
alternative hypothesis: true autocorrelation is greater than 0


Omnibus (D'Agostino-Pearson) - fBasics::dagoTest:

Title:
 D'Agostino Normality Test

Test Results:
  STATISTIC:
    Chi2 | Omnibus: 0.4194
    Z3  | Skewness: 0.6388
    Z4  | Kurtosis: 0.1065
  P VALUE:
    Omnibus  Test: 0.8108
    Skewness Test: 0.523
    Kurtosis Test: 0.9152
```

```
res_tests(mod_A0,  "Modelo A0 (Y ~ X, sem intercepto)")
```

```
==============================
Testes dos resíduos - Modelo A0 (Y ~ X, sem intercepto)
==============================
Jarque-Bera:
```

```
    Jarque Bera Test

data:  e
X-squared = 2.6548, df = 2, p-value = 0.2652


Assimetria (skewness): -0.0498
Curtose (kurtosis):   1.8755  (Normal ~ 3)

Durbin-Watson (lmtest::dwtest):

    Durbin-Watson test

data:  mod
DW = 0.22013, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0


Omnibus (D'Agostino-Pearson) - fBasics::dagoTest:

Title:
 D'Agostino Normality Test

Test Results:
  STATISTIC:
    Chi2 | Omnibus: 8.7587
    Z3   | Skewness: -0.1602
    Z4   | Kurtosis: -2.9552
  P VALUE:
    Omnibus  Test: 0.01253
    Skewness Test: 0.8728
    Kurtosis Test: 0.003125
```

```
res_tests(mod_log, "Modelo B (log(Y) ~ X)")
```

```
==============================
Testes dos resíduos - Modelo B (log(Y) ~ X)
==============================
Jarque-Bera:

    Jarque Bera Test
```

```
data:  e
X-squared = 3.1899, df = 2, p-value = 0.2029


Assimetria (skewness): 0.5779
Curtose (kurtosis):   3.4418  (Normal ~ 3)

Durbin-Watson (lmtest::dwtest):

    Durbin-Watson test

data:  mod
DW = 1.5831, p-value = 0.04934
alternative hypothesis: true autocorrelation is greater than 0


Omnibus (D'Agostino-Pearson) - fBasics::dagoTest:

Title:
 D'Agostino Normality Test

Test Results:
  STATISTIC:
    Chi2 | Omnibus: 4.236
    Z3  | Skewness: 1.7691
    Z4  | Kurtosis: 1.0518
  P VALUE:
    Omnibus  Test: 0.1203
    Skewness Test: 0.07688
    Kurtosis Test: 0.2929
```

**Comentários:**

- **Modelo A (com intercepto):**
  - O teste Omnibus e o Jarque-Bera não são significativos (valores-p altos), sugerindo que os resíduos não violam fortemente a normalidade.
  - O Durbin-Watson próximo de 2 indica ausência de autocorrelação serial.
  - Conclusão: resíduos adequados, modelo consistente.
- **Modelo A0 (sem intercepto):**

  - O Durbin-Watson é muito baixo ( 0,2), indicando forte autocorrelação positiva dos resíduos.

  - Embora Omnibus/Jarque-Bera não rejeitem a normalidade, a dependência serial torna o modelo problemático.

  - Conclusão: estatisticamente inadequado, reforçando que a exclusão do intercepto distorce o ajuste.

- **Modelo B (log da resposta):**

  - Testes de normalidade (Omnibus, JB) continuam não significativos, mas a assimetria (Skew negativo) sugere leve desvio.

  - O Durbin-Watson 1,4 aponta alguma autocorrelação positiva.

  - Conclusão: apesar das pequenas imperfeições, o modelo log-transformado melhora a homocedasticidade em relação ao Modelo A.

Em dados simulados sem mecanismo temporal explícito, autocorrelação forte geralmente é um **sinal de especificação inadequada** (por exemplo, restrições erradas como $\beta_0 = 0$ podem "organizar" os resíduos e induzir padrões). Em aplicações reais, autocorrelação também pode ocorrer por dependência temporal genuína; nesse caso, o MRLS pode precisar ser estendido (tema para capítulos posteriores).

(iii) **Modelos aninhados: teste F (A0 vs A)**

```
# Teste F (modelos aninhados): A0 (restrito) vs A (com intercepto)
cat("ANOVA (Teste F: A0 vs A):\n")
```

```
ANOVA (Teste F: A0 vs A):
```

```
print(anova(mod_A0, mod_A))
```

```
Analysis of Variance Table

Model 1: Y ~ 0 + X
Model 2: Y ~ X
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     49 377.84
2     48  37.90  1    339.94 430.52 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O modelo sem intercepto (A0) é um **caso particular** do modelo com intercepto (A), obtido ao impor a restrição $\beta_0 = 0$. Logo, o teste $F$ avalia se "forçar a origem" piora o ajuste além do esperado por acaso.

**Comentários:**
- O teste F compara os modelos **A0 (restrito)** e **A (com intercepto)**.
- O resultado altamente significativo (valor-p próximo de zero) indica que o **intercepto é necessário**.
- Assim, rejeitamos o modelo sem intercepto e preferimos o modelo com intercepto.

(iv) **Comparação numérica ($R^2$, $R^2_{aj}$, AIC, BIC)**

```
library(tidyverse)

cmp1 <- tibble(
  Modelo = c("A: Y~X (c/ intercepto)", "A0: Y~X (sem intercepto)",
             "B: log(Y)~X"),
  R2     = c(summary(mod_A)$r.squared,     summary(mod_A0)$r.squared,
             summary(mod_log)$r.squared),
  R2_aj  = c(summary(mod_A)$adj.r.squared, summary(mod_A0)$adj.r.squared,
             summary(mod_log)$adj.r.squared),
  AIC    = c(AIC(mod_A), AIC(mod_A0), AIC(mod_log)),
  BIC    = c(BIC(mod_A), BIC(mod_A0), BIC(mod_log))
)

cmp1
```

```
# A tibble: 3 x 5
  Modelo                      R2 R2_aj   AIC   BIC
  <chr>                    <dbl> <dbl> <dbl> <dbl>
1 A: Y~X (c/ intercepto)   0.942 0.941 134.  140.
2 A0: Y~X (sem intercepto) 0.945 0.944 247.  251.
3 B: log(Y)~X              0.936 0.935 -95.1 -89.4
```

Lembre-se que um $R^2$ elevado pode coexistir com **resíduos ruins**. Isso acontece porque $R^2$ mede "quanto o modelo explica" em termos de variabilidade total, mas não garante que as hipóteses do MRLS estejam razoavelmente satisfeitas. Em particular, um modelo pode ter alto $R^2$ e ainda assim produzir inferências pouco confiáveis (erros-padrão distorcidos).

**Comentários:**
- O **Modelo A0** apresenta $R^2$ elevado, mas penalizações via AIC/BIC mostram que é muito inferior (valores bem maiores). Logo, acredita-se que os coeficientes ficaram distorcidos, superestimando a inclinação.

- O **Modelo A** combina bom ajuste ($R^2_{aj}$ alto) e parcimônia.
- O **Modelo B (log)** apresenta os menores valores de AIC/BIC.

(v) **Diagnóstico gráfico comparativo**

**Dispersões com retas ajustadas dos dois modelos restantes**

```
suppressPackageStartupMessages({
  library(ggplot2)
  library(dplyr)
  library(tidyr)
})

dfA  <- df1 %>% mutate(valor = Y,        modelo = "A: Y ~ X (c/ intercepto)")
dfA0 <- df1 %>% mutate(valor = Y,        modelo = "A0: Y ~ X (sem intercepto)")
dfB  <- df1 %>% mutate(valor = log(Y),   modelo = "B: log(Y) ~ X")

df_plot <- bind_rows(dfA, dfA0, dfB)

ggplot(df_plot, aes(x = X, y = valor)) +
  geom_point(alpha = 0.9, size = 2) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  facet_wrap(~modelo, ncol = 3, scales = "free_y") +
  labs(x = "X", y = NULL) +
  theme_minimal(base_size = 12)
```

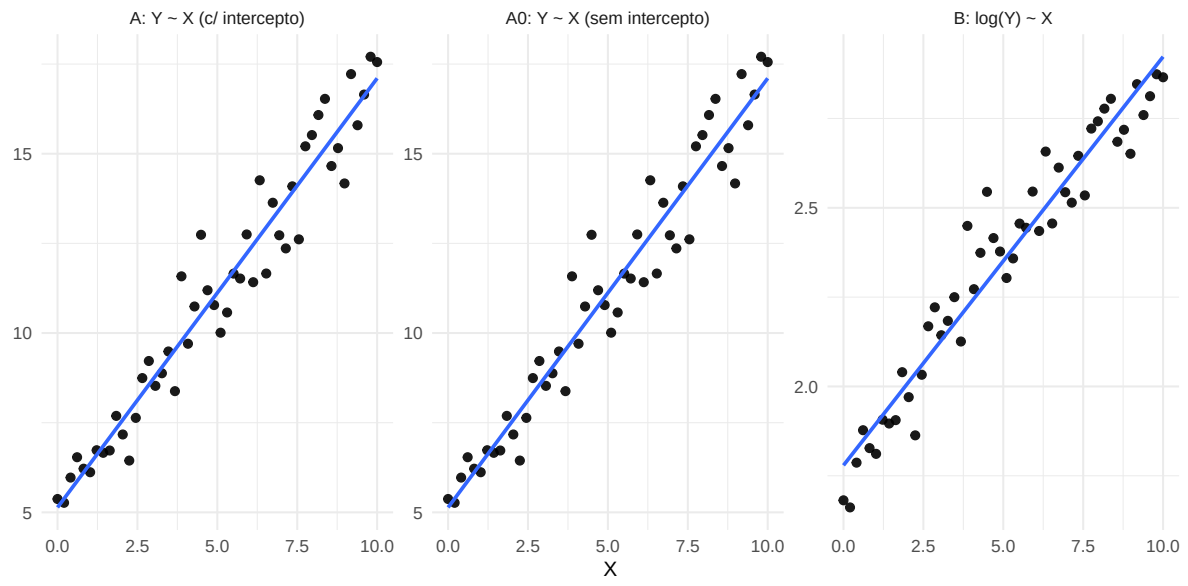



Figura 10.1: Exemplo 1 — Dispersão com reta OLS: Modelo A (Y~X, com intercepto), Modelo A0 (Y~X, sem intercepto) e Modelo B (log(Y)~X).

**Comentários:**
- Em ambos os **Modelos (A e B)**, não observa-se aumento relevante da variabilidade de $Y$ conforme $X$ cresce.

**Resíduos vs. valores ajustados**

```
  library(ggplot2)
  library(dplyr)

dados <- bind_rows(
  tibble(fit = fitted(mod_A),   res = resid(mod_A),   modelo = "A: Y ~ X (c/ intercepto)"),
  tibble(fit = fitted(mod_A0),  res = resid(mod_A0),  modelo = "A0: Y ~ X (sem intercepto)"),
  tibble(fit = fitted(mod_log), res = resid(mod_log), modelo = "B: log(Y) ~ X")
)

ggplot(dados, aes(x = fit, y = res)) +
  geom_point(alpha = 0.9, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~modelo, ncol = 3, scales = "free_x") +
  labs(x = "Ajustados", y = "Resíduo") +
  theme_minimal(base_size = 12)
```

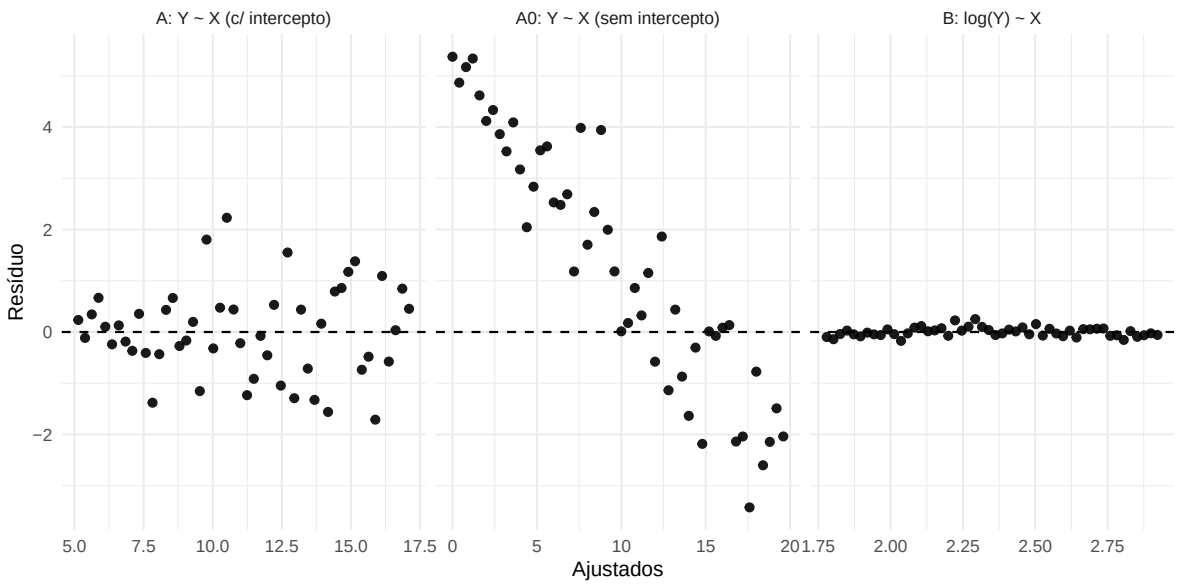



Figura 10.2: Exemplo 1 — Resíduos vs Ajustados: comparação entre A, A0 e B.

**Comentário:**
Neste gráfico, buscamos uma nuvem aproximadamente aleatória em torno de 0. Três estruturas são especialmente informativas:

- **Funil** (dispersão aumentando/diminuindo com o nível ajustado): sinal de **heteroscedasticidade**.
- **Curvatura** (padrão em arco): sinal de **forma funcional inadequada** (não linearidade não capturada).
- **Faixas** (bandas horizontais): podem surgir por discretização/limitação de medida em $Y$.

Na comparação entre modelos, o melhor candidato é o que *reduz* essas estruturas, sem criar novas.

**Resíduos estudentizados vs. valores ajustados**

```
  library(ggplot2)
  library(dplyr)

dados <- bind_rows(
  tibble(fit = fitted(mod_A),   stud = rstudent(mod_A),   modelo = "A: Y ~ X (c/ intercepto)"
  tibble(fit = fitted(mod_A0),  stud = rstudent(mod_A0),  modelo = "A0: Y ~ X (sem intercepto
  tibble(fit = fitted(mod_log), stud = rstudent(mod_log), modelo = "B: log(Y) ~ X")
```

```
)

ggplot(dados, aes(x = fit, y = stud)) +
  geom_point(alpha = 0.9, size = 2) +
  geom_hline(yintercept = 0,  linetype = "dashed", linewidth = 0.6) +
  geom_hline(yintercept = c(-2, 2), linetype = "dotted", linewidth = 0.6) +
  facet_wrap(~modelo, ncol = 3, scales = "free_x") +
  labs(x = "Ajustados", y = "t* (estudentizado)") +
  theme_minimal(base_size = 12)
```

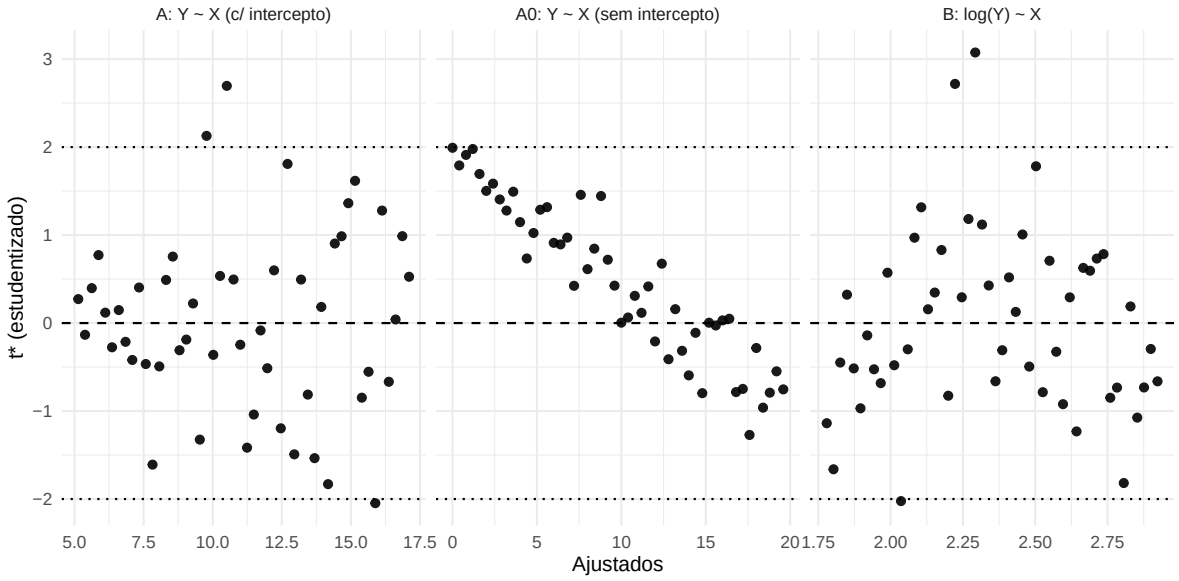



Figura 10.3: Exemplo 1 — Resíduos estudentizados vs Ajustados: comparação entre A, A0 e B.

**Comentário:**
Estes gráficos apresentam informações similares aos ilustradas nos gráficos anteriores

**QQ-plot dos resíduos**

```
library(ggplot2)
library(dplyr)

# dados do QQ-plot + parâmetros da reta ao estilo qqline (base R)
qq_df_line <- function(mod, modelo){
  r <- rstandard(mod)
  q <- qqnorm(r, plot.it = FALSE)
```

```r
  df <- tibble(theo = q$x, samp = q$y, modelo = modelo)

  # mesma regra do qqline: reta baseada nos quartis (25% e 75%)
  yq <- quantile(r, probs = c(0.25, 0.75), na.rm = TRUE)
  xq <- qnorm(c(0.25, 0.75))

  slope <- (yq[2] - yq[1])/(xq[2] - xq[1])
  intercept <- yq[1] - slope*xq[1]

  list(df = df, line = tibble(modelo = modelo, intercept = intercept, slope = slope))
}

outA  <- qq_df_line(mod_A,   "A: Y ~ X (c/ intercepto)")
outA0 <- qq_df_line(mod_A0,  "A0: Y ~ X (sem intercepto)")
outB  <- qq_df_line(mod_log, "B: log(Y) ~ X")

dados <- bind_rows(outA$df, outA0$df, outB$df)
linhas <- bind_rows(outA$line, outA0$line, outB$line)

ggplot(dados, aes(x = theo, y = samp)) +
  geom_point(alpha = 0.9, size = 2) +
  geom_abline(data = linhas, aes(intercept = intercept, slope = slope),
              linewidth = 0.8) +
  facet_wrap(~modelo, ncol = 3, scales = "free") +
  labs(x = "Quantis teoricos (Normal)",
       y = "Residuos padronizados") +
  theme_minimal(base_size = 12)
```

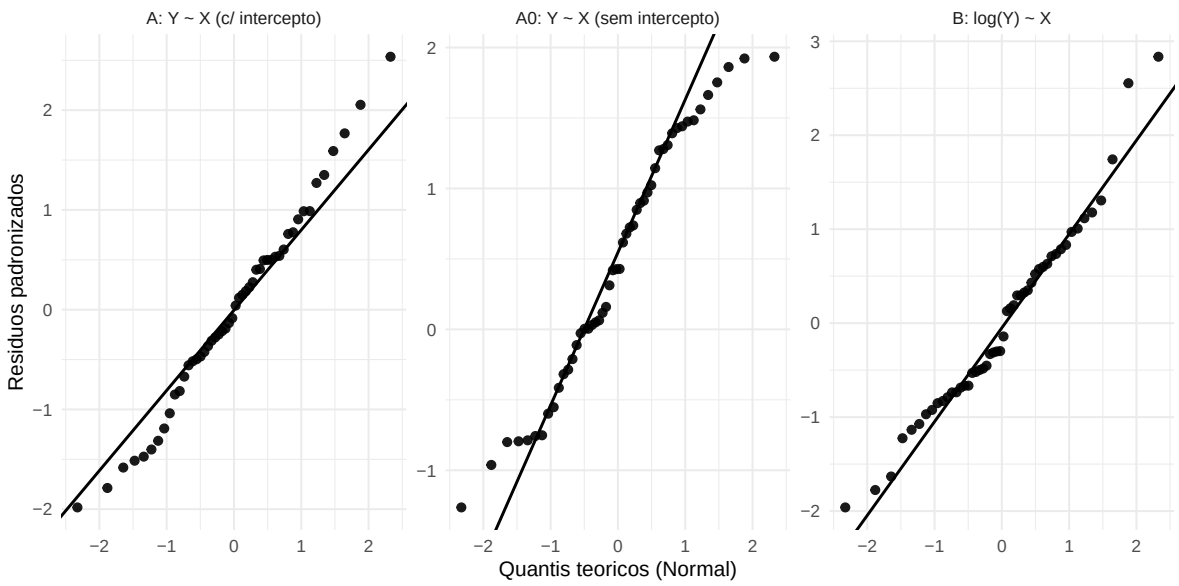



Figura 10.4: Exemplo 1 — QQ-plot (rstandard) com reta tipo qqline: comparação entre A, A0 e B.

**Comentário:**
O QQ-plot avalia *normalidade aproximada* por meio do alinhamento entre os quantis amostrais dos resíduos e os quantis teóricos da Normal. A interpretação deve ser feita por padrões:

- **Alinhamento global próximo da reta**: evidência visual a favor de resíduos aproximadamente normais (pelo menos no "miolo" da distribuição).
- **Desvios sistemáticos nas caudas** (pontos afastando-se da reta apenas no início e no fim): indicam **caudas mais pesadas ou mais leves** que a Normal; isso costuma afetar sobretudo inferência em amostras pequenas.
- **Padrão em "S"**: sugere **assimetria** (skewness diferente de 0).
- **Um ou poucos pontos muito afastados**: podem ser indício de **outliers** (verificar também resíduos estudentizados e Cook).

Ao comparar modelos, prefira o que apresenta **menos estrutura sistemática** no QQ-plot, especialmente quando isso é coerente com as medidas numéricas de assimetria/curtose e com testes como Jarque–Bera.

**Histograma dos resíduos**

```r
  library(ggplot2)
  library(dplyr)

dados <- bind_rows(
  tibble(r = resid(mod_A),   modelo = "A: Y ~ X (c/ intercepto)"),
  tibble(r = resid(mod_A0),  modelo = "A0: Y ~ X (sem intercepto)"),
  tibble(r = resid(mod_log), modelo = "B: log(Y) ~ X")
)

ggplot(dados, aes(x = r)) +
  geom_histogram(bins = 12, alpha = 0.7, color = "white") +
  facet_wrap(~modelo, ncol = 3, scales = "free") +
  labs(x = "Resíduo", y = "Frequência") +
  theme_minimal(base_size = 12)
```

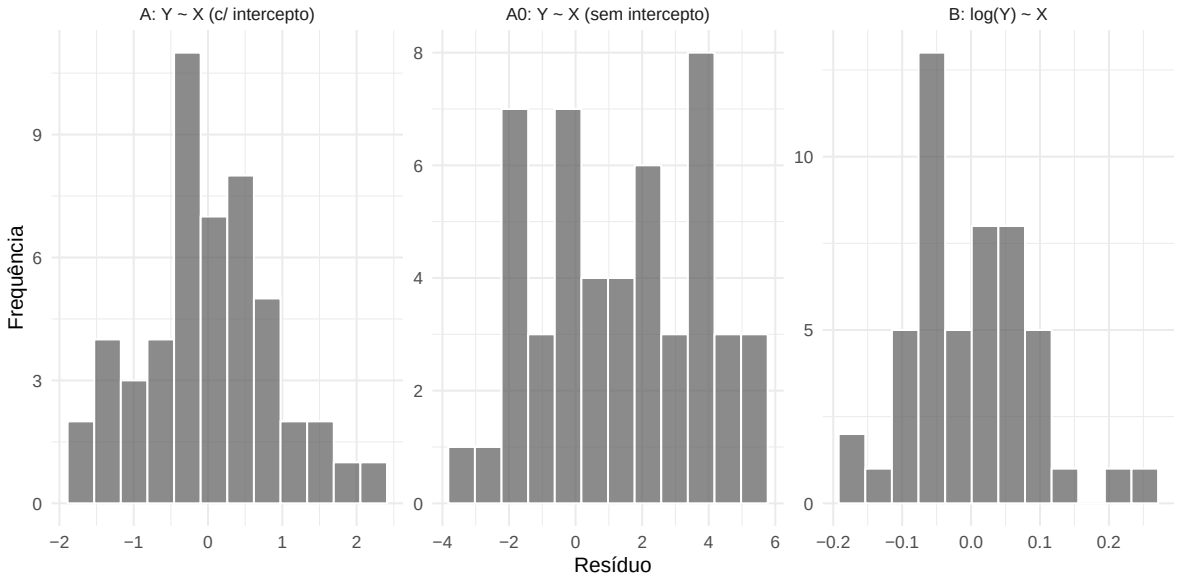



Figura 10.5: Exemplo 1 — Histograma dos resíduos: comparação entre A, A0 e B.

**Comentário:**

O histograma é um diagnóstico auxiliar: ele ajuda a visualizar **assimetria** e **caudas**. Em geral:

- histograma muito assimétrico sugere assimetria nos resíduos;

- caudas longas ou “ombros” podem sugerir caudas pesadas e/ou outliers.

A interpretação deve ser combinada com QQ-plot e com medidas como assimetria/curtose.

**Resíduos vs. X**

```
suppressPackageStartupMessages({
  library(ggplot2)
  library(dplyr)
})

dados <- bind_rows(
  tibble(X = df1$X, res = resid(mod_A),   modelo = "A: Y ~ X (c/ intercepto)"),
  tibble(X = df1$X, res = resid(mod_A0),  modelo = "A0: Y ~ X (sem intercepto)"),
  tibble(X = df1$X, res = resid(mod_log), modelo = "B: log(Y) ~ X")
)

ggplot(dados, aes(x = X, y = res)) +
  geom_point(alpha = 0.9, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~modelo, ncol = 3, scales = "free_y") +
  labs(x = "X", y = "Resíduo") +
  theme_minimal(base_size = 12)
```

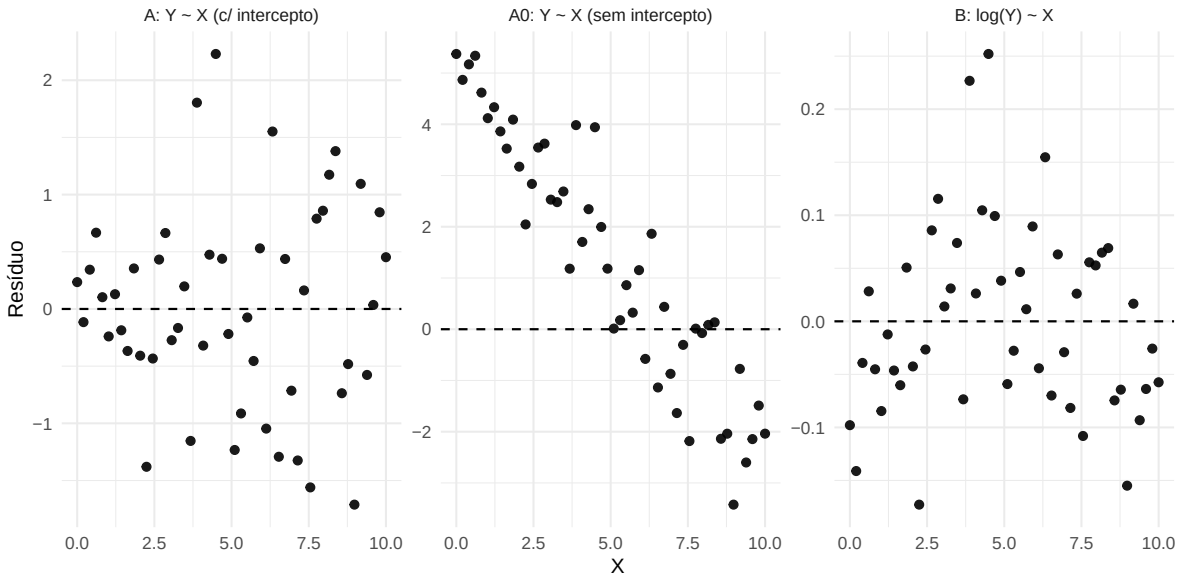



Figura 10.6: Exemplo 1 — Resíduos vs X: comparação entre A, A0 e B.

**Comentário:**
- Ambos os **Modelos (A e B)** apresentam dispersão uniforme em torno de zero, corroborando a homocedasticidade.

O diagnóstico gráfico serve como "primeira triagem": se houver funil, curvatura ou caudas muito pesadas, isso aparece visualmente de modo imediato. Somente depois disso faz sentido dar peso às comparações numéricas.

(vi) **Conclusões do exemplo**

- O modelo sem intercepto tende a se ajustar mal, pois ignora o consumo em stand-by. Este modelo foi descartado no teste com modelo aninhados.

- O modelo com intercepto melhora substancialmente o ajuste quando comparado com o modelo sem intercepto.

- Assim como o modelo com intercepto (A), o modelo com variável transformada (B) também mostrou um bom ajuste.

- Os resíduos dos modelos A e B mostraram que podemos considerar que ambos os modelos foram bem especificados.

- O modelo B apresentou AIC/BIC menores. No entanto, como apresentado anteriormente, não é correto comparar diretamente os valores de um modelo ajustado em $Y$ com outro ajustado em $\log(Y)$, pois a verossimilhança é diferente.

- Considerando que os modelos A e B foram bem especificados e nao teve uma métrica de qualidade de ajuste muito favorável a um deles, é aconselhado o uso do modelo mais simples e na escala original da variável (Modelo A - com intercepto).

### 10.4.2 Exemplo 2 — Escolhendo a melhor variável explicativa

(i) **Cenário e modelos candidatos** um pesquisador deseja explicar a produtividade agrícola ($Y$) a partir de três variáveis candidatas:

- $X_1$ = quantidade de fertilizante aplicada (kg/ha)

- $X_2$ = volume de irrigação (mm)

- $X_3$ = horas de sol na safra (h)

O objetivo é descobrir **qual dessas variáveis explica melhor** $Y$ de forma individual.

Os três MRLS candidatos são:

1. **Modelo A:** $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

2. **Modelo B:** $Y = \beta_0 + \beta_2 X_2 + \varepsilon$

3. **Modelo C:** $Y = \beta_0 + \beta_3 X_3 + \varepsilon$

**Passos da análise comparativa:**

1. **Ajuste de cada modelo** separadamente.

2. **Diagnóstico dos resíduos** em cada caso, verificando linearidade, homoscedasticidade e normalidade.

3. **Comparação de medidas de ajuste**: $R^2$, $R^2_{aj}$, AIC e BIC.

4. **Discussão substantiva**: qual variável faz mais sentido teoricamente como explicativa de $Y$.

**Critérios práticos de decisão:**
- Se todos os modelos tiverem resíduos adequados, a comparação pode ser feita pelos critérios formais (AIC/BIC, $R^2_{aj}$).
- Se apenas um modelo tiver resíduos consistentes com as hipóteses do MRLS, ele deve ser preferido.
- Mesmo que dois modelos apresentem ajustes estatisticamente próximos, a escolha deve considerar a **interpretação prática**.

(ii) **Simulação dos dados**

```
suppressPackageStartupMessages({
  library(dplyr)
})

set.seed(2025)

# Simulação: produtividade (Y) explicada por X1, X2, X3 (candidatas)
n2 <- 120
X1 <- runif(n2,   0, 100)   # fertilizante, 0-100 kg/ha
X2 <- runif(n2,   0,  60)   # irrigação, 0-60 mm
X3 <- runif(n2, 200, 350)   # horas de sol, 200-350 h

# Verdade de geração
Y2 <- 20 + 0.45*X1 + 0.12*X2 + 0.02*X3 + rnorm(n2, 0, 10)

df2 <- tibble(Y = Y2, X1 = X1, X2 = X2, X3 = X3)
```

```r
head(df2)
```

```
# A tibble: 6 x 4
      Y    X1    X2    X3
  <dbl> <dbl> <dbl> <dbl>
1  68.4  73.3  43.4  239.
2  59.9  47.6  39.9  245.
3  67.5  51.4  44.3  229.
4  59.2  49.8  33.2  332.
5  49.9  78.0  41.1  250.
6  42.0  50.4  43.0  241.
```

Neste exemplo, as três variáveis candidatas têm escalas diferentes, mas isso não impede o ajuste de três MRLS separados. O que muda é a interpretação dos coeficientes e a magnitude dos erros-padrão. Como o objetivo é "melhor preditor individual", estamos comparando **três modelos alternativos não aninhados**.

Este exemplo buscar mostrar que, mesmo no contexto de MRLS, a comparação entre variáveis explicativas pode guiar a seleção do **melhor preditor individual**.

(iii) **Roteiro de tarefas (atividade guiada)**

1. **Ajuste os três modelos candidatos** separadamente:
   - Modelo A: $ Y =  _0 +  _1 X_1 +  $
   - Modelo B: $ Y =  _0 +  _2 X_2 +  $
   - Modelo C: $ Y =  _0 +  _3 X_3 +  $
2. **Inspecione os resíduos** de cada modelo:
   - Gráficos de resíduos versus ajustados.
   - QQ-plot para verificar normalidade.
   - Histograma dos resíduos.
   - Resíduos versus a variável explicativa $X$.
3. **Compare as medidas de ajuste** entre os modelos:
   - $R^2_{aj}$ (ajustado)
   - AIC

- BIC

4. **Discuta os resultados obtidos**:
   - Qual modelo apresentou resíduos mais consistentes com as hipóteses do MRLS?
   - Qual modelo apresentou melhor desempenho segundo $R^2_{aj}$, AIC e BIC?
   - Existe coerência entre os diagnósticos gráficos e as medidas numéricas?
5. **Reflexão substantiva:**
   - Do ponto de vista prático, qual variável é a melhor candidata a explicar a produtividade agrícola ($Y$) individualmente e por quê?
   - Considere plausibilidade causal e relevância no contexto agrícola (fertilizante, irrigação ou horas de sol).
6. **Desafio opcional:**
   - Re-simule os dados com outro *seed* ou altere o nível de ruído.
   - Observe se a escolha do melhor modelo permanece a mesma ou se muda.

```r
# Ajustes dos três MRLS candidatos
m1 <- lm(Y ~ X1, data = df2)
m2 <- lm(Y ~ X2, data = df2)
m3 <- lm(Y ~ X3, data = df2)

cat("=== Modelo A: Y ~ X1 ===\n"); print(summary(m1))
```

```
=== Modelo A: Y ~ X1 ===



Call:
lm(formula = Y ~ X1, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-30.5982  -6.0204   0.6442   6.9029  27.5385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.49842    1.84972   17.03   <2e-16 ***
```

```
X1           0.41940    0.03076   13.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.73 on 118 degrees of freedom
Multiple R-squared:  0.6117,    Adjusted R-squared:  0.6084
F-statistic: 185.9 on 1 and 118 DF,  p-value: < 2.2e-16
```

```
cat("\n=== Modelo B: Y ~ X2 ===\n"); print(summary(m2))
```

```
=== Modelo B: Y ~ X2 ===


Call:
lm(formula = Y ~ X2, data = df2)

Residuals:
    Min      1Q  Median      3Q     Max
-45.332 -10.528   2.142  12.276  35.247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.69985    2.63414  18.108  < 2e-16 ***
X2           0.19641    0.07436   2.641  0.00937 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.17 on 118 degrees of freedom
Multiple R-squared:  0.05583,   Adjusted R-squared:  0.04783
F-statistic: 6.977 on 1 and 118 DF,  p-value: 0.009374
```

```
cat("\n=== Modelo C: Y ~ X3 ===\n"); print(summary(m3))
```

```
=== Modelo C: Y ~ X3 ===


Call:
lm(formula = Y ~ X3, data = df2)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-49.032 -12.257   1.166  11.872  31.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.31156    8.78394   6.638 1.02e-09 ***
X3          -0.01715    0.03168  -0.541    0.589
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.59 on 118 degrees of freedom
Multiple R-squared:  0.002478,  Adjusted R-squared:  -0.005975
F-statistic: 0.2932 on 1 and 118 DF,  p-value: 0.5892
```

```r
cmp2 <- tibble(
  Modelo = c("A: Y~X1", "B: Y~X2", "C: Y~X3"),
  R2     = c(summary(m1)$r.squared,     summary(m2)$r.squared,     summary(m3)$r.squared),
  R2_aj  = c(summary(m1)$adj.r.squared, summary(m2)$adj.r.squared, summary(m3)$adj.r.squared
  AIC    = c(AIC(m1), AIC(m2), AIC(m3)),
  BIC    = c(BIC(m1), BIC(m2), BIC(m3))
)

cmp2
```

```
# A tibble: 3 x 5
  Modelo       R2    R2_aj   AIC   BIC
  <chr>     <dbl>    <dbl> <dbl> <dbl>
1 A: Y~X1 0.612    0.608    891.  899.
2 B: Y~X2 0.0558   0.0478   997. 1006.
3 C: Y~X3 0.00248 -0.00598 1004. 1012.
```

**Diagnóstico gráfico comparativo:**
- dispersão $Y$ vs cada $X_j$ com reta OLS (um painel por modelo);
- resíduos vs ajustados (um painel por modelo);
- QQ-plot (um painel por modelo);
- histograma dos resíduos (um painel por modelo);
- resíduos vs $X_j$ (um painel por modelo).

```r
suppressPackageStartupMessages({
  library(ggplot2)
  library(dplyr)
  library(tidyr)
})

# Organizar dados em formato longo para facilitar facetas
long_xy <- df2 %>%
  pivot_longer(cols = c(X1, X2, X3), names_to = "Xname", values_to = "X") %>%
  mutate(modelo = recode(Xname, X1 = "Modelo A: Y~X1", X2 = "Modelo B: Y~X2", X3 = "Modelo C

# Função auxiliar para extrair resíduos e ajustados por modelo
aug1 <- tibble(fit = fitted(m1), res = resid(m1), modelo = "Modelo A: Y~X1")
aug2 <- tibble(fit = fitted(m2), res = resid(m2), modelo = "Modelo B: Y~X2")
aug3 <- tibble(fit = fitted(m3), res = resid(m3), modelo = "Modelo C: Y~X3")
aug  <- bind_rows(aug1, aug2, aug3)

# 1) Y vs X com reta OLS
p1 <- ggplot(long_xy, aes(x = X, y = Y)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  facet_wrap(~modelo, ncol = 3, scales = "free_x") +
  labs(x = NULL, y = "Y") +
  theme_minimal(base_size = 12)

# 2) Resíduos vs ajustados
p2 <- ggplot(aug, aes(x = fit, y = res)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~modelo, ncol = 3, scales = "free_x") +
  labs(x = "Ajustados", y = "Resíduo") +
  theme_minimal(base_size = 12)

# 3) QQ-plot (estilo plot.lm, which = 2)

qq_df_lmstyle <- function(mod, modelo){
  r <- rstandard(mod)                 # mesmo tipo de resíduo do plot.lm
  q <- qqnorm(r, plot.it = FALSE)

  pts <- tibble(
    theo = q$x,
    samp = q$y,
```

```
    modelo = modelo
  )

  # reta tipo qqline(): passa pelos quartis 25% e 75%
  yq <- quantile(r, c(0.25, 0.75))
  xq <- qnorm(c(0.25, 0.75))
  slope <- (yq[2] - yq[1])/(xq[2] - xq[1])
  intercept <- yq[1] - slope*xq[1]

  line <- tibble(
    modelo = modelo,
    intercept = as.numeric(intercept),
    slope = as.numeric(slope)
  )

  list(pts = pts, line = line)
}

o1 <- qq_df_lmstyle(m1, "Modelo A: Y~X1")
o2 <- qq_df_lmstyle(m2, "Modelo B: Y~X2")
o3 <- qq_df_lmstyle(m3, "Modelo C: Y~X3")

qq_all  <- bind_rows(o1$pts,  o2$pts,  o3$pts)
qq_line <- bind_rows(o1$line, o2$line, o3$line)

p3 <- ggplot(qq_all, aes(x = theo, y = samp)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_abline(
    data = qq_line,
    aes(intercept = intercept, slope = slope),
    linewidth = 0.8
  ) +
  facet_wrap(~modelo, ncol = 3, scales = "free") +
  labs(x = "Quantis teóricos (Normal)", y = "Resíduos padronizados") +
  theme_minimal(base_size = 12)

# 4) Histogramas dos resíduos
res_all <- bind_rows(
  tibble(r = resid(m1), modelo = "Modelo A: Y~X1"),
  tibble(r = resid(m2), modelo = "Modelo B: Y~X2"),
  tibble(r = resid(m3), modelo = "Modelo C: Y~X3")
)
```

```r
p4 <- ggplot(res_all, aes(x = r)) +
  geom_histogram(bins = 12, alpha = 0.7, color = "white") +
  facet_wrap(~modelo, ncol = 3, scales = "free") +
  labs(x = "Resíduo", y = "Frequência") +
  theme_minimal(base_size = 12)

# 5) Resíduos vs X (por modelo)
rx <- bind_rows(
  df2 %>% transmute(X = X1, res = resid(m1), modelo = "Modelo A: Y~X1"),
  df2 %>% transmute(X = X2, res = resid(m2), modelo = "Modelo B: Y~X2"),
  df2 %>% transmute(X = X3, res = resid(m3), modelo = "Modelo C: Y~X3")
)

p5 <- ggplot(rx, aes(x = X, y = res)) +
  geom_point(alpha = 0.85, size = 2) +
  geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.6) +
  facet_wrap(~modelo, ncol = 3, scales = "free_x") +
  labs(x = NULL, y = "Resíduo") +
  theme_minimal(base_size = 12)

# Mostrar os 5 gráficos (um por vez, na ordem)
print(p1)
```

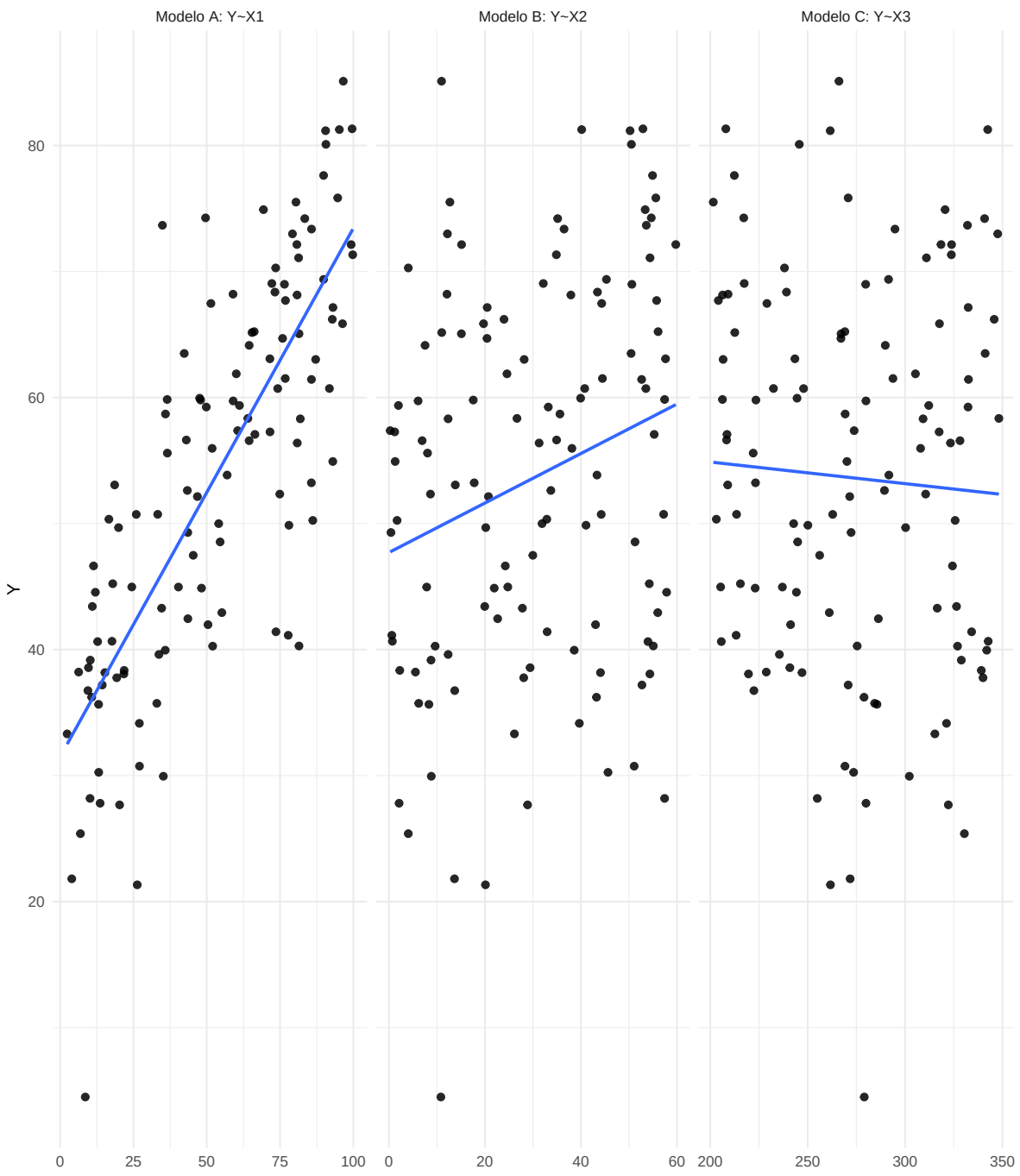



Figura 10.7: Diagnóstico gráfico comparativo (Exemplo 2): cada linha corresponde a um modelo (A, B, C).

```
print(p2)
```

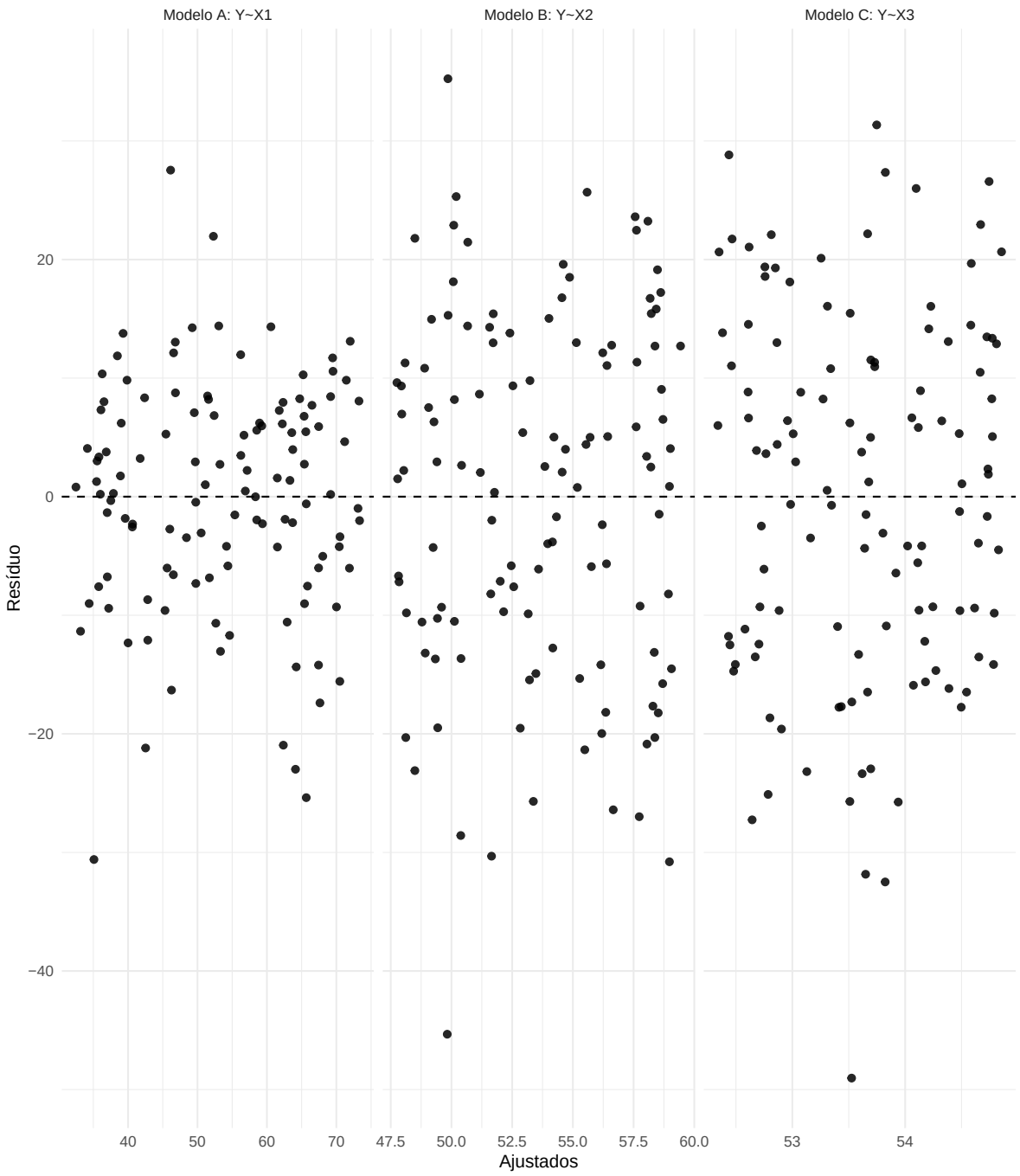



Figura 10.8: Diagnóstico gráfico comparativo (Exemplo 2): cada linha corresponde a um modelo (A, B, C).

```
print(p3)
```

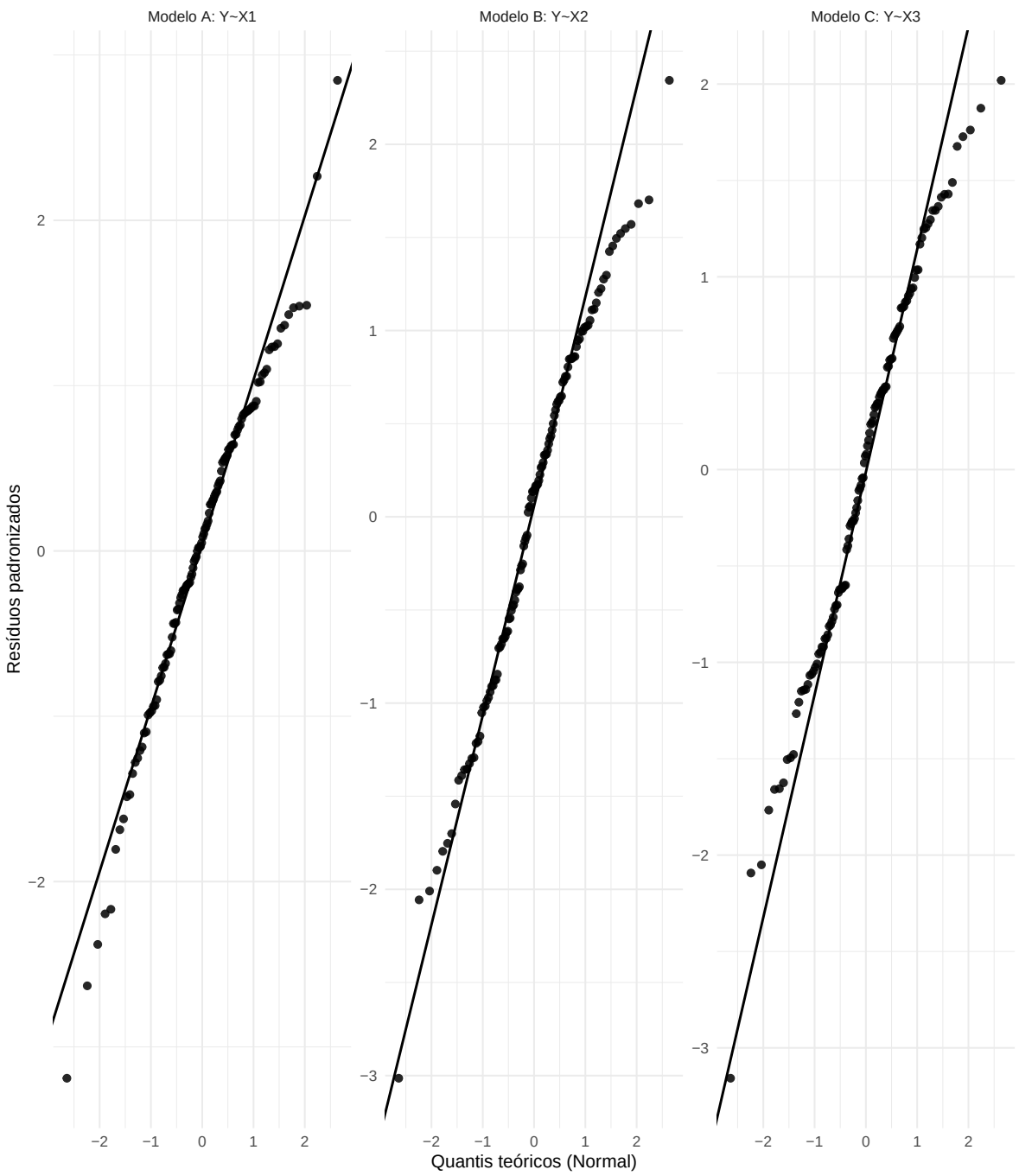



Figura 10.9: Diagnóstico gráfico comparativo (Exemplo 2): cada linha corresponde a um modelo (A, B, C).

```
print(p4)
```

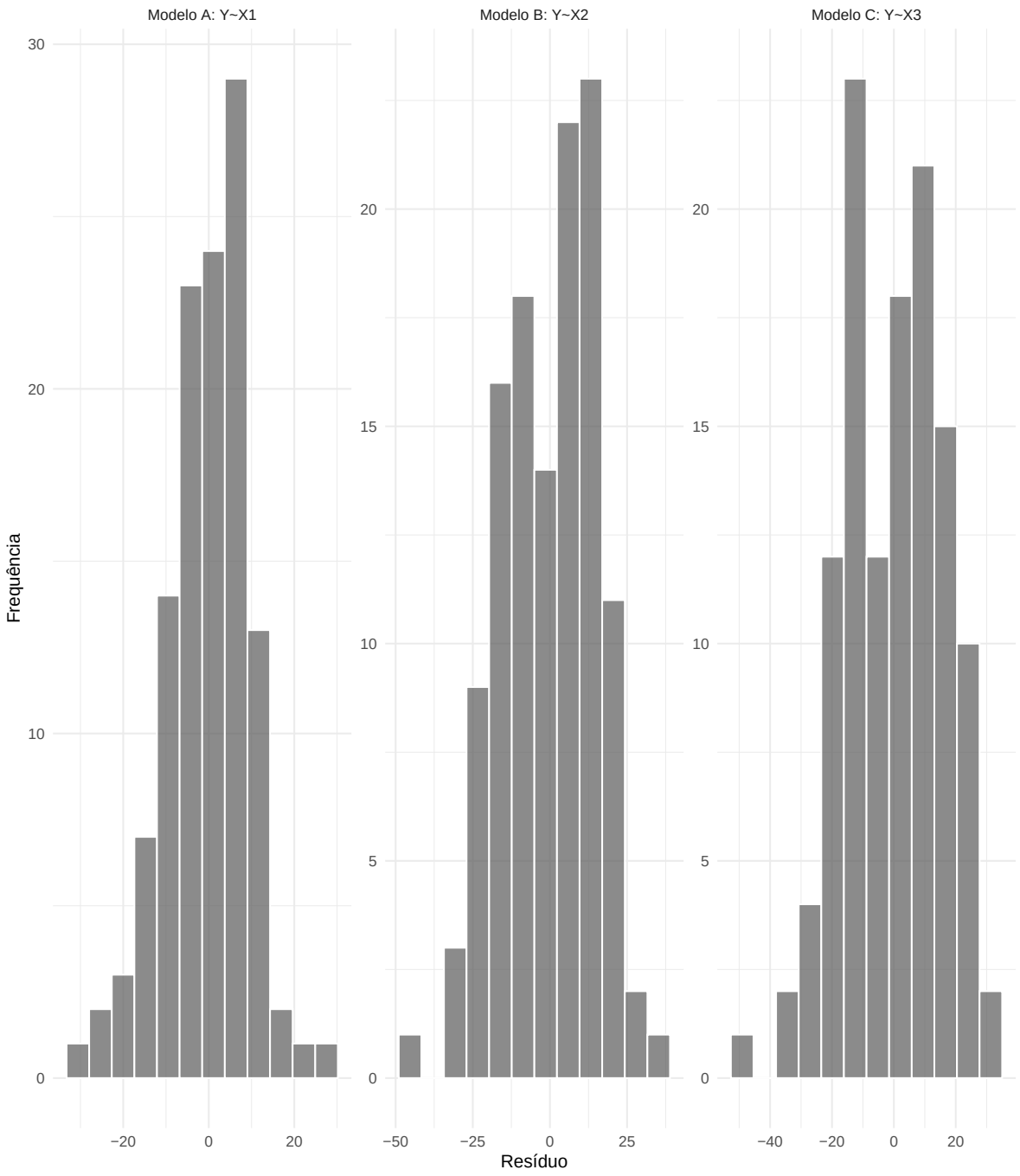



Figura 10.10: Diagnóstico gráfico comparativo (Exemplo 2): cada linha corresponde a um modelo (A, B, C).

```
print(p5)
```

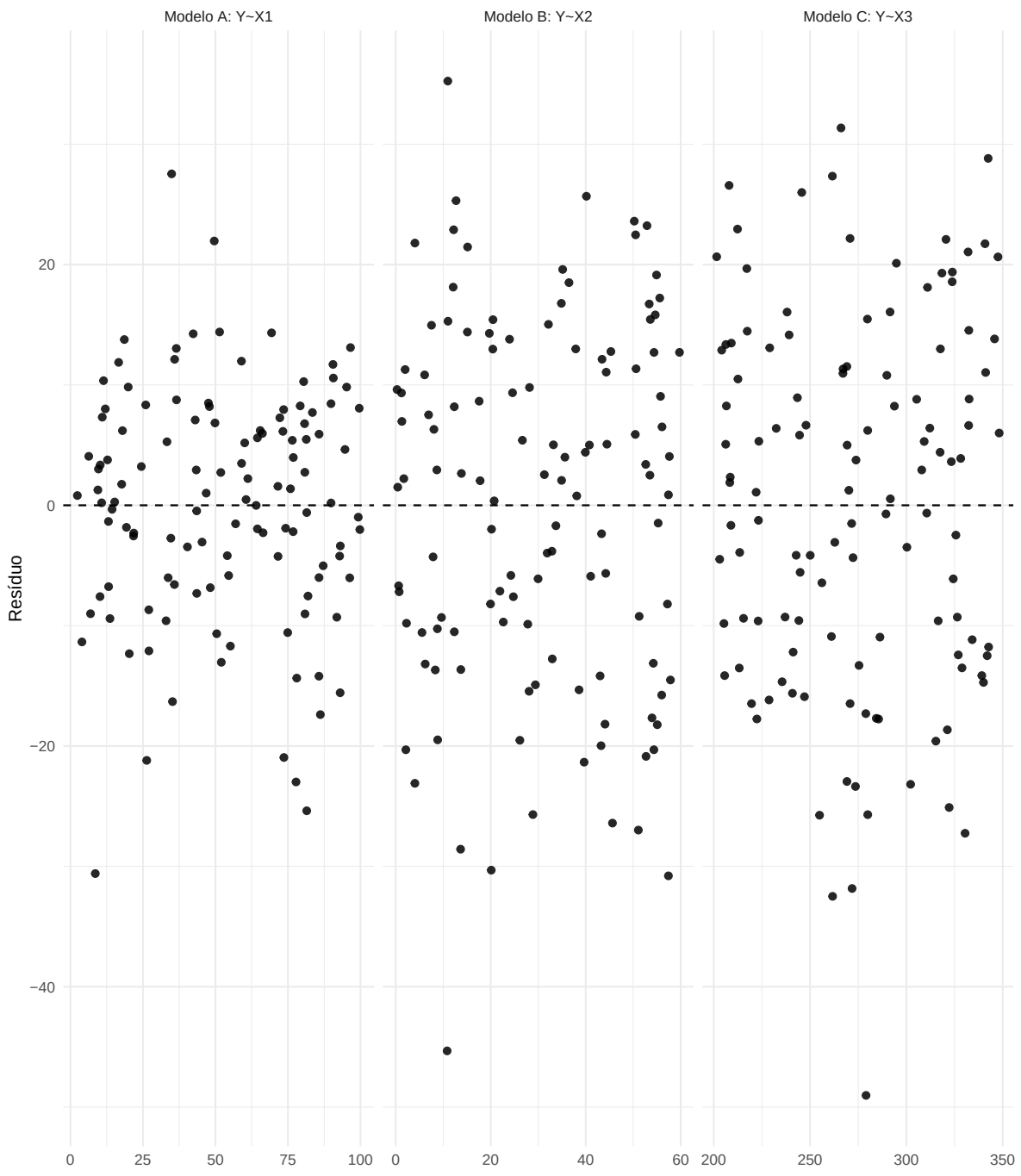



Figura 10.11: Diagnóstico gráfico comparativo (Exemplo 2): cada linha corresponde a um modelo (A, B, C).

- Se um modelo “vence” nos critérios numéricos, mas apresenta funil/curvatura/caudas

pesadas, ele deve perder força como candidato.

- Se dois modelos forem próximos numericamente, a decisão pode depender do contexto (medição, causalidade plausível, custo de obter a variável, etc.).

(iv) **Fechamento e síntese**

Os dois exemplos mostraram situações complementares na prática do MRLS. O **Exemplo 1** destacou como diferentes especificações de um mesmo modelo, com intercepto, sem intercepto e com transformação em (Y), podem levar a conclusões distintas sobre ajuste, resíduos e interpretação. Já o **Exemplo 2** explorou a escolha entre diferentes variáveis candidatas, mostrando como a comparação de modelos separados orienta a seleção do preditor mais adequado.

Em conjunto, os exemplos reforçam que a escolha do modelo não deve se basear apenas em indicadores numéricos, mas sim no equilíbrio entre **consistência estatística, parcimônia e coerência substantiva** com o fenômeno estudado.

# 11 Exercícios e atividades

Em breve!!!

# Parte III

# Parte III — Modelo de Regressão Linear Simples (MRLM)

# 12 Exercícios e atividades

Em breve!!!

# Parte IV

# Parte IV — Apêndices

# 13 Lista de Siglas e Símbolos

## 13.1 Lista de Siglas

| Sigla | Significado |
|---|---|
| MRLS | Modelo de Regressão Linear Simples |
| MRLM | Modelo de Regressão Linear Múltipla |
| MQO / OLS | Mínimos Quadrados Ordinários / *Ordinary Least Squares* |
| MV | Máxima Verossimilhança |
| WLS | *Weighted Least Squares* (Regressão Linear Ponderada) |
| GLS | *Generalized Least Squares* (MQ Generalizados) |
| BLUE | *Best Linear Unbiased Estimator* (Melhor Estimador Linear Não-Viesado) |
| GLM | *Generalized Linear Model* (Modelo Linear Generalizado) |
| GAM | *Generalized Additive Model* (Modelo Aditivo Generalizado) |
| GAMLSS | *Generalized Additive Model for Location, Scale and Shape* |
| ANOVA | Análise de Variância |
| IC | Intervalo de Confiança |
| EP | Erro-Padrão |
| IC95% | Intervalo de Confiança a 95% |
| H0, H1 | Hipótese nula, Hipótese alternativa |
| PDF, CDF | Função Densidade de Probabilidade; Função de Distribuição Acumulada |
| AIC, BIC, AICc | Critérios de Informação (Akaike, Bayesiano, Akaike corrigido) |
| $R^2$, $\bar{R}^2$ | Coeficiente de determinação e ajustado |
| SQTot, SQReg, SQRes | Somas de Quadrados (Total, Regressão, Resíduos) |
| PRESS | *Predicted Residual Sum of Squares* (LOOCV) |
| CV-$k$ | $k$-fold Cross-Validation |
| VIF | *Variance Inflation Factor* |
| df | Graus de liberdade |
| Var, Cov, Corr | Variância, Covariância, Correlação |
| FDR | *False Discovery Rate* |

## 13.2 Lista de Símbolos

| Símbolo | Descrição |
|---|---|
| $Y,\ \mathbf{Y}$ | Variável resposta (escalar / vetor de observações) |
| $X,\ \mathbf{X}$ | Matriz de covariáveis (matriz de planejamento) |
| $\beta_0,\ \beta_1,\ \ldots,\ \beta_p$ | Parâmetros do modelo (intercepto e inclinações) |
| $\beta$ | Vetor de parâmetros $(\beta_0,\ \beta_1,\ \ldots,\ \beta_p)^\top$ |
| $\varepsilon_i,\ \varepsilon$ | Termo(s) de erro aleatório (escalar / vetor) |
| $\widehat{\beta}_j,\ \widehat{\beta}$ | Estimadores (MQO/MV) dos parâmetros |
| $\widehat{Y}_i,\ \widehat{\mathbf{Y}}$ | Valores ajustados pelo modelo (escalar / vetor) |
| $\widehat{\varepsilon}_i,\ \widehat{\varepsilon}$ | Resíduos (observado – ajustado; escalar / vetor) |
| $\widehat{\sigma}^2$ | Estimador da variância residual ($\widehat{\sigma}^2 = \text{SQRes}/(n-p-1)$) |
| $\mathbf{H}$ | Matriz "chapéu" (projeção): $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ |
| $\mathbf{M}$ | Matriz dos resíduos: $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$ |
| $h_{ii}$ | Alavancagem (diagonal de $\mathbf{H}$) |
| $r_i$ | Resíduo *studentizado*: $r_i = \dfrac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}$ |
| $D_i$ | Distância de Cook |
| $\mathbf{I}_n$ | Matriz identidade de dimensão $n$ |
| $\text{col}(\mathbf{X})$ | Espaço coluna de $\mathbf{X}$ |
| $\text{rank}(\mathbf{X})$ | Posto (rank) de $\mathbf{X}$ |
| $\mathbf{X}^+$ | Inversa generalizada de Moore–Penrose |
| $\text{tr}(\mathbf{A})$ | Traço da matriz $\mathbf{A}$ |
| $\det(\mathbf{A})$ | Determinante de $\mathbf{A}$ |
| $\mathbf{A}^{-1},\ \mathbf{A}^\top$ | Inversa e transposta de $\mathbf{A}$ |
| $\mathbb{R}^n$ | Espaço vetorial real $n$-dimensional |
| $\mathcal{N}_n(\mu, \Sigma)$ | Distribuição Normal $n$-variada |
| $\mu,\ \Sigma$ | Média e matriz de covariância na Normal multivariada |
| $n,\ p$ | Nº de observações; Nº de variáveis explicativas |
| $\bar{X},\ \bar{Y}$ | Médias amostrais de $X$ e $Y$ |
| $S_{xx},\ S_{xy}$ | Somas de quadrados e produto cruzado (MRLS) |
| $\mathbf{C},\ d$ | Matriz de contrastes e vetor-alvo (testes conjuntos: $H_0 : \mathbf{C}\beta = d$) |

# 14 Estrutura Matricial dos Modelos de Regressão Linear

A formulação moderna dos modelos de regressão linear é essencialmente matricial. Essa notação vetorial revela a estrutura geométrica do problema de estimação, explicita as condições necessárias para identificabilidade dos parâmetros e permite analisar propriedades estatísticas dos estimadores de forma sistemática (ver Harville (1997)).

Este apêndice consolida os principais elementos de Álgebra Linear utilizados ao longo do estudo de regressão, com ênfase nas estruturas que reaparecem na estimação por mínimos quadrados, na inferência e na análise de diagnóstico.

## 14.1 Operações Fundamentais com Matrizes e Vetores

A linguagem matricial é uma forma compacta de escrever o modelo de regressão que permite enxergar o problema como um problema geométrico em $\mathbb{R}^n$. Cada vetor corresponde a um ponto ou direção nesse espaço, e cada matriz representa uma transformação linear.

Sejam $\mathbf{A}$ e $\mathbf{B}$ matrizes de dimensões compatíveis e $\mathbf{x}, \mathbf{y}$ vetores coluna em $\mathbb{R}^n$.

Para fixar ideias, considere explicitamente:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

### 14.1.1 Soma Matricial

Se

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

então

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}.$$

A soma é definida elemento a elemento e exige dimensões idênticas.

### 14.1.2 Produto Matricial

Se

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

então

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}.$$

O produto matricial corresponde à composição de transformações lineares.
No modelo linear

$$\mathbf{Y} = \mathbf{X}\beta,$$

a matriz $\mathbf{X}$, de dimensão $n \times (p+1)$ transforma o vetor de parâmetros $\beta$, de dimensão $(p+1) \times 1$ em um vetor no espaço das respostas. Assim, $\mathbf{X}$ pode ser interpretada como uma transformação que leva parâmetros em $\mathbb{R}^{p+1}$ para vetores ajustados em $\mathbb{R}^n$.

### 14.1.3 Produto Interno e Norma

O produto interno entre vetores é dado por

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

Quando $\mathbf{x} = \mathbf{y}$, obtemos

$$\mathbf{x}^\top \mathbf{x} = \sum_{i=1}^{n} x_i^2,$$

que define o quadrado da norma euclidiana:

$$\|\mathbf{x}\|^2 = \mathbf{x}^\top \mathbf{x}.$$

Essa noção de norma é central na regressão, pois a estimação por mínimos quadrados consiste em minimizar

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2.$$

Portanto, o problema de estimação é um problema geométrico de minimizar distância no espaço $\mathbb{R}^n$. A formulação geométrica da regressão em termos de subespaços e projeções é desenvolvida em detalhe em Harville (1997).

### 14.1.4 Forma Quadrática

Uma expressão da forma

$$\mathbf{x}^\top \mathbf{A} \mathbf{x}$$

é chamada forma quadrática.

Para visualizar, considere

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Então

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = a x_1^2 + 2 b x_1 x_2 + c x_2^2.$$

Observe que surgem termos quadráticos e termos mistos. Em regressão, as somas de quadrados explicada e residual podem ser escritas exatamente como formas quadráticas do vetor $\mathbf{Y}$ (ver Rencher e Christensen (2012)).

### 14.1.5 Transposição

A transposta de uma matriz é obtida trocando linhas por colunas:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}.$$

A transposição é essencial para definir produtos internos e garantir que expressões como $\mathbf{X}^\top\mathbf{X}$ sejam matrizes quadradas.

### 14.1.6 Inversão

Uma matriz quadrada $\mathbf{A}$ é invertível se existe $\mathbf{A}^{-1}$ tal que

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.$$

Por exemplo, para uma matriz $2 \times 2$,

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

se $ad - bc \neq 0$, então

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Na regressão, a invertibilidade de $\mathbf{X}^\top\mathbf{X}$ é condição necessária para a existência do estimador de mínimos quadrados único.Para o caso de posto deficiente e o uso de decomposição SVD e pseudoinversas, ver Golub e Van Loan (2013).

### 14.1.7 Propriedades Importantes

Duas identidades frequentemente utilizadas são

$$(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top, \qquad (\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}.$$

Essas propriedades são fundamentais na dedução de resultados como:

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1},$$

e na demonstração das propriedades das matrizes de projeção.

Essas operações constituem o mecanismo estrutural que permitirá:

- interpretar o estimador como projeção ortogonal;
- escrever somas de quadrados como formas quadráticas;
- analisar variâncias e covariâncias de estimadores;
- compreender a geometria do ajuste e do diagnóstico.

Nos itens seguintes, essas operações serão organizadas dentro da estrutura específica do modelo linear múltiplo.

## 14.2 Estruturas Matriciais Relevantes

Determinados tipos de matrizes surgem de forma recorrente na teoria da regressão linear. Cada uma delas corresponde a uma propriedade geométrica ou estatística que será explorada na estimação, na inferência e na análise de diagnóstico.

A tabela a seguir resume algumas dessas estruturas fundamentais.

| Tipo | Definição | Relevância |
|---|---|---|
| Identidade $\mathbf{I}_n$ | Diagonal principal composta por 1's | Elemento neutro da multiplicação |
| Simétrica | $\mathbf{A} = \mathbf{A}^\top$ | Autovalores reais |
| Idempotente | $\mathbf{A}^2 = \mathbf{A}$ | Projeções |
| Ortogonal | $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ | Preserva norma |
| Diagonal | Elementos fora da diagonal iguais a zero | Simplifica formas quadráticas |
| Definida positiva | $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$ para todo $\mathbf{x} \neq 0$ | Garantia de invertibilidade e convexidade |

A seguir, detalham-se as propriedades e implicações dessas estruturas no contexto do modelo linear.

### 14.2.1 Matriz Identidade

A matriz identidade de ordem $n$ é dada por

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Ela satisfaz

$$\mathbf{I}_n \mathbf{x} = \mathbf{x} \quad \text{para todo } \mathbf{x} \in \mathbb{R}^n.$$

No modelo linear, a identidade aparece, por exemplo, na matriz de covariância dos erros sob homocedasticidade:

$$\text{Cov}(\varepsilon) = \sigma^2 \mathbf{I}_n.$$

Sob normalidade, essa estrutura implica independência e variância constante dos erros.

### 14.2.2 Matrizes Simétricas

Uma matriz é simétrica se

$$\mathbf{A} = \mathbf{A}^\top.$$

Por exemplo,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$$

é simétrica.

Matrizes simétricas possuem autovalores reais e podem ser diagonalizadas por matrizes ortogonais. Essa propriedade é crucial para compreender a decomposição espectral de $\mathbf{X}^\top \mathbf{X}$ e analisar problemas como multicolinearidade (ver Harville (1997)).

No modelo linear, as matrizes $\mathbf{X}^\top \mathbf{X}$, $\mathbf{H}$ e $\mathbf{M}$ são simétricas.

### 14.2.3 Matrizes Idempotentes

Uma matriz é idempotente se

$$\mathbf{A}^2 = \mathbf{A}.$$

Isso implica que aplicar a transformação duas vezes produz o mesmo resultado que aplicá-la uma vez.

Por exemplo, a matriz

$$\mathbf{P} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

é idempotente.

No modelo linear, a matriz de projeção

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

satisfaz

$$\mathbf{H}^2 = \mathbf{H}.$$

Isso significa que $\mathbf{H}$ projeta vetores no subespaço $\text{col}(\mathbf{X})$. Uma vez projetado, aplicar novamente a projeção não altera o vetor.

Essa propriedade é fundamental para compreender:

- decomposição ortogonal;
- independência entre componentes projetadas sob normalidade;
- decomposição da soma de quadrados total.

### 14.2.4 Matrizes Ortogonais

Uma matriz $\mathbf{Q}$ é ortogonal se

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}.$$

Isso implica que

$$\|\mathbf{Q}\mathbf{x}\| = \|\mathbf{x}\|.$$

Ou seja, matrizes ortogonais preservam comprimentos e ângulos.

Essa propriedade é central na decomposição espectral de matrizes simétricas:

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top,$$

em que $\Lambda$ é diagonal contendo os autovalores.A análise numérica dessas decomposições é tratada em profundidade em Golub e Van Loan (2013).

Na regressão, essa decomposição permite analisar:

- a estrutura de $\mathbf{X}^\top\mathbf{X}$;
- a estabilidade numérica da estimação;
- o efeito da multicolinearidade.

### 14.2.5 Matrizes Diagonais

Uma matriz diagonal possui a forma

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}.$$

Formas quadráticas envolvendo matrizes diagonais simplificam-se para

$$\mathbf{x}^\top \mathbf{D}\mathbf{x} = \sum_{i=1}^{n} d_i x_i^2.$$

Essa simplificação é útil na análise de variâncias e na interpretação de decomposições espectrais.

### 14.2.6 Matrizes Definidas Positivas

Uma matriz simétrica $\mathbf{A}$ é definida positiva se

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \quad \text{para todo } \mathbf{x} \neq \mathbf{0}.$$

Equivalentemente, todos os seus autovalores são positivos.

Essa propriedade possui consequências fundamentais:

1. $\mathbf{A}$ é invertível;
2. a função $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ é estritamente convexa;
3. problemas de minimização associados possuem solução única.

No modelo de regressão linear, a matriz

$$\mathbf{X}^\top \mathbf{X}$$

é simétrica e definida positiva se, e somente se, as colunas de $\mathbf{X}$ forem linearmente independentes. Isso equivale à condição

$$\text{rank}(\mathbf{X}) = p + 1.$$

Quando essa condição é satisfeita, o estimador de mínimos quadrados

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

existe e é único.

Se $\mathbf{X}^\top \mathbf{X}$ não for definida positiva, o problema de estimação perde identificabilidade, caracterizando multicolinearidade perfeita.

A compreensão dessas estruturas matriciais permite interpretar o modelo linear como:

- um problema geométrico de projeção em subespaços;
- um problema analítico de minimização convexa;
- um problema espectral envolvendo autovalores e autovetores.

Essas perspectivas convergem na teoria de estimação, inferência e diagnóstico.

## 14.3 Estrutura Geométrica do Modelo Linear

Considere o modelo linear múltiplo em notação matricial:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

em que

- $\mathbf{Y} \in \mathbb{R}^n$ é o vetor de respostas observadas,
- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ é a matriz de planejamento,
- $\beta \in \mathbb{R}^{p+1}$ é o vetor de parâmetros,
- $\varepsilon \in \mathbb{R}^n$ é o vetor de erros.

Explicitamente, pode-se escrever

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Cada coluna de $\mathbf{X}$ é um vetor em $\mathbb{R}^n$. Assim, $\mathbf{X}$ pode ser vista como um conjunto de $p+1$ vetores que geram um subespaço de $\mathbb{R}^n$.

### 14.3.1 Espaço Coluna e Identificabilidade

O **espaço coluna** de $\mathbf{X}$ é definido como

$$\operatorname{col}(\mathbf{X}) = \left\{\mathbf{X}\beta : \beta \in \mathbb{R}^{p+1}\right\}.$$

Esse conjunto é um subespaço vetorial de $\mathbb{R}^n$, cujo posto é

$$\operatorname{rank}(\mathbf{X}) \leq p + 1.$$

Se as colunas de $\mathbf{X}$ forem linearmente independentes, então

$$\operatorname{rank}(\mathbf{X}) = p + 1,$$

e o espaço coluna tem dimensão $p+1$.

Essa condição é equivalente à positividade definida de $\mathbf{X}^\top\mathbf{X}$ e garante a identificabilidade única dos parâmetros.

### 14.3.2 O Problema de Mínimos Quadrados como Problema de Projeção

O estimador de mínimos quadrados é definido como a solução do problema

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2.$$

Geometricamente, isso significa encontrar o vetor em $\text{col}(\mathbf{X})$ que esteja mais próximo de $\mathbf{Y}$ na métrica euclidiana.

Seja

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

Então

$$\hat{\mathbf{Y}} \in \text{col}(\mathbf{X})$$

e

$$\mathbf{Y} - \hat{\mathbf{Y}} \perp \text{col}(\mathbf{X}).$$

Ou seja,

$$\mathbf{X}^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

Essa condição é exatamente a forma matricial das equações normais.

### 14.3.3 Interpretação Ortogonal e Soma Direta

Seja $V = \mathbb{R}^n$ equipado com o produto interno usual

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}.$$

Se $S$ é um subespaço de $\mathbb{R}^n$, define-se seu complemento ortogonal como

$$S^\perp = \left\{ \mathbf{z} \in \mathbb{R}^n : \mathbf{z}^\top \mathbf{s} = 0 \text{ para todo } \mathbf{s} \in S \right\}.$$

Diz-se que $\mathbb{R}^n$ é a **soma direta ortogonal** de dois subespaços $S$ e $T$ se:

1. todo vetor de $\mathbb{R}^n$ pode ser escrito como soma de um vetor em $S$ e um vetor em $T$;

2. essa decomposição é única;
3. $S$ e $T$ são ortogonais, isto é, $\mathbf{s}^\top \mathbf{t} = 0$ para todo $\mathbf{s} \in S$ e $\mathbf{t} \in T$.

Nessa situação escreve-se

$$\mathbb{R}^n = S \oplus T,$$

em que o símbolo $\oplus$ indica soma direta.

No contexto do modelo linear, tomando

$$S = \mathrm{col}(\mathbf{X}), \qquad T = \mathrm{col}(\mathbf{X})^\perp,$$

obtém-se

$$\mathbb{R}^n = \mathrm{col}(\mathbf{X}) \oplus \mathrm{col}(\mathbf{X})^\perp.$$

Isso significa que qualquer vetor $\mathbf{Y} \in \mathbb{R}^n$ pode ser decomposto de maneira única como

$$\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\varepsilon},$$

onde

- $\hat{\mathbf{Y}} \in \mathrm{col}(\mathbf{X})$,
- $\hat{\varepsilon} \in \mathrm{col}(\mathbf{X})^\perp$.

A ortogonalidade implica

$$\hat{\mathbf{Y}}^\top \hat{\varepsilon} = 0,$$

ou, equivalentemente,

$$\mathbf{X}^\top \hat{\varepsilon} = \mathbf{0}.$$

Essa decomposição é puramente geométrica e independe de qualquer hipótese probabilística sobre os erros. Ela constitui o núcleo estrutural do método de mínimos quadrados e fundamenta:

- a decomposição da soma de quadrados total;
- a contagem de graus de liberdade;
- a independência entre componentes projetadas sob normalidade.

Assim, o modelo linear pode ser interpretado como a decomposição ortogonal do vetor de respostas em duas componentes pertencentes a subespaços complementares.

### 14.3.4 Relação com Posto e Dimensão

Se $\text{rank}(\mathbf{X}) = r$, então:

- $\dim(\text{col}(\mathbf{X})) = r$,
- $\dim(\text{col}(\mathbf{X})^{\perp}) = n - r$.

No modelo linear completo com intercepto e colunas independentes,

$$r = p + 1,$$

e, portanto, o espaço residual tem dimensão

$$n - (p + 1).$$

Essa contagem de dimensões será reinterpretada mais adiante como graus de liberdade na decomposição das somas de quadrados.

### 14.3.5 Conexão com Diagnóstico

A estrutura geométrica permite compreender diversos elementos de diagnóstico:

- Vetores de alta alavancagem correspondem a observações cuja projeção sobre $\text{col}(\mathbf{X})$ é dominante.
- Resíduos grandes correspondem a componentes significativas no subespaço ortogonal.
- A decomposição da variabilidade total decorre da ortogonalidade entre componentes projetadas.

O modelo linear é, portato, uma decomposição geométrica do vetor de respostas em dois componentes ortogonais.

## 14.4 Matrizes de Projeção e Decomposição Ortogonal

A matriz de projeção associada ao modelo linear múltiplo é definida por

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top.$$

Essa matriz desempenha papel relevante na teoria da regressão, pois formaliza algebricamente a projeção ortogonal sobre o subespaço $\text{col}(\mathbf{X})$.

#### 14.4.0.1 Verificação das Propriedades Estruturais

**Simetria**

$$\mathbf{H}^\top = \left[\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right]^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}.$$

Utilizou-se o fato de que $\mathbf{X}^\top\mathbf{X}$ é simétrica e que a transposta de um produto inverte a ordem dos fatores.

**Idempotência**

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top.$$

Como

$$\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{I},$$

segue que

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H}.$$

A idempotência caracteriza transformações que, uma vez aplicadas, não alteram mais o vetor.

#### 14.4.0.2 Interpretação Geométrica

Para qualquer vetor $\mathbf{y} \in \mathbb{R}^n$,

$$\mathbf{H}\mathbf{y} \in \text{col}(\mathbf{X}).$$

Além disso,

$$\mathbf{y} - \mathbf{H}\mathbf{y} \in \text{col}(\mathbf{X})^{\perp}.$$

Portanto,

$$\mathbf{y} = \mathbf{H}\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

Definindo

$$\mathbf{M} = \mathbf{I}_n - \mathbf{H},$$

obtém-se a projeção complementar sobre o subespaço ortogonal.

#### 14.4.0.3 Propriedades da Matriz Residual

A matriz

$$\mathbf{M} = \mathbf{I}_n - \mathbf{H}$$

satisfaz:

- $\mathbf{M}^{\top} = \mathbf{M}$,
- $\mathbf{M}^2 = \mathbf{M}$,
- $\mathbf{H}\mathbf{M} = \mathbf{0}$,
- $\mathbf{M}\mathbf{H} = \mathbf{0}$.

Essas propriedades garantem que os subespaços são ortogonais e complementares.

#### 14.4.0.4 Decomposição do Vetor de Respostas

Aplicando as matrizes ao vetor $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{HY} + \mathbf{MY} = \hat{\mathbf{Y}} + \hat{\varepsilon}.$$

em que

- $\hat{\mathbf{Y}} = \mathbf{HY}$ pertence ao espaço coluna;
- $\hat{\varepsilon} = \mathbf{MY}$ pertence ao complemento ortogonal.

A ortogonalidade implica

$$\hat{\mathbf{Y}}^{\top}\hat{\varepsilon} = 0.$$

Essa identidade é a base da decomposição da soma de quadrados total.

#### 14.4.0.5 Traço, Posto e Autovalores

Para matrizes idempotentes e simétricas, os autovalores são apenas 0 ou 1.

Se

$$\text{rank}(\mathbf{X}) = p + 1,$$

então:

- $\mathbf{H}$ possui $p + 1$ autovalores iguais a 1,
- e $n - (p + 1)$ autovalores iguais a 0.

Assim,

$$\text{tr}(\mathbf{H}) = p + 1.$$

De forma análoga,

$$\text{tr}(\mathbf{M}) = n - p - 1.$$

Como o traço de uma matriz idempotente simétrica coincide com seu posto, essas quantidades correspondem às dimensões dos subespaços projetados (ver Harville (1997)).

#### 14.4.0.6 Conexão com Somas de Quadrados

A soma de quadrados ajustada pode ser escrita como

$$\mathbf{Y}^\top \mathbf{H} \mathbf{Y}.$$

A soma de quadrados residual é

$$\mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

Como $\mathbf{H}$ e $\mathbf{M}$ projetam sobre subespaços ortogonais,

$$\mathbf{Y}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{H} \mathbf{Y} + \mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

Essa identidade é puramente geométrica e antecede qualquer consideração probabilística.

#### 14.4.0.7 Relação com Diagnóstico

A diagonal de $\mathbf{H}$ contém as alavancagens:

$$h_{ii} = (\mathbf{H})_{ii}.$$

Essas quantidades medem o grau de influência estrutural da $i$-ésima observação no ajuste, pois determinam o peso da projeção sobre o espaço coluna (ver Weisberg (2005)).

Valores elevados de $h_{ii}$ indicam observações que ocupam posições extremas no espaço das covariáveis.

A matriz de projeção sintetiza, portanto, três dimensões fundamentais do modelo linear:

1. Estrutura geométrica (projeção ortogonal);
2. Estrutura algébrica (idempotência e posto);
3. Estrutura estatística (somas de quadrados e graus de liberdade).

## 14.5 Diferenciação Matricial e Estimação por Mínimos Quadrados

A estimação por mínimos quadrados consiste na minimização da soma de quadrados dos resíduos,

$$S(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2 = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta).$$

Essa função é uma forma quadrática em $\beta$. Para compreender sua estrutura, é útil expandi-la explicitamente:

$$S(\beta) = \mathbf{Y}^\top \mathbf{Y} - 2\beta^\top \mathbf{X}^\top \mathbf{Y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta.$$

Observa-se que:

- $\mathbf{Y}^\top \mathbf{Y}$ é constante em relação a $\beta$;
- $\beta^\top \mathbf{X}^\top \mathbf{Y}$ é termo linear;
- $\beta^\top \mathbf{X}^\top \mathbf{X}\beta$ é forma quadrática.

A função objetivo é, portanto, um polinômio quadrático convexo sempre que $\mathbf{X}^\top \mathbf{X}$ for definida positiva.

### 14.5.1 Derivadas Matriciais Fundamentais

As identidades de cálculo matricial utilizadas a seguir são sistematizadas em Abadir e Magnus (2005).

1. $$\frac{\partial(\mathbf{a}^\top \mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$$

2. $$\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$$

Em particular, se $\mathbf{A}$ é simétrica,

$$\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

3. $$\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{b})}{\partial \mathbf{x}} = \mathbf{A}^\top \mathbf{b}$$

4.
$$\frac{\partial \text{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}^\top$$

5.
$$\frac{\partial (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta)}{\partial \beta} = -2\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta).$$

### 14.5.2 Derivação das Equações Normais

Aplicando a derivada à função $S(\beta)$ (ver Abadir e Magnus (2005)):

$$\nabla_\beta S(\beta) = -2\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta).$$

Igualando o gradiente a zero:

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}) = 0.$$

Reorganizando,

$$\mathbf{X}^\top \mathbf{X}\widehat{\beta} = \mathbf{X}^\top \mathbf{Y}.$$

Essas são as **equações normais**.

Se $\mathbf{X}^\top \mathbf{X}$ é invertível, isto é, se as colunas de $\mathbf{X}$ são linearmente independentes, obtém-se a solução única:

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

### 14.5.3 Interpretação Algébrica

A condição

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}) = 0$$

implica

$$\mathbf{X}^\top \widehat{\varepsilon} = 0,$$

ou seja, o vetor residual é ortogonal a cada coluna de $\mathbf{X}$.

### 14.5.4 Interpretação Geométrica

A minimização de $S(\beta)$ equivale a resolver

$$\min_{\mathbf{v}\in\text{col}(\mathbf{X})} \|\mathbf{Y}-\mathbf{v}\|^2.$$

Portanto,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

é a projeção ortogonal de $\mathbf{Y}$ sobre $\text{col}(\mathbf{X})$. Essa caracterização independe de qualquer hipótese probabilística.

## 14.6 Estrutura Matricial para Diagnóstico e Inferência

A formulação matricial do modelo linear não se limita à obtenção do estimador de mínimos quadrados. Ela estrutura integralmente os procedimentos de diagnóstico e prepara o terreno para a inferência estatística.

Recordemos a decomposição fundamental:

$$\mathbf{Y} = \mathbf{HY} + \mathbf{MY} = \hat{\mathbf{Y}} + \hat{\varepsilon}.$$

Essa identidade organiza o vetor de respostas em duas componentes ortogonais pertencentes a subespaços complementares.

### 14.6.1 Alavancagem e Estrutura da Matriz $\mathbf{H}$

A diagonal da matriz de projeção

$$h_{ii} = (\mathbf{H})_{ii}$$

mensura o quanto a $i$-ésima observação contribui estruturalmente para sua própria projeção.

Explicitamente,

$$h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i,$$

em que $\mathbf{x}_i^\top$ é a $i$-ésima linha da matriz $\mathbf{X}$.

Propriedades fundamentais:

- $0 \leq h_{ii} \leq 1$;
- $\sum_{i=1}^{n} h_{ii} = p + 1$;
- observações com valores elevados de $h_{ii}$ ocupam posições extremas no espaço das covariáveis.

### 14.6.2 Estrutura dos Resíduos

Os resíduos podem ser escritos como

$$\hat{\varepsilon} = \mathbf{M}\mathbf{Y}.$$

Como $\mathbf{M}$ é simétrica e idempotente,

$$\mathbf{M}^2 = \mathbf{M}, \qquad \mathbf{M}^\top = \mathbf{M}.$$

Além disso,

$$\mathbf{X}^\top \hat{\varepsilon} = \mathbf{0},$$

o que garante que os resíduos são ortogonais às colunas de $\mathbf{X}$.

Essa ortogonalidade fundamenta:

- a decomposição das somas de quadrados;
- a independência geométrica entre componentes ajustadas e residuais;
- a contagem de graus de liberdade.

### 14.6.3 Somas de Quadrados em Forma Matricial

A soma de quadrados total pode ser escrita como

$$\mathbf{Y}^\top \mathbf{Y}.$$

A soma de quadrados explicada pelo modelo é

$$\mathbf{Y}^\top \mathbf{H}\mathbf{Y}.$$

A soma de quadrados residual é

$$\mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

Como $\mathbf{H}$ e $\mathbf{M}$ projetam sobre subespaços ortogonais,

$$\mathbf{Y}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{H} \mathbf{Y} + \mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

Essa identidade é puramente algébrica e independe de qualquer suposição probabilística.

### 14.6.4 Postos e Graus de Liberdade

Como visto anteriormente,

$$\operatorname{tr}(\mathbf{H}) = p + 1, \qquad \operatorname{tr}(\mathbf{M}) = n - p - 1.$$

Para matrizes simétricas idempotentes, o traço coincide com o posto. Portanto:

- o subespaço ajustado tem dimensão $p + 1$;
- o subespaço residual tem dimensão $n - p - 1$.

Essas dimensões serão reinterpretadas, no contexto probabilístico, como graus de liberdade associados às somas de quadrados.

# 15 Distribuição Normal

Este apêndice tem um papel estrutural na fundamentação matemática dos modelos de regressão linear. A Distribuição Normal, especialmente em sua forma multivariada, fornece a base probabilística que torna possível derivar distribuições amostrais exatas para estimadores, contrastes lineares e estatísticas de teste em amostras finitas. Uma exposição formal e sistemática dessas propriedades pode ser encontrada em Anderson (2003) e Casella e Berger (2002).

A ideia central que deve acompanhar o leitor ao longo deste apêndice é a seguinte:

> Em regressão, não estudamos variáveis isoladas, mas vetores aleatórios e suas transformações lineares e quadráticas.

Essa perspectiva vetorial não é apenas notacional. Ela altera profundamente a forma de pensar sobre variabilidade, dependência e inferência.

## 15.1 Distribuição Normal Univariada

Uma variável aleatória $Y$ tem distribuição Normal univariada com média $\mu \in \mathbb{R}$ e variância $\sigma^2 > 0$ se sua função densidade (fpd) é

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad y \in \mathbb{R}.$$

Essa distribuição surge com frequência em modelagem estatística porque aparece como **distribuição limite** em muitos contextos, especialmente quando uma quantidade observada pode ser representada como a soma (ou média) de um grande número de contribuições aleatórias.

O **Teorema Central do Limite** estabelece que, sob condições adequadas, a soma devidamente padronizada de variáveis aleatórias independentes, ou fracamente dependentes, converge em distribuição para a Normal, independentemente da forma das distribuições individuais; essa fundamentação é essencialmente assintótica e aproximada, não constituindo uma identidade estrutural exata. Uma formulação rigorosa desse resultado pode ser consultada em Casella e Berger (2002).

No contexto de modelos de regressão, a suposição de Normalidade **não é obrigatória** nem define o modelo em si. Um modelo de regressão linear pode ser formulado sem qualquer

hipótese distributiva explícita sobre o erro, bastando condições sobre esperança, variância e independência.

A Normalidade é frequentemente adotada porque constitui o **caso mais simples e matematicamente tratável**, permitindo obter distribuições exatas para estimadores, estatísticas de teste e intervalos de confiança em amostras finitas. Em outros contextos, distribuições alternativas podem ser mais adequadas, levando a extensões naturais da regressão linear, como os modelos lineares generalizados.

Os parâmetros da Normal univariada admitem interpretações diretas, mas é importante compreendê-las com precisão estatística.

O parâmetro $\mu$ representa o **valor esperado teórico** da variável aleatória $Y$, isto é, o ponto em torno do qual a distribuição se concentra em média. Trata-se de uma quantidade populacional, definida independentemente de qualquer amostra específica, e que resume a tendência central do fenômeno sob o modelo probabilístico adotado.

O parâmetro $\sigma^2$ representa a **variância populacional** da variável aleatória, quantificando a dispersão em torno de $\mu$. Essa variabilidade reflete a incerteza inerente ao fenômeno modelado e não carrega, nesse estágio, qualquer interpretação ligada a explicação ou não explicação por covariáveis. Essa distinção só surgirá no contexto de modelos condicionais, como a regressão.

Essas interpretações ficam claras ao observarmos duas propriedades fundamentais da distribuição Normal:

- Esperança:
$$\mathbb{E}[Y] = \mu$$
- Variância:
$$\mathrm{Var}(Y) = \sigma^2$$

Essas igualdades não são meras convenções, mas decorrem da integração direta da densidade.

Uma característica estrutural importante da Normal é sua estabilidade por transformações lineares. Se $Y \sim N(\mu, \sigma^2)$ e definimos

$$Z = aY + b,$$

com $a \neq 0$, então

$$Z \sim N(a\mu + b, a^2\sigma^2).$$

Essa propriedade, demonstrada em Casella e Berger (2002), é o primeiro indício da importância da Normal na regressão: combinações lineares preservam a forma distributiva.

Uma transformação particularmente importante, tanto do ponto de vista teórico quanto prático, é a **padronização**. Definindo

$$Z = \frac{Y - \mu}{\sigma},$$

obtém-se uma nova variável aleatória com distribuição

$$Z \sim N(0, 1),$$

conhecida como **Normal padrão**.

A padronização desempenha um papel central em inferência estatística porque remove as unidades de medida e a escala original da variável, permitindo comparar desvios em termos relativos. Em modelos de regressão, essa ideia reaparece de forma sistemática: estatísticas de teste, resíduos padronizados e intervalos de confiança são construídos a partir de quantidades que mensuram desvios em relação a uma média teórica, expressos em unidades de desvio-padrão.

Assim, compreender profundamente o significado de $\mu$, $\sigma^2$ e da padronização é essencial para interpretar corretamente os resultados inferenciais que surgirão nos modelos de regressão.

## 15.2 Distribuição Normal Bivariada

Ao avançarmos para o caso bivariado, deixamos de estudar variáveis aleatórias isoladas e passamos a lidar explicitamente com **dependência entre variáveis aleatórias**. Esse é um passo conceitual fundamental, pois modelos estatísticos mais complexos e abrangentes, incluindo os modelos de regressão, são construídos exatamente a partir de relações entre variáveis.

Considere o vetor aleatório

$$\mathbf{Y} = (Y_1, Y_2)^\top.$$

Dizemos que $\mathbf{Y}$ segue uma **Distribuição Normal bivariada** se

$$\mathbf{Y} \sim N_2(\mu, \Sigma),$$

onde o vetor de médias é dado por

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

e a matriz de covariância é

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Neste ponto ocorre uma mudança conceitual importante. Enquanto no caso univariado a variância era um único número, agora a **matriz** $\Sigma$ passa a concentrar toda a informação sobre dispersão e dependência:

- os termos da diagonal ($\sigma_1^2$ e $\sigma_2^2$) descrevem a variabilidade individual de cada componente;
- os termos fora da diagonal descrevem a associação linear entre as variáveis, resumida pelo coeficiente de correlação $\rho$.

Assim, a estrutura de dependência entre $Y_1$ e $Y_2$ não é um elemento acessório, mas parte integrante da própria definição da distribuição conjunta.

A função densidade de probabilidade conjunta, como apresentado em Anderson (2003), pode ser escrita de forma compacta como

$$f(\mathbf{y}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\right\}.$$

Essa expressão merece uma leitura cuidadosa. O termo que aparece no expoente,

$$(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu),$$

é um escalar obtido a partir de vetores e matrizes, resultado de uma operação que combina transposição, multiplicação matricial e produto interno.

Neste momento, **não é necessário compreender formalmente esse termo como uma "forma quadrática"** essa noção será estudada com cuidado em um apêndice específico.

Intuitivamente, essa quantidade mede quão distante o vetor $\mathbf{y}$ está do centro $\mu$, mas **não usando a distância euclidiana usual**. Em vez disso, a distância é avaliada levando em conta a estrutura de variabilidade e dependência entre as componentes do vetor, codificada na matriz de covariância $\Sigma$.

Dessa forma, desvios ao longo de direções em que há maior variabilidade conjunta são penalizados de maneira diferente de desvios ao longo de direções com menor variabilidade. É essa ponderação que faz com que a distribuição apresente contornos elípticos, em vez de circulares.

A formalização matemática desse tipo de expressão, bem como seu papel central na regressão, nas somas de quadrados e nas estatísticas de teste, será apresentada posteriormente, quando estudarmos explicitamente as distribuições associadas a expressões desse tipo.

Geometricamente, isso se traduz no fato de que as curvas de mesma densidade dessa distribuição são **elipses centradas em** $\mu$.

A forma, o tamanho e a orientação dessas elipses dependem diretamente de $\Sigma$:

- quando $\rho = 0$, as elipses são alinhadas com os eixos coordenados;
- quando $\rho \neq 0$, as elipses tornam-se inclinadas, refletindo a associação linear entre $Y_1$ e $Y_2$.

Essa interpretação geométrica será essencial mais adiante, quando discutirmos **projeções, decomposições ortogonais e ajuste de modelos de regressão**, nos quais a ideia de "direções relevantes" no espaço dos dados desempenha papel central.

Mesmo nesse cenário conjunto, algumas propriedades permanecem familiares e ajudam a consolidar a intuição:

- As **distribuições marginais** continuam sendo Normais univariadas:

$$Y_1 \sim N(\mu_1, \sigma_1^2), \qquad Y_2 \sim N(\mu_2, \sigma_2^2).$$

Essas marginais mostram que, marginalmente, cada componente do vetor se comporta como uma variável Normal comum, mas isso **não elimina** a possibilidade de dependência entre elas quando observadas conjuntamente.

- As **distribuições condicionais** também são Normais:

$$Y_1 \mid Y_2 = y_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2),\, (1-\rho^2)\sigma_1^2\right).$$

Aqui aparece uma ideia conceitualmente profunda e extremamente importante para o que virá depois: **a média condicional de uma variável Normal é uma função linear da variável condicionante**.

Essa linearidade não é um artifício do modelo, nem uma escolha conveniente; ela é uma consequência direta da estrutura da Normalidade conjunta. Em modelos de regressão, essa propriedade será reinterpretada como a relação entre a resposta e as covariáveis, agora formulada de maneira explícita e sistemática.

Portanto, compreender a Normal bivariada é compreender, em um cenário simples, a origem probabilística da ideia de regressão como relação média condicional.

## 15.3 Distribuição Normal Multivariada

No caso geral, consideramos um vetor aleatório

$$\mathbf{Y} \in \mathbb{R}^n,$$

que segue uma **Distribuição Normal multivariada** se

$$\mathbf{Y} \sim N_n(\mu, \Sigma),$$

onde $\mu$ é o vetor de médias e $\Sigma$ é a matriz de covariância, simétrica e definida positiva.

A função densidade associada é

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\right\}.$$

Neste ponto, é importante fazer uma mudança consciente na forma de pensar. Não estamos mais lidando com observações isoladas, mas com **vetores aleatórios**, e a incerteza passa a ser descrita por **estruturas geométricas em espaços de dimensão maior**.

O vetor $\mu$ representa o centro da distribuição no espaço $\mathbb{R}^n$, enquanto a matriz $\Sigma$ determina como a variabilidade se organiza em torno desse centro. Mais especificamente, $\Sigma$ define:

- direções ao longo das quais a variabilidade conjunta é maior;
- direções ao longo das quais a variabilidade conjunta é menor;
- dependências lineares entre as componentes do vetor.

Essas direções não precisam coincidir com os eixos coordenados originais, e essa observação será fundamental quando discutirmos projeções e decomposições em regressão múltipla.

A expressão que aparece no expoente da densidade envolve novamente uma quantidade do tipo

$$(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu),$$

que produz um escalar a partir de vetores e matrizes. Assim como no caso bivariado, **não é necessário, neste momento, compreender formalmente essa expressão como uma forma quadrática**. Por ora, basta interpretar essa quantidade como uma medida de distância multivariada entre $\mathbf{y}$ e o centro $\mu$, ajustada pela estrutura de covariância.

Essa forma de medir distância explica por que as regiões de maior densidade da Normal multivariada são elipsoides em $\mathbb{R}^n$, generalizando as elipses vistas no caso bivariado.

Algumas propriedades fundamentais seguem diretamente dessa definição e merecem ser destacadas, pois reaparecerão continuamente ao longo do estudo de modelos de regressão.

A esperança e a covariância do vetor aleatório são dadas por

$$\mathbb{E}[\mathbf{Y}] = \mu, \qquad \mathrm{Cov}(\mathbf{Y}) = \Sigma.$$

Essas expressões formalizam a interpretação de $\mu$ como centro da distribuição e de $\Sigma$ como descrição completa da variabilidade conjunta.

Uma propriedade absolutamente central da Normal multivariada é sua **estabilidade por transformações lineares**. Se tomarmos uma transformação do tipo

$$\mathbf{Z} = \mathbf{A}\mathbf{Y} + \mathbf{a},$$

então a variável transformada também segue uma distribuição Normal multivariada:

$$\mathbf{Z} \sim N_m(\mathbf{A}\mu + \mathbf{a},\ \mathbf{A}\Sigma\mathbf{A}^\top).$$

Essa propriedade merece atenção especial. Ela afirma que **qualquer combinação linear de um vetor Normal multivariado continua sendo Normal**, independentemente da dimensão envolvida.

Esse resultado será a pedra angular da teoria de regressão linear. Quando estudarmos regressão, veremos que os estimadores dos coeficientes, os valores ajustados e diversos contrastes estatísticos são obtidos exatamente como transformações lineares do vetor de respostas. A Normalidade dessas quantidades decorre diretamente desta propriedade, e não de argumentos ad hoc.

Outra quantidade natural que surge no contexto da Normal multivariada é a chamada **distância de Mahalanobis**:

$$Q = (\mathbf{Y} - \mu)^\top \Sigma^{-1} (\mathbf{Y} - \mu),$$

para o qual vale

$$Q \sim \chi^2_n.$$

Mais uma vez, não é necessário aprofundar formalmente esse resultado neste momento. Conceitualmente, ele afirma que a distância multivariada entre $\mathbf{Y}$ e seu centro, quando devidamente padronizada pela matriz de covariância, possui uma distribuição conhecida.

Esse fato será explorado de forma sistemática em regressão, onde somas de quadrados, estatísticas de teste e medidas de ajuste surgirão como casos particulares desse tipo de expressão.

Assim, a Distribuição Normal multivariada fornece não apenas um modelo probabilístico para vetores de dados, mas também a base matemática para compreender por que as quantidades centrais da regressão admitem distribuições explícitas e interpretáveis.

## 15.4 Partição da Normal Multivariada

Um dos recursos mais poderosos da Normal multivariada é a possibilidade de **particionar o vetor aleatório** em blocos menores e ainda assim manter uma descrição probabilística completa e explícita.

Considere o vetor aleatório particionado como

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \sim N_n\left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

Aqui, a partição é puramente conceitual: estamos apenas reorganizando o vetor em dois blocos, sem alterar o modelo probabilístico subjacente. Ainda assim, essa simples reorganização permite responder a perguntas fundamentais sobre o comportamento do vetor aleatório.

Em particular, ela nos permite distinguir claramente dois tipos de informação:

- **comportamento marginal**, isto é, como cada bloco se distribui quando considerado isoladamente;
- **comportamento condicional**, isto é, como um bloco se distribui quando o outro é observado.

As distribuições marginais seguem diretamente da definição da Normal multivariada:

$$\mathbf{Y}_1 \sim N_{n_1}(\mu_1, \Sigma_{11}), \qquad \mathbf{Y}_2 \sim N_{n_2}(\mu_2, \Sigma_{22}).$$

Essas expressões mostram que, ao "olharmos apenas para uma parte do vetor", o comportamento probabilístico dessa parte continua sendo Normal, com média e covariância correspondentes aos blocos apropriados de $\mu$ e $\Sigma$ (Anderson (2003); Casella e Berger (2002)). No entanto, essa visão marginal ignora completamente a dependência entre os blocos.

A riqueza da Normal multivariada aparece de forma ainda mais clara ao analisarmos o comportamento **condicional**. A distribuição de $\mathbf{Y}_1$ dado que $\mathbf{Y}_2 = \mathbf{y}_2$ é

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2 \sim N_{n_1}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

Essa expressão concentra vários conceitos importantes em um único resultado.

Primeiro, observe que a **média condicional** de $\mathbf{Y}_1$ não é simplesmente $\mu_1$. Ela é ajustada pelo termo

$$\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2),$$

que incorpora a informação trazida pela observação de $\mathbf{Y}_2$. Esse ajuste depende exclusivamente da estrutura de covariância entre os blocos, e não de escolhas arbitrárias de modelagem.

Segundo, note que a **matriz de covariância condicional**

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

é sempre menor, no sentido de variância, do que a covariância marginal $\Sigma_{11}$. Isso formaliza matematicamente uma ideia intuitiva: **ao observar parte do vetor, reduzimos a incerteza sobre o restante**.

Esse resultado mostra que, na Normal multivariada, o condicionamento produz dois efeitos simultâneos e bem definidos:

- a média é deslocada de forma linear em função da parte observada;
- a variabilidade é reduzida de maneira controlada pela estrutura de dependência.

Essas duas propriedades; linearidade da média condicional e redução da variância, não são hipóteses adicionais nem aproximações, elas são consequências diretas da Normalidade conjunta.

Embora ainda não estejamos estudando modelos de regressão, é importante registrar que essa lógica será reinterpretada mais adiante quando os coeficientes de um modelo passarem a ser entendidos como **efeitos condicionais**, isto é, como variações esperadas em uma componente do vetor quando outras são mantidas fixas.

Assim, a partição da Normal multivariada fornece o arcabouço probabilístico que sustenta a noção de regressão como estudo de relações condicionais, mesmo antes de qualquer equação de regressão ser escrita explicitamente.

## 15.5 Covariância zero implica independência

Em geral, para vetores aleatórios arbitrários, a condição de covariância nula **não implica independência**. Isto é, pode ocorrer que duas variáveis tenham covariância igual a zero e, ainda assim, sejam dependentes.

Entretanto, a família Normal possui uma propriedade estrutural especial:

> Se um vetor aleatório é Normal multivariado, então quaisquer componentes (ou combinações lineares de componentes) que tenham covariância zero são independentes.

Mais precisamente, se

$$\mathbf{Y} \sim N_n(\mu, \Sigma)$$

e $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, então

$$\mathrm{Cov}(\mathbf{a}^\top \mathbf{Y},\, \mathbf{b}^\top \mathbf{Y}) = 0 \quad \Longrightarrow \quad \mathbf{a}^\top \mathbf{Y} \text{ e } \mathbf{b}^\top \mathbf{Y} \text{ são independentes.}$$

Essa propriedade é específica da distribuição Normal e não vale em geral para outras distribuições multivariadas. Uma demonstração pode ser encontrada em Anderson (2003) e em Casella e Berger (2002).

Essa característica será essencial na regressão linear, pois permite concluir, sob Normalidade dos erros, que:

- o estimador dos coeficientes é independente do vetor de resíduos;
- diferentes somas de quadrados associadas a projeções ortogonais são independentes;
- estatísticas baseadas em decomposições ortogonais possuem distribuições independentes.

A independência decorre da ortogonalidade geométrica no espaço das observações, combinada com a estrutura da Normal multivariada.

### 15.5.1 Independência no caso bivariado

No caso particular bivariado, seja

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_2(\mu, \Sigma),$$

com

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Então vale a equivalência:

$$\rho = 0 \quad \Longleftrightarrow \quad Y_1 \text{ e } Y_2 \text{ são independentes.}$$

Essa equivalência é uma consequência direta da forma explícita da densidade conjunta e constitui uma propriedade distintiva da Normal bivariada. Em distribuições gerais, correlação zero não implica independência.

### 15.5.2 Caso marginal

Se

$$\mathbf{Y} \sim N_n(\mu, \Sigma),$$

então qualquer subconjunto de componentes de $\mathbf{Y}$ também possui distribuição Normal multivariada.

Formalmente, se particionarmos

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix},$$

então as distribuições marginais são

$$\mathbf{Y}_1 \sim N_{n_1}(\mu_1, \Sigma_{11}), \qquad \mathbf{Y}_2 \sim N_{n_2}(\mu_2, \Sigma_{22}).$$

Essa estabilidade marginal é consequência direta da definição da Normal multivariada e pode ser verificada integrando-se a densidade conjunta ou utilizando o resultado de que combinações lineares preservam Normalidade.

### 15.5.3 Partições e Independência entre Blocos

Considere novamente a partição

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \sim N_n\left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

Então:

$$\Sigma_{12} = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{Y}_1 \text{ e } \mathbf{Y}_2 \text{ são independentes.}$$

Isto é, na Normal multivariada, blocos são independentes se, e somente se, sua matriz de covariância cruzada for nula.

Essa propriedade tem papel central na teoria da regressão linear clássica. Quando se demonstra que duas quantidades são obtidas por projeções ortogonais e que a matriz de covariância cruzada entre elas é nula, a Normalidade garante automaticamente independência.

Essa combinação entre:

- ortogonalidade algébrica,
- covariância nula,
- estrutura Normal,

é o mecanismo matemático que sustenta a independência entre soma de quadrados do modelo e soma de quadrados do erro, fundamento da estatística $F$.

## 15.6 Papel da Distribuição Normal na fundamentação dos modelos de regressão

Os resultados apresentados neste apêndice fornecem uma base probabilística razoável para a formulação e a análise dos modelos clássicos de regressão linear. O objetivo aqui é explicitar as estruturas matemáticas que a tornam analisável de forma rigorosa.

Em modelos de regressão linear com erros normalmente distribuídos, considera-se que o vetor de respostas pode ser escrito como

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

o que implica diretamente que

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n).$$

como discutido em Kutner et al. (2005). Essa especificação não define o modelo de regressão em si, que pode ser formulado sob hipóteses mais gerais, mas estabelece um **caso fundamental** no qual resultados exatos de inferência podem ser obtidos em amostras finitas.

A partir dessa estrutura probabilística decorrem, de forma sistemática, várias propriedades centrais da regressão linear clássica:

- os estimadores dos coeficientes surgem como **transformações lineares** do vetor aleatório $\mathbf{Y}$;
- os resíduos e as somas de quadrados associadas ao ajuste do modelo surgem como **expressões quadráticas** em $\mathbf{Y}$;
- as distribuições amostrais das estatísticas utilizadas para inferência são obtidas a partir das distribuições dessas transformações lineares e quadráticas.

Deste modo, estatísticas do tipo $t$ e $F$ não são introduzidas de maneira ad hoc, mas emergem naturalmente da combinação entre a Normal multivariada e as operações algébricas realizadas sobre o vetor de respostas.

Outros apêndices exploram explicitamente essas estruturas, estudando as distribuições associadas a transformações lineares e quadráticas de vetores aleatórios conjuntamente normais. Esse desenvolvimento permitirá compreender, de forma unificada, a origem das principais ferramentas inferenciais utilizadas em modelos de regressão linear.

# 16 Formas Lineares e Quadráticas na Normal Multivariada

Em regressão linear clássica, muitos estimadores e estatísticas fundamentais são **formas lineares** e **formas quadráticas** de um vetor aleatório Normal multivariado. Essa conexão não é meramente técnica: ela é estrutural e explica por que distribuições exatas em amostras finitas podem ser obtidas sob normalidade. A fundamentação matricial e geométrica desses resultados pode ser encontrada, em diferentes níveis de formalidade, em Harville (1997) e Anderson (2003).

## 16.1 Preliminares: Normal multivariada e transformações lineares

Seja

$$\mathbf{Y} \sim N_n(\mu, \Sigma),$$

com $\Sigma$ simétrica definida positiva.

Uma propriedade central da Normal multivariada é sua **estabilidade por transformações lineares**: se $\mathbf{A}$ é uma matriz fixa $m \times n$ e $\mathbf{a} \in \mathbb{R}^m$, então

$$\mathbf{Z} = \mathbf{A}\mathbf{Y} + \mathbf{a} \sim N_m(\mathbf{A}\mu + \mathbf{a},\, \mathbf{A}\Sigma\mathbf{A}^\top).$$

Esse resultado é apresentado de forma sistemática em Anderson (2003) e constitui a base probabilística da teoria da regressão linear.

Como caso particular, para um vetor fixo $\mathbf{c} \in \mathbb{R}^n$,

$$L = \mathbf{c}^\top \mathbf{Y} \sim N(\mathbf{c}^\top \mu,\, \mathbf{c}^\top \Sigma \mathbf{c})\,.$$

A dedução segue diretamente da propriedade anterior e também pode ser vista como aplicação do fato de que combinações lineares de vetores Normais permanecem Normais, como discutido em Casella e Berger (2002).

Essa estrutura será aplicada diretamente a:

- $\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$;
- contrastes $\mathbf{C}\widehat{\beta}$;
- predições lineares e combinações de coeficientes.

## 16.2 Padronização multivariada e redução ao caso I

Para estudar formas quadráticas, é útil reduzir o problema ao caso esférico.

Como $\Sigma$ é definida positiva, existe uma matriz simétrica definida positiva $\Sigma^{1/2}$ tal que

$$\Sigma^{1/2}\Sigma^{1/2} = \Sigma, \qquad (\Sigma^{1/2})^{-1} = \Sigma^{-1/2}.$$

A existência dessa raiz quadrada decorre da decomposição espectral de matrizes simétricas definidas positivas, tratada em detalhe em Harville (1997).

Defina

$$\mathbf{Z} = \Sigma^{-1/2}(\mathbf{Y} - \mu).$$

Então

$$\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n).$$

Essa transformação separa de maneira conceitualmente limpa os papéis de média e dispersão, reduzindo a análise probabilística ao caso esférico. Além disso, ela mostra que a chamada distância de Mahalanobis

$$(\mathbf{Y} - \mu)^\top \Sigma^{-1} (\mathbf{Y} - \mu)$$

é simplesmente a norma euclidiana ao quadrado de um vetor Normal padrão:

$$(\mathbf{Y} - \mu)^\top \Sigma^{-1} (\mathbf{Y} - \mu) = \mathbf{Z}^\top \mathbf{Z}.$$

Essa ponte entre geometria (normas e projeções) e distribuição (Qui-quadrado) é central na teoria da inferência sob normalidade.

## 16.3 Forma linear: distribuição, interpretação e conexão com regressão

Uma **forma linear** de um vetor aleatório é uma expressão do tipo

$$L = \mathbf{c}^\top \mathbf{Y},$$

onde $\mathbf{c}$ é fixo.

Se $\mathbf{Y} \sim N_n(\mu, \Sigma)$, então:

- $L$ é Normal;
- $\mathbb{E}(L) = \mathbf{c}^\top \mu$;
- $\mathrm{Var}(L) = \mathbf{c}^\top \Sigma \mathbf{c}$.

Na regressão linear, quando se assume normalidade dos erros, tanto $\widehat{\beta}$ quanto qualquer contraste $\mathbf{C}\widehat{\beta}$ são formas lineares em $\mathbf{Y}$. Por isso, possuem distribuição Normal exata em amostras finitas.

## 16.4 Forma quadrática: definição e interpretação

Uma **forma quadrática** em $\mathbf{Y}$ é uma expressão do tipo

$$Q = \mathbf{Y}^\top \mathbf{A} \mathbf{Y},$$

onde $\mathbf{A}$ é uma matriz fixa $n \times n$.

Primeira observação fundamental: apenas a parte simétrica de $\mathbf{A}$ influencia $Q$. De fato,

$$\mathbf{Y}^\top \mathbf{A} \mathbf{Y} = \mathbf{Y}^\top \left( \frac{\mathbf{A} + \mathbf{A}^\top}{2} \right) \mathbf{Y}.$$

Logo, pode-se assumir $\mathbf{A}$ simétrica sem perda de generalidade, fato discutido na literatura matricial estatística como em Harville (1997).

Em regressão, as somas de quadrados são precisamente formas quadráticas:

$$\text{SQReg} = \mathbf{Y}^\top \mathbf{H} \mathbf{Y}, \qquad \text{SQRes} = \mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

O comportamento probabilístico de $Q$ depende criticamente das propriedades estruturais de $\mathbf{A}$. Quando $\mathbf{A}$ é uma projeção ortogonal (simétrica e idempotente), a forma quadrática se reduz a uma soma de quadrados de componentes Normais padrão em um subespaço.

## 16.5 O caso central: $\mathbf{Z}^\top \mathbf{Z}$ e a distribuição $\chi^2$

Se $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, então suas componentes são independentes e

$$Z_i \sim N(0, 1).$$

Consequentemente,

$$\mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

Esse resultado é a base de todas as somas de quadrados na regressão sob normalidade, como apresentado em textos de inferência como Casella e Berger (2002).

## 16.6 Projeções ortogonais e Qui-quadrado

Seja $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ e seja $\mathbf{A}$ simétrica e idempotente:

$$\mathbf{A}^\top = \mathbf{A}, \qquad \mathbf{A}^2 = \mathbf{A}.$$

Se $r = \operatorname{rank}(\mathbf{A}) = \operatorname{tr}(\mathbf{A})$, então

$$\mathbf{Z}^\top \mathbf{A} \mathbf{Z} \sim \chi^2_r.$$

Esse resultado pode ser deduzido via decomposição espectral de $\mathbf{A}$ e é tratado com rigor em Harville (1997) e Rencher e Christensen (2012).

Geometricamente, $\mathbf{A}$ projeta sobre um subespaço $S$ de dimensão $r$. Assim, $\mathbf{Z}^\top \mathbf{A} \mathbf{Z}$ mede o comprimento ao quadrado da componente projetada de $\mathbf{Z}$ em $S$, que se comporta como soma de quadrados de $r$ Normais padrão.

Esse resultado aplica-se diretamente às matrizes $\mathbf{H}$ (posto $p+1$) e $\mathbf{M}$ (posto $n-p-1$). Quando $\mathbf{Z}$ representa a versão padronizada do vetor de erros, as somas de quadrados do modelo e do resíduo assumem distribuições Qui-quadrado com graus de liberdade dados por esses postos.

## 16.7 Independência via ortogonalidade

Se $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ e $\mathbf{A}$, $\mathbf{B}$ são simétricas idempotentes tais que

$$\mathbf{A}\mathbf{B} = \mathbf{0},$$

então

$$\mathbf{Z}^\top \mathbf{A} \mathbf{Z} \;\perp\; \mathbf{Z}^\top \mathbf{B} \mathbf{Z}.$$

A independência decorre da decomposição ortogonal do espaço combinada com a simetria esférica da Normal padrão.

Critério análogo vale para independência entre forma linear $L = \mathbf{a}^\top \mathbf{Z}$ e forma quadrática $Q = \mathbf{Z}^\top \mathbf{A} \mathbf{Z}$: se

$$\mathbf{A}\mathbf{a} = \mathbf{0},$$

então $L$ e $Q$ são independentes.

## 16.8 MRLM Normal

Considere o modelo de regressão linear múltipla sob normalidade e homocedasticidade:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Como consequência direta da estabilidade da Normal multivariada por transformações lineares, tem-se

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n).$$

As quantidades centrais do modelo podem ser classificadas estruturalmente da seguinte forma, destacando-se suas distribuições quando há uma forma conhecida.

### 16.8.1 Estimador como forma linear em $\mathbf{Y}$

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Trata-se de uma transformação linear do vetor aleatório $\mathbf{Y}$. Logo,

$$\hat{\beta} \sim N_{p+1}\left(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right).$$

Em particular, para cada componente $\hat{\beta}_j$,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}} \sim N(0,1).$$

### 16.8.2 Resíduos como transformação linear

$$\hat{\varepsilon} = \mathbf{M}\mathbf{Y}, \qquad \mathbf{M} = \mathbf{I}_n - \mathbf{H}.$$

Como $\hat{\varepsilon}$ é transformação linear de um vetor Normal multivariado, segue que

$$\hat{\varepsilon} \sim N_n\left(\mathbf{M}\mathbf{X}\beta,\, \sigma^2 \mathbf{M}\right).$$

Como $\mathbf{M}\mathbf{X} = \mathbf{0}$, obtém-se

$$\hat{\varepsilon} \sim N_n(\mathbf{0},\, \sigma^2 \mathbf{M}).$$

Assim, os resíduos possuem estrutura Normal degenerada em um subespaço de dimensão $n-p-1$.

### 16.8.3 Soma de quadrados residual como forma quadrática

$$\text{SQRes} = \hat{\varepsilon}^\top \hat{\varepsilon} = \mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

Como $\mathbf{M}$ é simétrica idempotente com

$$\text{rank}(\mathbf{M}) = n - p - 1,$$

segue que, após padronização por $\sigma^2$,

$$\frac{\text{SQRes}}{\sigma^2} = \frac{\mathbf{Y}^\top \mathbf{M} \mathbf{Y}}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Equivalentemente,

$$\text{SQRes} \sim \sigma^2 \, \chi^2_{n-p-1}.$$

Além disso, definindo

$$\hat{\sigma}^2 = \frac{\text{SQRes}}{n-p-1},$$

tem-se

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

### 16.8.4 Decomposição ortogonal da soma de quadrados total

$$\mathbf{Y}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{H} \mathbf{Y} + \mathbf{Y}^\top \mathbf{M} \mathbf{Y}.$$

Sob o modelo Normal, após padronização por $\sigma^2$, as duas parcelas tornam-se formas quadráticas associadas a projeções ortogonais complementares.

Escrevendo $\mu = \mathbf{X}\beta$, obtém-se:

- **Parte ajustada (forma quadrática não-central)**

$$\frac{\mathbf{Y}^\top \mathbf{H}\mathbf{Y}}{\sigma^2} \sim \chi^2_{p+1}(\lambda), \qquad \lambda = \frac{\mu^\top \mu}{\sigma^2}.$$

- **Parte residual (forma quadrática central)**

$$\frac{\mathbf{Y}^\top \mathbf{M}\mathbf{Y}}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Como $\mathbf{HM} = \mathbf{0}$ e ambas são projeções ortogonais, essas duas formas quadráticas são independentes:

$$\frac{\mathbf{Y}^\top \mathbf{H}\mathbf{Y}}{\sigma^2} \perp \frac{\mathbf{Y}^\top \mathbf{M}\mathbf{Y}}{\sigma^2}.$$

### 16.8.5 Surgimento das distribuições $t$ e $F$

A partir dessas distribuições, emergem naturalmente as estatísticas de inferência.

Para cada coeficiente,

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}} = \frac{\dfrac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}}}}{\sqrt{\frac{\mathrm{SQRes}/\sigma^2}{n-p-1}}} \sim t_{\,n-p-1}.$$

A estatística $t$ surge como razão entre uma Normal padrão e a raiz de uma Qui-quadrado independente dividida por seus graus de liberdade.

De modo análogo, a estatística global de significância do modelo pode ser escrita como

$$F = \frac{(\mathbf{Y}^\top \mathbf{H}\mathbf{Y})/(p+1)}{(\mathbf{Y}^\top \mathbf{M}\mathbf{Y})/(n-p-1)} \sim F_{p+1,\,n-p-1},$$

no caso central sob a hipótese nula apropriada.

Assim, as distribuições $t$ e $F$ não são introduzidas ad hoc: elas emergem diretamente da estrutura geométrica das projeções ortogonais combinada com a Normalidade multivariada do vetor de respostas.

# 17 Tratamento de dados para regressão (pré-modelagem)

Este apêndice organiza, em forma de roteiro estruturado, os procedimentos que devem ser realizados **antes** do ajuste de qualquer modelo de regressão. O objetivo é preparar bases de dados tabulares (em que linhas representam observações e colunas representam variáveis) para que estejam **coerentes, consistentes, completas e adequadas à modelagem estatística**.

Os métodos de regressão (lineares, generalizados ou penalizados) **pressupõem** que a variabilidade observada nos dados represente o fenômeno real sob estudo, e não erros de registro, inconsistências de escala ou problemas de codificação. O tratamento pré-modelagem constitui, portanto, o **primeiro passo do raciocínio estatístico aplicado**: garantir que o modelo descreva o mundo observado e não artefatos do processo de coleta ou organização dos dados.

**Princípios norteadores da preparação de dados**

1. **Parcimônia:** trate o necessário, não o possível. Cada transformação modifica o significado estatístico das variáveis e pode alterar a interpretação dos resultados.
2. **Transparência:** todo procedimento deve ser reprodutível e documentado (log de alterações + dicionário de variáveis).
3. **Consistência:** mantenha coerência entre bases, períodos e versões (nomes de variáveis, unidades de medida, tipos e categorias).
4. **Domínio do contexto:** conheça o fenômeno estudado antes de decidir se determinado valor é erro, exceção legítima ou padrão relevante.
5. **Validação cruzada:** confirme decisões de tratamento por mais de uma ótica — estatística e substantiva. (Iannone (2024))

**Ferramentas e escopo**

- Trabalharemos com **R** (por exemplo, utilizando `tidyverse` ou `data.table`) e com **dados estruturados em formato tabular**.
- **Não** abordaremos, neste apêndice, dados não estruturados ou semi-estruturados (texto livre, JSON heterogêneo, imagens, áudio ou bases documentais).
- Quando utilizarmos expressões como "detectar", "verificar" ou "avaliar", estaremos nos referindo a **operações descritivas e exploratórias**, sem ajuste formal de modelos nesta etapa.

**Fluxo geral do processo analítico**

coleta → organização → tratamento → **(entregável: base tratada + dicionário + log de decisões)** → modelagem → diagnóstico → interpretação

O tratamento de dados é, portanto, uma etapa intermediária e estruturante, que conecta a coleta bruta à modelagem estatística.

## 17.1 Exemplos do dia a dia e Checklist rápido

**Exemplos rápidos do dia a dia**

- **CSV que não abre "corretamente"**: o arquivo foi salvo com `;` em vez de `,` como separador, e o decimal usa `,` (ex.: `3,14`). Resultado: números são interpretados como texto.
  **Tratamento**: especificar corretamente o delimitador e o símbolo decimal na leitura; padronizar separadores.
- **Acentos e caracteres estranhos**: colunas aparecem como `ação`, `aÃ§Ã£o` ou nomes quebrados.
  **Tratamento**: ajustar o `encoding` (ex.: UTF-8 vs latin-1) e padronizar nomes de variáveis.
- **Tipos incorretos**: idades lidas como texto (`"21"`), datas como strings (`"2025-10-24"`) e variáveis 0/1 lidas como numéricas quando deveriam ser categóricas. **Tratamento**: converter explicitamente para os tipos adequados (numérico, data/hora, fator) e validar o resultado.
- **Códigos de "faltante" disfarçados**: valores como `999`, `-1`, `NA`, `""` representam ausência, mas estão misturados com dados válidos. (Wickham et al. (2024))
  **Tratamento**: recodificar para ausentes padronizados e decidir entre excluir, imputar ou manter com justificativa.
- **Duplicatas e chaves quebradas**: a mesma unidade amostral aparece repetida; totais e médias ficam incorretos.
  **Tratamento**: identificar chaves únicas, remover ou conciliar duplicatas e documentar a decisão.
- **Categorias inconsistentes**: `Masculino`, `M`, `masc` e `male` aparecem como níveis distintos.
  **Tratamento**: unificar rótulos, definir categoria de referência substantiva e gerar dummies adequadamente.
- **Datas em formatos mistos**: `31/12/2025` e `12-31-2025` na mesma coluna.
  **Tratamento**: normalizar formato, verificar timezone quando relevante e extrair componentes (ano/mês/dia) se necessário.

- **Valores impossíveis**: `altura = -5`, `proporção > 1`, `idade = 300`.
  **Tratamento**: definir faixas válidas segundo o domínio e corrigir ou descartar observações inconsistentes.
- **Escalas heterogêneas**: receita em reais e custo em milhares de reais; área em $m^2$ e $km^2$.
  **Tratamento**: padronizar unidades ou aplicar reescala/padronização (ex.: z-score).
- **Outliers evidentes**: uma observação muito superior às demais; zero estrutural inesperado.
  **Tratamento**: verificar plausibilidade no contexto, decidir entre correção, winsorização, transformação ou manutenção justificada.
- **Resposta incompatível com o objetivo**: deseja-se regressão linear contínua, mas a variável resposta é binária ou contagem.
  **Tratamento**: verificar se o tipo de resposta é compatível com o modelo pretendido (linear, logístico, Poisson etc.).

Observe que, assim como a cozinha precisa estar organizada para a receita "dar certo", a base precisa estar **coerente, completa e padronizada** para que a modelagem produza resultados interpretáveis e estatisticamente válidos.

**Checklist rápido: Faça antes de modelar**

1. **Defina o objetivo e o tipo de resposta**: contínua, binária ou contagem; confirme que a variável resposta é compatível com o modelo pretendido.
2. **Garanta leitura correta**: verifique separador, encoding e símbolo decimal; confirme que variáveis numéricas não foram importadas como texto.
3. **Acerte tipos e unidades**: converta datas, inteiros e reais; padronize unidades e nomes de variáveis.
4. **Mapeie e trate dados faltantes**: identifique `NA`, vazios e códigos artificiais (ex.: 999); decida entre excluir, imputar ou manter com justificativa. (Wickham et al. (2024))
5. **Elimine inconsistências**: remova ou una duplicatas, valide chaves, corrija valores impossíveis.
6. **Cuide de outliers**: identifique pontos extremos; corrija, transforme ou mantenha. Todas as ações com outliers devem ser realizadas com justificativa documentada.
7. **Codifique variáveis qualitativas**: unifique rótulos, defina categoria de referência e evite colinearidade perfeita na matriz de projeto.
8. **Cheque condições mínimas para regressão**:
   - Variabilidade das covariáveis;
   - Ausência de colinearidade perfeita ($X^\top X$ não singular);
   - Número de observações maior que o número de parâmetros ($n > p + 1$);

- Gere sumários e gráficos básicos;

- **Salve a versão tratada com dicionário de variáveis e log de decisões.**

**Cartões de decisão**

- **Faltantes:**
  Qual a proporção? A variável é essencial? Qual o mecanismo provável (MCAR/MAR/MNAR)? (Little e Rubin (2019))
  → excluir / imputar (média, mediana, por grupo, métodos múltiplos) / manter com justificativa. (Kuhn et al. (2024); Buuren e Groothuis-Oudshoorn (2024))

- **Outliers:**
  Erro de digitação ou unidade? Valor plausível no domínio?
  → corrigir / winsorizar / transformar / manter e justificar.

- **Categóricas:**
  Existem níveis raros (alta cardinalidade)?
  → agrupar em "outros"; escolher referência substantiva; padronizar rótulos.

- **Escalas:**
  Magnitudes muito diferentes entre covariáveis?
  → aplicar padronização (z-score) ou reescala; revisar unidades.

- **Reprodutibilidade:**
  Toda decisão foi registrada no **log de tratamento**?

##Leitura e organização dos dados

Ler corretamente um arquivo de dados (formato, delimitador, encoding, símbolo decimal) é a primeira barreira contra erros que podem comprometer toda a análise. Uma leitura incorreta pode fazer com que números sejam interpretados como texto, datas como caracteres ou colunas inteiras sejam deslocadas. Esses problemas, se não detectados no início, propagam-se até a modelagem e podem invalidar inferências.

Em regressão, erros de tipagem e leitura afetam diretamente a construção da matriz de projeto $\mathbf{X}$ e da variável resposta $\mathbf{y}$. Portanto, esta etapa é estrutural, não meramente operacional.

**Conexões com a literatura e boas práticas**

Autores como Charnet et al. (2008) e Sheather (2009) enfatizam que a qualidade dos dados influencia diretamente as estimativas, testes e intervalos de confiança. Essa etapa integra o que se denomina **análise exploratória de dados (AED)**: conhecer o material empírico antes de ajustar qualquer modelo.

Explorar não é modelar, é compreender a estrutura do dado.

**Funções úteis no R**

Pacotes e funções frequentemente utilizados nessa etapa incluem:

- `readr`: `read_csv()`, `read_delim()`, `guess_encoding()`, `locale()`
- `readxl`: `read_excel()`
- `dplyr`: `glimpse()`, `summarise()`, `count()`
- `fs` e `here`: organização de diretórios e caminhos reproduzíveis
- `vroom`: leitura rápida de arquivos grandes
- `stringi`: diagnóstico e conversão de encoding

O objetivo não é "usar funções", mas garantir que a base esteja corretamente interpretada pelo R antes de qualquer transformação.

**O que fazer**

- Mapear a **origem dos dados** (site, repositório, disciplina, experimento) e sua **licença de uso**.
- Conferir o **formato do arquivo** (CSV, XLSX, SAV, DTA etc.) e o **delimitador** utilizado (`,`, `;`, tabulação).
- Ajustar corretamente o **encoding** (UTF-8, latin-1) e o **símbolo decimal** (`,` ou `.`).
- Definir um **esquema de pastas do projeto** (por exemplo: `dados_brutos/`, `dados_interinos/`, `dados_tratados/`) e convenções padronizadas de nomes.

**Boas práticas**

- Manter a pasta `dados_brutos/` **imutável**. Nunca sobrescrever dados originais.
- Realizar todas as modificações em cópias armazenadas em `dados_tratados/`.
- Padronizar nomes de colunas: minúsculas, sem acento, em formato `snake_case`.
- Criar um arquivo `README_dados.md` com:
  - Fonte dos dados
  - Data de download
  - Responsável
  - Descrição geral das variáveis

**Erros comuns (e como evitar)**

- Arquivo CSV com `;` como separador lido como se fosse `,` → verificar explicitamente o delimitador.
- Valores como `1,23` interpretados como texto → ajustar `locale(decimal_mark = ",")` ou converter adequadamente.
- Datas ambíguas (ex.: `01/02/2020`) → padronizar formato (preferencialmente ISO 8601: `2020-02-01`).
- Colunas numéricas importadas como `character` → converter explicitamente e validar.

**Checklist rápido**

Antes de seguir para qualquer transformação ou modelagem, verifique:

☐ Colunas numéricas estão realmente no tipo numérico?
☐ Datas foram reconhecidas como datas?
☐ Há duplicatas segundo a chave da base?
☐ Nomes de colunas estão padronizados?
☐ A leitura preservou o número esperado de linhas e colunas?

Se qualquer resposta for negativa, o tratamento ainda não começou e a leitura ainda não terminou.

## 17.2 Estrutura e tipagem das variáveis

A estrutura e o tipo das variáveis determinam como a matriz de projeto $\mathbf{X}$ será construída. Uma tipagem incorreta não é apenas um erro técnico: ela altera o significado estatístico do modelo, pode introduzir colinearidade artificial e comprometer estimativas, testes e interpretações.

De acordo com Sheather (2009), a correta identificação do tipo de cada variável é condição essencial para que o modelo represente adequadamente a relação entre preditores e resposta. Uma variável categórica tratada como numérica impõe uma estrutura inexistente; uma variável numérica tratada como categórica multiplica desnecessariamente parâmetros.

Em regressão, a tipagem define como cada coluna entra em $\mathbf{X}$: como valor contínuo, como conjunto de dummies ou como transformação temporal.

**Ferramentas R úteis**

Funções base do R:

- `str()` — inspeciona a estrutura da base
- `class()` — verifica o tipo de objeto
- `as.numeric()`, `as.integer()` — conversões numéricas
- `as.Date()` — conversão de datas

Pacotes auxiliares:

- `lubridate`: `ymd()`, `dmy()`, `ymd_hms()`
- `dplyr`: `mutate()`, `across()`
- `forcats`: manipulação de fatores

Para validação: - `is.na()` — valores ausentes
- `is.finite()` — valores numéricos válidos
- Verificações de faixa (ex.: idade  0)

O objetivo não é apenas converter, mas **verificar se a conversão preserva o significado substantivo da variável**.

**Classifique as variáveis**

Antes de qualquer modelagem, identifique claramente:

- Numéricas **contínuas** (salário, temperatura, peso)
- Numéricas **discretas** (número de visitas, contagem de eventos)
- **Binárias** (0/1, sim/não)
- **Categóricas nominais** (sexo, região)
- **Categóricas ordinais** (baixo, médio, alto)
- **Temporais** (datas, horários)
- **Identificadores (IDs)** que não devem entrar como preditores numéricos

Identificadores numéricos (ex.: matrícula, CPF, código do lote) **não são variáveis quantitativas** e não devem ser tratados como tal na regressão.

**Ajustes típicos**

- Converter strings numéricas para número de forma explícita.
- Converter datas para classe apropriada e, quando necessário, extrair componentes (ano, mês).
- Transformar variáveis 0/1 em fator quando a interpretação for categórica.
- Definir fatores com níveis claros e ordenados quando houver estrutura ordinal.
- Padronizar rótulos inconsistentes (ex.: `M`, `Masculino`, `male` → um único padrão).

Após os ajustes, reavalie a estrutura da base e confirme que cada coluna está no tipo esperado antes de avançar para o tratamento de faltantes ou criação de dummies.

## 17.3 Tratamento de dados faltantes (missing)

Dados faltantes alteram o tamanho efetivo da amostra, modificam a estrutura da matriz de projeto $\mathbf{X}$ e podem introduzir viés nas estimativas quando o mecanismo de ausência não

é completamente aleatório (MCAR). Em regressão, a presença de faltantes pode funcionar como um **mecanismo implícito de seleção amostral**, afetando tanto a estimação quanto a interpretação dos coeficientes.

Excluir observações com valores ausentes equivale, muitas vezes, a analisar uma subamostra potencialmente não representativa.

**Erros clássicos**

- Imputar zero indiscriminadamente (por exemplo, via `replace_na`) sem considerar o significado substantivo.
- Realizar junções (*joins*) entre tabelas sem validar chaves, criando duplicações artificiais.
- Remover colunas inteiras apenas por conterem `NA`, sem avaliar sua relevância analítica.
- Imputar a variável resposta sem justificativa metodológica.

**Sugestões da literatura**

Segundo Little e Rubin (2019), o tratamento adequado depende do mecanismo de ausência:

- **MCAR (Missing Completely At Random)**: ausência independente de variáveis observadas e não observadas.
- **MAR (Missing At Random)**: ausência depende apenas de variáveis observadas.
- **MNAR (Missing Not At Random)**: ausência depende de informação não observada.

Em situações simples e com baixa proporção de faltantes, imputações por média ou mediana podem ser aceitáveis como aproximação. Em contextos mais complexos, recomenda-se imputação múltipla ou métodos baseados em modelos, preservando a incerteza associada ao processo.

**Como minimizar problemas**

- Mapear sistematicamente os valores ausentes com `is.na()` e quantificar proporções por variável.
- Identificar códigos artificiais de ausência (`999`, `-1`, `""`) e recodificá-los para `NA`.
- Avaliar a importância substantiva da variável antes de decidir pela exclusão.
- Documentar cada remoção ou imputação realizada.
- Para duplicatas associadas a junções, utilizar verificações de chave e funções como `duplicated()` ou `distinct()`.

A decisão deve ser estatística **e** substantiva.

**Boas práticas**

- Manter um log explícito das decisões tomadas.
- Comparar estatísticas descritivas antes e depois da imputação.
- Evitar alterar a distribuição da variável de forma não justificada.
- Não imputar automaticamente a variável resposta nesta etapa, salvo sob estratégia metodológica claramente definida.

**Procedimento recomendado**

1. Calcular a taxa de faltantes por coluna.
2. Identificar padrões estruturais de ausência.
3. Classificar o possível mecanismo (MCAR/MAR/MNAR).
4. Definir estratégia:
    - excluir observações,
    - imputar (média, mediana, por grupo ou múltipla),
    - manter e modelar posteriormente o mecanismo de ausência.
5. Registrar todas as decisões no log de tratamento.

Tratamento de dados faltantes não é apenas limpeza — é uma decisão inferencial que pode alterar os resultados do modelo.

## 17.4 Detecção de valores extremos e inconsistências

Valores extremos (outliers) podem ser resultado de erro de digitação, inconsistência de unidade ou representar fenômenos reais raros. Em regressão, esses pontos podem influenciar de forma desproporcional os estimadores $\widehat{\beta}$, os resíduos e os diagnósticos de ajuste.

Nem todo valor extremo é um erro, mas todo valor extremo exige investigação.

**Erros clássicos**

- Remover automaticamente qualquer ponto extremo sem verificar sua origem.
- Estabelecer cortes arbitrários sem registrar critérios e justificativas.
- Transformar a variável resposta sem considerar a nova interpretação dos coeficientes.
- Ignorar a possibilidade de que o ponto seja informativo para o fenômeno estudado.

Segundo Charnet et al. (2008) e Sheather (2009), valores extremos podem distorcer estimativas, ampliar variâncias e comprometer testes de hipóteses.

Barnett e Lewis (1994) oferecem fundamentos formais para detecção de outliers em análises univariadas e multivariadas, distinguindo entre:

- Observações extremas na distribuição marginal;
- Pontos de alta alavancagem (leverage);
- Observações influentes (que alteram substancialmente o ajuste do modelo).

Embora diagnósticos formais de influência sejam discutidos em outros capítulos, a identificação preliminar já deve ocorrer nesta etapa.

**Ferramentas R**

Para inspeção inicial:

- `quantile()` — limites interquartílicos
- `sd()` e `scale()` — padronização e z-score
- `summary()` — inspeção geral

Visualizações:

- `ggplot2::geom_boxplot()`
- `ggplot2::geom_histogram()`
- `ggplot2::geom_point()`

A visualização frequentemente revela padrões que estatísticas isoladas não mostram.

**Como resolver ou minimizar efeitos**

- Confirmar se o valor resulta de erro de digitação ou unidade (ex.: centímetros vs metros).
- Corrigir inconsistências quando verificadas documentalmente.
- Aplicar winsorização (`DescTools::Winsorize`) quando a estratégia for limitar extremos mantendo observações.
- Utilizar transformações (log, raiz quadrada) quando houver forte assimetria.
- Manter a observação quando for substantivamente plausível e documentar a decisão.

A remoção deve ser exceção, não regra.

**Observação importante**

Eliminar valores extremos altera a distribuição da variável, o tamanho da amostra e potencialmente a matriz $\mathbf{X}$. Toda decisão deve ser registrada no log de tratamento.

## 17.5 Padronização e transformações numéricas

Padronizar variáveis numéricas torna os preditores comparáveis em escala e pode melhorar a estabilidade numérica de procedimentos de estimação. Transformações adequadas, por sua vez, reduzem assimetria, estabilizam variância e facilitam interpretações coerentes com o fenômeno estudado.

Em regressão linear clássica, a padronização não altera o ajuste global do modelo nem o $R^2$, mas modifica a escala dos coeficientes e sua interpretação. Já em métodos penalizados (Ridge e LASSO), a padronização é praticamente indispensável.

Charnet et al. (2008) ressaltam que variáveis em escalas muito distintas podem gerar instabilidade numérica e dificultar a comparação entre efeitos.

Sheather (2009) destaca que reescalar variáveis pode facilitar a interpretação de coeficientes em regressões múltiplas, especialmente quando as unidades originais são muito grandes ou muito pequenas.

Montgomery, Peck, e Vining (2021) enfatizam que diferenças extremas de magnitude entre covariáveis podem afetar diagnósticos e procedimentos computacionais.

**Ferramentas R**

- `scale()` — padronização pelo z-score (média zero e desvio padrão um).
- Reescala min–max — pode ser feita manualmente via transformações aritméticas.
- `MASS::boxcox()` — identificação de transformações do tipo Box–Cox.
- `log()` ou `log1p()` — transformações logarítmicas (úteis para assimetria positiva).
- Transformações via `dplyr::mutate()` para aplicação sistemática.

A escolha da transformação deve ser guiada pelo comportamento empírico da variável e pelo contexto substantivo.

**Soluções recomendadas na literatura**

- Para variáveis altamente assimétricas, aplicar transformações logarítmicas pode aproximar a normalidade e reduzir heterocedasticidade.
- Para contagens moderadas, Paula (2004) sugere transformação por raiz quadrada como alternativa simples.
- Para distribuições fortemente assimétricas, considerar Box–Cox quando apropriado.
- Evitar misturar variáveis em escalas radicalmente diferentes sem padronização prévia.

**Quando padronizar**

- Para comparar magnitudes relativas entre preditores.
- Para métodos penalizados (Ridge, LASSO), em que a penalização depende da escala.
- Para algoritmos baseados em distância ou otimização numérica.
- Quando unidades originais dificultam interpretação direta.

Padronizar não é obrigação universal; é uma decisão metodológica que deve preservar a interpretabilidade do modelo.

## 17.6 Codificação de variáveis categóricas (dummies)

Uma variável categórica com $k$ níveis não pode entrar diretamente como coluna numérica em $\mathbf{X}$. É necessário convertê-la em variáveis indicadoras (dummies), usualmente em número $k-1$, para evitar colinearidade perfeita.

A escolha da codificação determina a interpretação dos coeficientes estimados.

**Erros comuns**

- Criar $k$ dummies para $k$ categorias (armadilha da variável dummy), gerando singularidade em $\mathbf{X}^\top\mathbf{X}$.
- Escolher categoria de referência sem critério substantivo.
- Manter níveis raros, produzindo colunas quase vazias e estimativas instáveis.
- Usar rótulos inconsistentes (acentos, abreviações, maiúsculas/minúsculas misturadas).
- Tratar variável ordinal como nominal sem refletir sobre a estrutura de ordem.

**Conexão com a modelagem**

Quando uma variável categórica é convertida corretamente em $k-1$ dummies, cada coeficiente estimado representa a diferença média entre aquela categoria e a categoria de referência, mantendo os demais preditores constantes.

Se todas as $k$ dummies forem incluídas juntamente com o intercepto, ocorre dependência linear exata, tornando $\mathbf{X}^\top\mathbf{X}$ não invertível no caso clássico de MRLM.

Portanto, a codificação correta não é apenas conveniência computacional, é condição para a existência do estimador de mínimos quadrados.

**Fundamentação teórica**

Charnet et al. (2008) destacam a importância da escolha da categoria de referência para interpretação dos coeficientes.

James et al. (2013) discutem como alta cardinalidade pode gerar modelos instáveis e sobreajustados.

Para variáveis ordinais, a estrutura de ordenação pode ser incorporada explicitamente, evitando perda de informação.

**Ferramentas R**

- `model.matrix()` — gera automaticamente a matriz de projeto com codificação apropriada.
- `fastDummies` — criação explícita de variáveis indicadoras.
- `recipes::step_dummy()` — codificação sistemática em pipelines.
- `forcats` — manipulação e reorganização de níveis.
- Fatores ordenados (`ordered`) para variáveis com hierarquia natural.

O R, por padrão, utiliza codificação por tratamento (*treatment contrasts*), mas outras codificações podem ser especificadas conforme necessidade.

**Boas práticas**

- Definir categoria de referência com base em critério substantivo (grupo controle, baseline, padrão).
- Agrupar níveis raros quando apropriado.
- Manter um dicionário de variáveis documentando níveis e significados.
- Padronizar rótulos antes da geração de dummies.
- Verificar o número final de parâmetros gerados após codificação.

**Atenção à alta cardinalidade**

Variáveis com muitos níveis distintos podem gerar dezenas ou centenas de colunas em $\mathbf{X}$, aumentando dimensionalidade e risco de sobreajuste.

Nesses casos, considere:

- Agrupamento por regras de negócio;
- Seleção de variáveis;
- Métodos penalizados (quando apropriado na etapa seguinte).

Codificar corretamente é garantir que a estrutura qualitativa do fenômeno seja traduzida adequadamente para o modelo quantitativo.

## 17.7 Verificação de condições para modelagem linear

Cumprir condições mínimas como variabilidade das covariáveis, ausência de colinearidade perfeita e tamanho amostral adequado ($n > p+1$) é o que permite aplicar os resultados teóricos da regressão linear e múltipla com segurança.

No modelo linear clássico, a existência do estimador de mínimos quadrados depende de $\mathbf{X}^\top\mathbf{X}$ ser invertível, o que exige que a matriz de projeto $\mathbf{X}$ tenha posto completo. Assim, esta verificação não é opcional — é estrutural. (Searle (2016))

Essas checagens antecedem os diagnósticos formais e reduzem problemas posteriores na estimação e interpretação.

**Ferramentas R úteis**

- `cor()` ou pacote `corrr` — inspeção de associações entre preditores.
- `qr()` ou `Matrix::rankMatrix()` — verificação do posto da matriz de projeto.
- `car::vif()` — cálculo do fator de inflação da variância (VIF). (Fox e Weisberg (2024))
- `summary()` e inspeção de variância — identificação de variáveis constantes.

Essas ferramentas ajudam a identificar:

- Colinearidade perfeita ou quase perfeita;
- Variáveis constantes ou quase constantes;
- Relações lineares redundantes entre preditores.

**Como mitigar problemas**

- Remover ou combinar variáveis altamente correlacionadas.
- Revisar a codificação de dummies para evitar dependência linear.
- Padronizar variáveis quando necessário.
- Eliminar duplicatas estruturais na base.
- Reduzir dimensionalidade quando $p$ se aproxima de $n$.

Toda verificação deve ser documentada no relatório de tratamento.

**Compatibilidade da variável resposta**

Antes da modelagem, é essencial verificar se o tipo da variável resposta é compatível com o modelo pretendido:

- **Contínua**: variável em escala intervalar ou razão.
- **Binária**: 0/1 ou fator com dois níveis.
- **Contagem**: inteiros não negativos.
- **Proporção**: valores no intervalo $[0, 1]$.

Escolher modelo inadequado ao tipo de resposta gera inferências inválidas.

**Tabela-guia: tipo de resposta e cuidados**

| Tipo de resposta | Exemplo | Tratamento pré-modelo | Modelo típico |
|---|---|---|---|
| Contínua | Preço, temperatura | Verificar outliers, padronização e unidades | MRLS, MRLM |
| Binária | Sucesso/fracasso | Conferir codificação 0/1, balanceamento e faltantes | Logístico, Probit |
| Contagem | Nº de eventos | Avaliar zeros estruturais e dispersão | Poisson, Binomial Negativa |
| Proporção | Taxa, share | Verificar limites 0/1 e denominadores | Beta, Quasi-binomial |

Modelar sem verificar essas condições equivale a aplicar teoria sob premissas não verificadas. O tratamento adequado garante que a transição para a modelagem seja matemática e estatisticamente legítima.

## 17.8 Sumários e visualizações exploratórias

Explorar os dados antes da modelagem permite identificar padrões, inconsistências e relações estruturais que orientam decisões de limpeza, transformação e especificação do modelo.

A visualização não substitui a modelagem, ela antecipa problemas e revela estruturas que podem afetar a construção de $\mathbf{X}$ e a escolha do modelo. (Tufte (2001))

**Ferramentas R**

Funções e pacotes úteis nesta etapa:

- `summary()` — estatísticas descritivas básicas.
- `table()` ou `dplyr::count()` — frequências de variáveis categóricas.
- `ggplot2` — histogramas, boxplots, gráficos de dispersão.
- `GGally::ggpairs()` — matriz gráfica de dispersão.
- `corrplot` ou `ggcorrplot` — visualização de matrizes de correlação.
- `plotly` — visualizações interativas (opcional).

O objetivo é compreender estrutura, dispersão, assimetria e possíveis relações lineares preliminares.

**Como resolver dificuldades comuns**

- Distribuições muito assimétricas → aplicar transformações (log, raiz quadrada) e reavaliar.
- Categorias vazias ou raras → reclassificar níveis ou agrupar.
- Escalas muito distintas → padronizar antes de comparar magnitudes.
- Correlações elevadas entre preditores → revisar especificação do modelo.

Visualizar é diagnosticar antes do diagnóstico formal.

**Produtos esperados**

Ao final da etapa exploratória, espera-se:

- Tabela de estatísticas descritivas por variável numérica (média, desvio padrão, p5, p50, p95).
- Frequência absoluta e relativa por variável categórica.
- Histogramas e boxplots para avaliar distribuição.
- Gráficos de dispersão entre $Y$ e cada $X$.
- Matriz de correlação entre preditores numéricos.

Esses produtos funcionam como evidência documental do entendimento da base antes da modelagem.

**Perguntas-guia**

- Alguma variável apresenta assimetria extrema?
- Existem valores fora de faixa plausível?
- Há categorias quase vazias?
- Quais pares de $X$ apresentam correlação elevada?
- A relação entre $Y$ e $X$ parece aproximadamente linear?

Responder a essas perguntas reduz erros na etapa seguinte.

**Mapa de funções em R (resumo operacional)**

- **Leitura**: `read_csv`, `read_delim`, `read_excel`.
- **Inspeção**: `glimpse`, `summary`, `count`.
- **Transformação**: `mutate`, `across`, `scale`.
- **Tipagem**: `as.numeric`, `as.Date`, `factor`.
- **Correlação**: `cor`.
- **Matriz de projeto**: `model.matrix`.
- **Diagnóstico estrutural**: `qr`, `rankMatrix`, `vif`.

A exploração sistemática consolida a transição entre tratamento de dados e modelagem estatística.

## 17.9 Salvamento e documentação da base tratada

A documentação do tratamento (log + dicionário) e o versionamento adequado garantem reprodutibilidade, transparência metodológica e facilitam revisão por pares.

Em regressão, a qualidade das inferências depende não apenas do modelo ajustado, mas da rastreabilidade das decisões tomadas antes da modelagem.

**O que entregar ao final do tratamento**

1. Base final tratada (formato padrão, por exemplo: CSV).
2. Dicionário de variáveis contendo:
    - Nome da variável
    - Tipo
    - Unidade (quando aplicável)
    - Níveis (para categóricas)
    - Origem ou transformação realizada
3. Log de decisões documentando:

- O que foi modificado
- Por que foi modificado
- Quando foi modificado

4. Script reproduzível contendo todas as etapas do tratamento (opcional, mas altamente recomendado).

O objetivo é que qualquer outro pesquisador consiga reconstruir exatamente a base utilizada na modelagem.

**Convenções úteis**

- Utilizar nome de arquivo com carimbo de data, por exemplo: `base_tratada_2025-10-26.csv`
- Manter script de preparo versionado e comentado.
- Utilizar estrutura organizada de pastas:
  - `dados_brutos/`
  - `dados_interinos/`
  - `dados_tratados/`

Nunca sobrescrever dados brutos.

**Pipelines modernos e reprodutibilidade**

Em aplicações contemporâneas, o tratamento de dados é frequentemente organizado em **pipelines**: sequências estruturadas de leitura → transformação → validação → saída tratada → modelagem.

Esse fluxo reduz erros humanos, aumenta consistência e fortalece a reprodutibilidade científica. No ecossistema R, essa abordagem é facilitada por ferramentas como `tidyverse`, `recipes` e estruturas de modelagem integradas.

Mesmo quando não formalizado em código automatizado, o processo deve ser concebido como um fluxo explícito, sequencial e documentado.

Modelar é o passo visível; documentar é o passo que garante credibilidade.

## 17.10 Checklist técnico: Condições mínimas para seguir à modelagem

Este checklist consolida os critérios estruturais que devem ser verificados ao final do tratamento de dados, antes do ajuste de qualquer modelo de regressão.

1. **Variabilidade das covariáveis:** nenhuma variável explicativa deve ser constante ou quase constante.

2. **Ausência de colinearidade perfeita:** a matriz $\mathbf{X}$ deve ter posto completo; logo, $\mathbf{X}^\top\mathbf{X}$ deve ser não singular.
3. **Tamanho amostral adequado:** $n > p + 1$ no caso do modelo linear com intercepto; caso contrário, considere reduzir dimensionalidade ou ampliar a amostra.
4. **Escalas compatíveis:** quando necessário, variáveis reescaladas ou padronizadas para evitar instabilidade numérica.
5. **Tipos de dados conferidos:** variáveis numéricas, categóricas e temporais corretamente tipadas.
6. **Ausência de erros estruturais:** duplicatas removidas, chaves validadas e inconsistências corrigidas.
7. **Base salva e documentada:** versão final armazenada com data, autor e dicionário de variáveis.
8. **Pronta para modelagem:** a base pode ser utilizada diretamente em MRLS, MRLM, MLG ou métodos penalizados sem retrabalho estrutural.

O não atendimento a qualquer desses critérios compromete a legitimidade estatística do modelo. A qualidade de toda inferência estatística depende dessa distinção, e ela começa antes do ajuste do primeiro modelo.

## 17.11 Bases para prática

Ao finalizar o tratamento de uma base, tente **ajustar um modelo simples apenas para validar tipos e dummies** (sem discutir resultados). Se rodar sem erros e os sumários fizerem sentido, a base está pronta para a próxima unidade.

**Escada de dificuldade das bases - Nível 1 (aquecimento):** Online Shoppers: tipagem + dummies + sumários.
- **Nível 2 (intermediário):** Ames/House Prices: faltantes moderados + reescala + dicionário.
- **Nível 3 (avançado):** Student Failure (messy): chaves/duplicatas + integração + plano de imputação.
- **Extra (discreta):** Bike Sharing: temporais + zeros estruturais + outliers climáticos.

Esta seção apresenta **bases públicas e didáticas** para exercícios de tratamento pré-modelagem. Cada item traz link de acesso, o que a base representa e **situações-problema** que motivam o tratamento. Ao final de cada base há um bloco **Tarefas sugeridas** para orientar o estudo.

### 17.11.1 House Prices: Advanced Regression Techniques (Kaggle)

- **Tipo de resposta:** Contínua (`SalePrice`).
- **Link:** https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

- **Sobre a base:** preços de casas em Ames (Iowa, EUA), com ~79 variáveis numéricas e categóricas.
- **Situação:**
  - Muitas colunas com valores ausentes (por ex.: `LotFrontage`, `Alley`, `PoolQC`).
  - Variáveis categóricas com codificação inconsistente e níveis raros.
  - Outliers em preço e área; unidades e escalas heterogêneas.
- Excelente caso para um **pipeline completo** de limpeza (missing, tipagem, codificação, reescala) antes da regressão contínua.
- **Tarefas sugeridas:**
  1. Mapear porcentagens de faltantes por coluna e decidir estratégia (excluir, imputar, manter).
  2. Padronizar nomes e unidades; verificar outliers em `SalePrice` e `GrLivArea`.
  3. Unificar níveis categóricos raros e definir dummies com categoria de referência.
  4. Salvar uma versão tratada e documentar as decisões.

### 17.11.2 Ames Housing Dataset

- **Tipo de resposta:** Contínua (`SalePrice`).
- **Link:** https://github.com/data-doctors/kaggle-house-prices-advanced-regression-techniques
- **Sobre a base:** variação/derivação do problema de habitação de Ames, amplamente usada em cursos.
- **Situação:**
  - Mistura de tipos (numéricos + categóricos), com valores ausentes e níveis raros.
  - Recomendação frequente de transformação logarítmica da resposta.
  - Outliers estruturais (ex.: casas muito acima da média).
- Reforça o **contraste de estratégias** de tratamento em relação à 10.1.
- **Tarefas sugeridas:**
  1. Comparar duas abordagens de tratamento de faltantes (ex.: imputação por mediana vs. KNN) e registrar impactos em sumários.
  2. Testar padronização vs. não padronização nas variáveis contínuas.
  3. Produzir um dicionário de variáveis claro.

### 17.11.3 Online Shoppers Purchasing Intention (UCI)

- **Tipo de resposta:** Binária (`Revenue`: sim/não).
- **Link:** https://archive.ics.uci.edu/ml/datasets/Online%2BShoppers%2BPurchasing%2BIntention%2BDataset

- **Sobre a base:** sessões de navegação em e-commerce; objetivo é prever se a sessão termina em compra.
- **Situação:**
  - Variáveis categóricas e temporais misturadas às numéricas; necessidade de codificação.
  - Possível desbalanceamento da classe `Revenue`.
  - Datas/temporais como texto exigindo normalização e extração de componentes.
- Bom **caso de tratamento moderado**, contrastando com bases mais "sujas".
- **Tarefas sugeridas:**
  1. Verificar distribuição de `Revenue` (balanceamento).
  2. Definir dummies consistentes e padronizar escalas numéricas.
  3. Criar variáveis derivadas temporais (mês, dia da semana) de forma reproducível.

### 17.11.4 Student Failure (Messy) Dataset (Kaggle)

- **Tipo de resposta:** Binária (`fail` = 1 se o aluno reprova/sai; 0 caso contrário).
- **Link:** https://www.kaggle.com/code/sashatarakanova/student-failure-modelling-with-a-messy-dataset
- **Sobre a base:** dados educacionais com múltiplas tabelas heterogêneas para prever reprovação.
- **Situação:**
  - Muitos valores ausentes; tabelas com chaves não padronizadas; duplicatas.
  - Categorias inconsistentes para o mesmo conceito (ex.: formas distintas de escrever "curso").
  - Necessidade de unificação/integração de fontes (join/merge) com validação.
- Ótimo para treinar **integração e saneamento** antes de qualquer modelagem binária.
- **Tarefas sugeridas:**
  1. Reconstruir uma chave única estável e eliminar duplicatas.
  2. Mapear e recodificar categorias equivalentes.
  3. Documentar um plano de imputação apropriado por variável.

### 17.11.5 Bike Sharing Dataset (UCI)

- **Tipo de resposta:** Discreta (contagem de bicicletas alugadas por hora/dia).
- **Link:** https://archive.ics.uci.edu/dataset/275/bike%2Bsharing%2Bdataset
- **Sobre a base:** uso de bicicletas compartilhadas, com variáveis meteorológicas, feriados, sazonalidade e efeitos de hora do dia.
- **Situação:**

  - Contagens com muitos zeros em horários de baixa demanda e picos em horários de pico; necessidade de identificar **zeros estruturais**.
  - Variáveis temporais em formato de texto exigindo conversão e extração (hora, dia da semana, feriado).
  - Possíveis outliers (eventos climáticos extremos) e variabilidade alta da resposta.
  - Padronização de escalas e codificação consistente de feriados/sazonalidade.
  - Prepara para o tratamento de **resposta discreta (contagem)**, anterior à escolha de modelos como Poisson ou Binomial Negativa.

- **Tarefas sugeridas:**

  1. Normalizar as variáveis temporais e criar indicadores (feriado, fim de semana, hora do rush).
  2. Caracterizar zeros estruturais vs. esparsidade aleatória e discutir implicações para a modelagem.
  3. Detectar outliers climáticos e decidir estratégia (transformação, winsorização ou justificativa de manutenção).
  4. Entregar uma versão tratada com dicionário e log de decisões.

## 17.12 Glossário

**Encoding (codificação de caracteres)**
Como o computador guarda letras com acentos e símbolos. Exemplos: **UTF-8**, **latin-1**. Se o encoding está errado, aparecem “ ” ou letras quebradas.

**Delimitador / Separador**
Símbolo que separa colunas em arquivos de texto. Exemplos: vírgula `,`, ponto e vírgula `;`, tab .

**Decimal (símbolo decimal)**
O símbolo que separa parte inteira e fracionária. Em pt-BR, vírgula (ex.: `3,14`); em en-US, ponto (`3.14`).

**CSV, XLSX, SAV, DTA**
Formatos de planilha/tabela. **CSV**: texto simples; **XLSX**: Excel; **SAV**: SPSS; **DTA**: Stata.

**Licença (de uso dos dados)**
Condições legais de uso/compartilhamento do dataset (ex.: CC-BY). Leia antes de usar.

**Faltante / Dados faltantes**
Informação ausente em uma célula. Pode aparecer como **NA**, **NaN**, vazio `""` ou códigos como `999`.

**MCAR, MAR, MNAR**
Tipos de mecanismo de ausência: **MCAR** (ausência completamente ao acaso), **MAR** (ao acaso condicional a outras variáveis), **MNAR** (não ao acaso; depende do próprio valor ausente).

**Duplicatas e chaves**
**Duplicata**: linha repetida. **Chave**: coluna (ou combinação) que identifica exclusivamente cada linha (ex.: `id`).

**Log de duplicatas / Log de tratamento**
Registro simples do que foi removido/alterado e por quê. Ajuda na reprodutibilidade.

**Outlier**
Valor muito fora do padrão do conjunto. Pode ser erro, evento raro ou caso especial.

**Winsorização (winsorize)**
Técnica que **limita** valores extremos a um limite (ex.: truncar no percentil 1% e 99%) para reduzir impacto de outliers.

**Reescala / Padronização (z-score)**
Colocar variáveis em escala comparável. **z-score**: subtrai a média e divide pelo desvio-padrão (fica média 0 e dp 1). **min–max**: leva para [0,1] pela fórmula `(x-min)/(max-min)`.

**Cardinalidade (de categorias)**
Número de níveis distintos de uma variável categórica. Alta cardinalidade = muitos níveis.

**Padronizar rótulos**
Escrever categorias de forma consistente (ex.: tudo minúsculo, sem acento, sem espaços extras), unificando sinônimos.

**Dummies (one-hot)**
Transformar uma variável categórica em colunas 0/1 (uma coluna a menos que o número de categorias, para evitar colinearidade perfeita).

**Variável ordinal**
Categorias que **têm ordem** (ex.: fundamental < médio < superior).

**Zeros estruturais**
Zeros esperados por construção (ex.: aluguel de bicicletas à 03h pode ser zero). Diferem de "zeros por acaso".

**Regularização (Ridge, LASSO)**
Técnicas que penalizam coeficientes para lidar com muitas variáveis e reduzir sobreajuste; exigem atenção à **escala** dos preditores.

**JSON**
Formato de texto para dados estruturados em pares chave:valor (não será foco aqui; usamos tabelas).

**Pipeline**
Sequência organizada de passos de tratamento: leitura → limpeza → transformação → verificação → saída tratada.

# Referências


Abadir, Karim M., e Jan R. Magnus. 2005. *Matrix Algebra*. Cambridge: Cambridge University Press.

Akaike, Hirotugu. 1974. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* 19 (6): 716–23.

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. 3º ed. New York: Wiley.

Barnett, Vic, e Toby Lewis. 1994. *Outliers in Statistical Data*. 3º ed. Chichester: Wiley.

Belsley, David A., Edwin Kuh, e Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Burnham, Kenneth P., e David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2º ed. New York: Springer.

Buuren, Stef van, e Karin Groothuis-Oudshoorn. 2024. *mice: Multivariate Imputation by Chained Equations*. CRAN. https://cran.r-project.org/package=mice.

Casella, George, e Roger L. Berger. 2002. *Statistical Inference*. 2º ed. Pacific Grove: Duxbury.

Charnet, Reinaldo, Carlos Alberto Freire, Eliane M. R. Charnet, e Helio Bonvino. 2008. *Análise de Modelos de Regressão Linear com Aplicações*. 2º ed. Campinas: EDUNICAMP.

Draper, Norman R., e Harry Smith. 1998. *Applied Regression Analysis*. 3º ed. New York: John Wiley & Sons.

Fox, John, e Sanford Weisberg. 2024. *vif: Variance Inflation Factors*. car package documentation. https://search.r-project.org/CRAN/refmans/car/html/vif.html.

Galton, Francis. 1886a. "Family likeness in stature". *Proceedings of the Royal Society of London* 40: 42–72.

———. 1886b. "Regression towards mediocrity in hereditary stature". *Journal of the Anthropological Institute* 15: 246–63.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, e Donald B. Rubin. 2014. *Bayesian Data Analysis*. 3º ed. CRC Press.

Giordano, Frank R., William P. Fox, e Steven B. Horton. 2013. *A First Course in Mathematical Modeling*. Cengage Learning.

Golub, Gene H., e Charles F. Van Loan. 2013. *Matrix Computations*. 4º ed. Baltimore: Johns Hopkins University Press.

Gujarati, Damodar N. 2006. *Econometria Básica*. 4º ed. Rio de Janeiro: Elsevier Campus.

Harville, David A. 1997. *Matrix Algebra From a Statistician's Perspective*. New York: Springer.

———. 2000. *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Hoffmann, Rodolfo. 2006. *Análise de Regressão: Uma Introdução à Econometria*. 2º ed. São Paulo: Hucitec.

———. 2016. *Análise de Regressão: Uma Introdução à Econometria*. 5º ed. Portal de Livros Abertos da USP. https://doi.org/10.11606/9788592105709.

Iannone, Richard. 2024. *pointblank: Data Validation*. R package documentation. https://rstudio.github.io/pointblank/.

James, Gareth, Daniela Witten, Trevor Hastie, e Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Kuhn, Max et al. 2024. *step_impute_mean: Imputation Steps*. recipes package documentation. https://recipes.tidymodels.org/reference/step_impute_mean.html.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, e William Li. 2005. *Applied Linear Statistical Models*. 5º ed. New York: McGraw-Hill.

Little, Roderick J. A., e Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. 3º ed. Hoboken: Wiley.

Meerschaert, Mark M. 2013. *Mathematical Modeling*. 4º ed. Academic Press.

Montgomery, Douglas C., Elizabeth A. Peck, e G. Geoffrey Vining. 2021. *Introduction to Linear Regression Analysis*. 6º ed. Hoboken: John Wiley & Sons.

Paula, Gilberto A. 2004. *Modelos de Regressão com Apoio Computacional*. São Paulo: IME-USP.

Pearson, Karl, e Alice Lee. 1903. "On the laws of inheritance". *Biometrika* 2: 357–462.

Rencher, Alvin C., e William F. Christensen. 2012. *Methods of Multivariate Analysis*. 3º ed. Hoboken: Wiley.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model". *The Annals of Statistics* 6 (2): 461–64.

Searle, Shayle R. 2016. *Matrix Algebra Useful for Statistics*. 2º ed. Hoboken: Wiley.

Sheather, Simon J. 2009. *A Modern Approach to Regression with R*. New York: Springer.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2º ed. Cheshire: Graphics Press.

Weisberg, Sanford. 2005. *Applied Linear Regression*. New York: Wiley.

Wickham, Hadley et al. 2024. *Missing values*. tidyr package documentation. https://tidyr.tidyverse.org/articles/missing-values.html.