Aula 6 – Breve recapitulação de Modelos Lineares Gerais

Perguntas comuns

- Um novo fertilizante realmente aumenta o crescimento médio das plantas em comparação com o fertilizante antigo?

- A concentração de um poluente em um rio está, em média, acima do limite de segurança estabelecido por lei?

- Um medicamento para baixar a pressão arterial realmente funciona? Como podemos provar isso comparando a pressão *antes* e *depois* do tratamento?

O que é um teste *t*

Ferramenta estatística que permite ***comparar médias*** *de* ***uma*** *ou* ***duas populações***. *Ele avalia* ***se as médias são*** *significativamente* ***diferentes****, levando em conta a média, o desvio padrão, e o tamanho da amostra*.

Teste t como um tipo particular de modelo linear geral

Introduzindo LMs

- Até aqui aprendemos três versões do Teste t de Student: para uma amostra, para duas amostras independentes e para dados pareados. Cada um parecia ter sua própria fórmula e seu próprio procedimento.
- Isso pode parecer uma 'coleção de testes', uma receita de bolo para cada situação.
- Mas e se eu dissesse que todos eles são, na verdade, **a mesma ideia estatística fundamental** disfarçada de maneiras diferentes?

Médias amostrais são estimadores não enviesados de médias populacionais

Posso comparar médias de duas populações

Existem vários tipos de teste t

Diferentes teste t são na verdade tipos particulares de um único modelo linear

Introduzindo LMs

- Agora, vamos revisitar o Teste t, mas sob uma nova ótica: a de **modelagem linear**.
- Vamos ver como todos os testes t são, na essência, modelos lineares simples.
- Entender isso não só unifica o conhecimento, mas abre as portas para quase todas as técnicas estatísticas que vocês usarão na carreira, como a ANOVA.

Equação de um modelo linear passo-a-passo

$$Dados = Modelo + Erro$$

Equação de um modelo linear passo-a-passo

$$\widehat{Dados}_i = Modelo_i$$

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Variável resposta

Variável preditora

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

intercepto

Inclinação (slope)

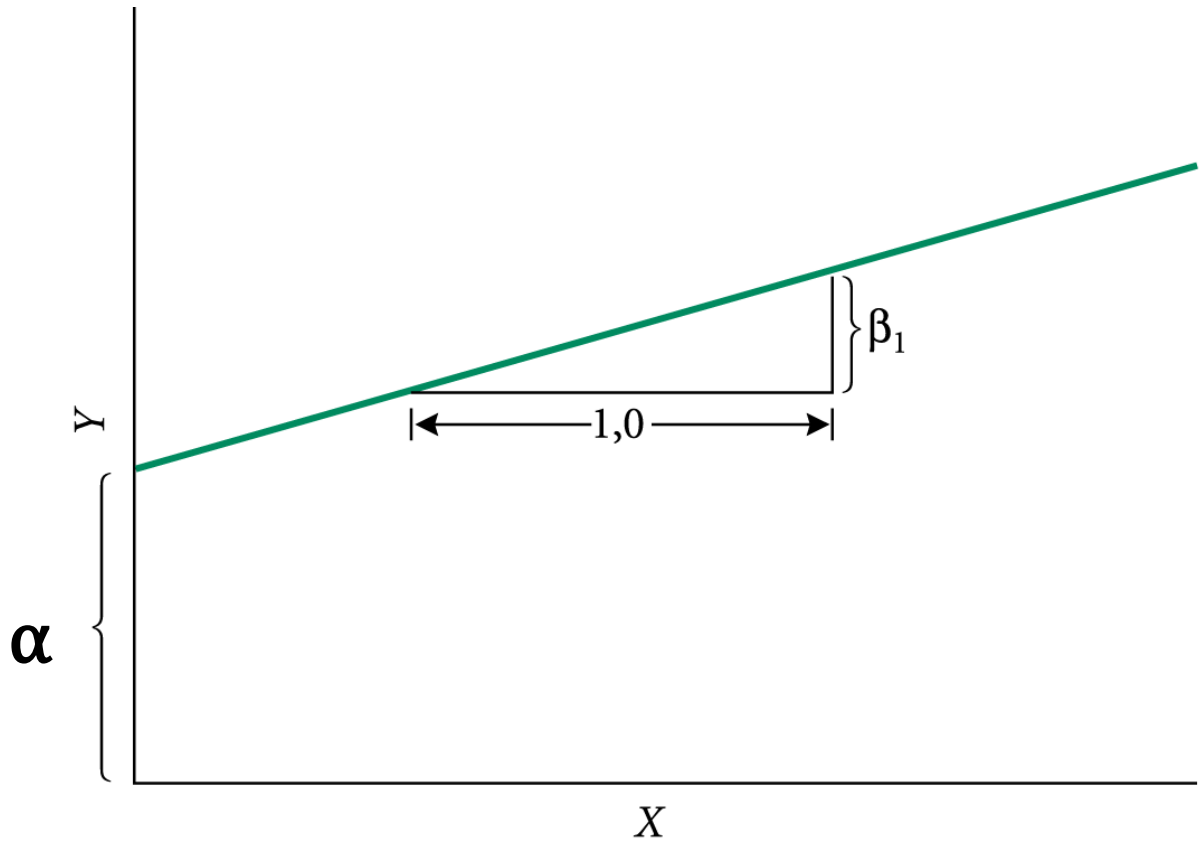


Figura 9.1 Relação linear entre as variáveis *X* e *Y*. A linha é descrita pela equação $Y = \beta_0 + \beta_1 X$, onde β_0 é o intercepto e β_1 é a inclinação da linha. O intercepto β_0 é o valor predito da equação quando $X = 0$. A inclinação da linha β_1 é o aumento na variável *Y* associado com o de uma unidade da variável *X* ($\Delta Y/\Delta X$). Se o valor de *X* é conhecido, o valor predito de *Y* pode ser calculado multiplicando *X* pela inclinação e somando o intercepto (β_0).

Equação de um modelo linear passo-a-passo

$$Erro_i = Dados_i - \widehat{Dados}_i$$

Exemplo

Altura dos filhos = $\alpha + \beta$*altura dos pais + resíduos

Pressão arterial = $\alpha + \beta$*grupo de tratamento + resíduos

Parâmetros do modelo

- Intercepto = valor médio da variável resposta quando a variável preditora é zero

- Inclinação (slope, coeficiente angular) = inclinação da reta predita pelo modelo. Representa o efeito de X sobre Y. Mede o quanto a variável resposta (Y) muda em unidades de β.
  - Pode-se calcular o IC95% para o slope
- Resíduo (erro): Representa a variabilidade biológica natural e tudo mais que não conseguimos explicar com nosso modelo.


MakeAGIF.com

Preditores categóricos

- Até agora, parece uma equação de reta da 8ª série. Mas o grande poder do modelo linear é que a variável preditora, X, **não precisa ser numérica**
- Podemos usar uma variável categórica, como ‘grupo tratado’. E a maneira de fazer isso é com algo chamado **variáveis Dummy**
  - Se temos dois grupos (ex: Controle e Tratamento), criamos uma variável X que assume o valor **0** para um grupo (o grupo de referência, ex: Controle) e **1** para o outro (ex: Tratamento).

Variáveis *dummy* como “tradutores”

- Mas e se o nosso preditor não for um número e sim uma categoria? Por exemplo:
  - **Grupo:** Tratamento vs. Controle
  - **Sexo:** Macho vs. Fêmea
  - **Local:** Área Preservada vs. Área Poluída
- Não podemos colocar a palavra “Tratamento” na equação e multiplicá-la por β. Precisamos de uma forma de “traduzir” essas categorias para uma linguagem numérica que o modelo linear entenda.
- A variável dummy é um “tradutor” que converte categorias em números (0 e 1), permitindo que usemos preditores categóricos em modelos matemáticos como o modelo linear.

Solução para o problema

- A **variável *dummy*** (também chamada de variável fictícia, indicadora ou binária) é a solução para este problema. Ela é uma variável numérica que representa uma informação categórica.
- **Como funciona?** A forma mais comum é a **codificação 0/1**:
- **Escolha um “Grupo de Referência”:** Um dos seus grupos será a base de comparação. A escolha é arbitrária, mas afeta a interpretação dos resultados. Geralmente, o grupo "Controle" ou o grupo mais comum é escolhido como referência.
- **Atribua os Códigos:**
  - O grupo de referência recebe o código **0**.
  - O outro grupo recebe o código **1**.
- Essa simples codificação transforma nossa variável categórica em um número que pode ser usado na equação do modelo linear. Ela atua como um “interruptor”: está “desligada” (0) para o grupo de referência e “ligada” (1) para o outro grupo.

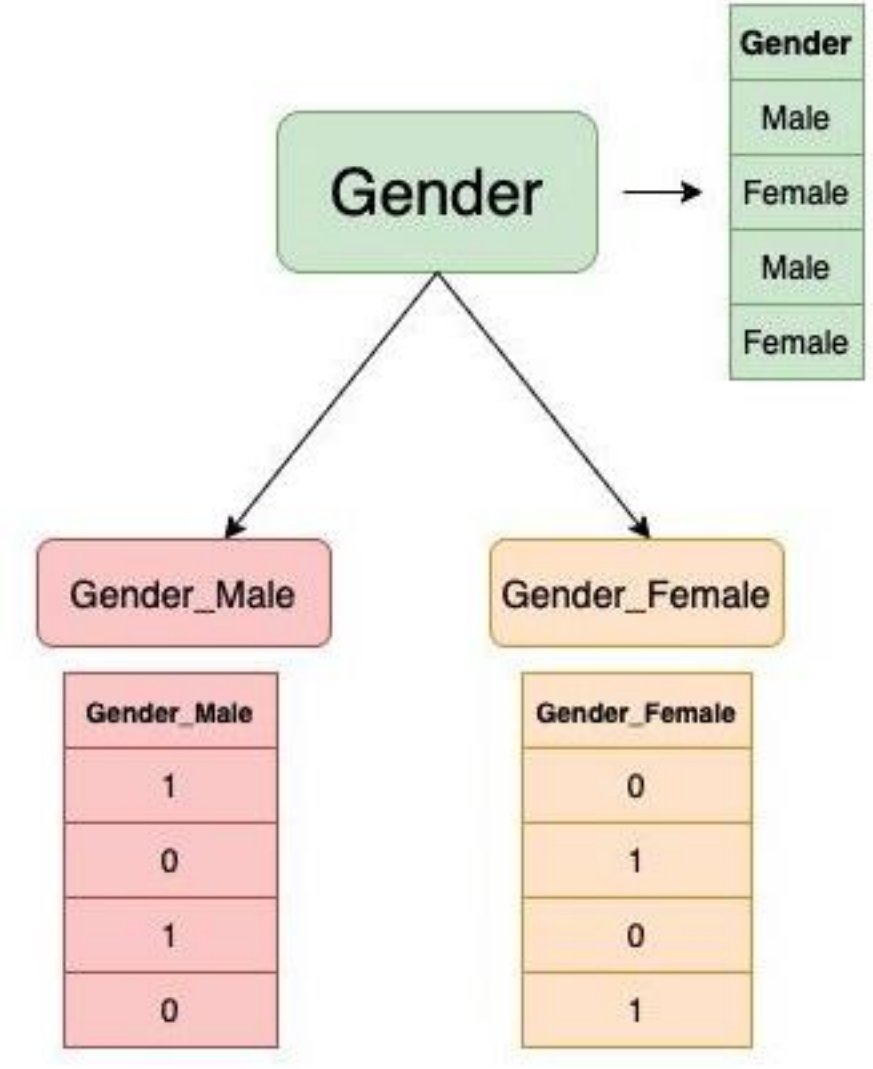
Codificando um fator como *dummy*

- Vamos usar um cenário biológico: um experimento para testar o efeito de um suplemento na dieta de ratos.
- **Variável Categórica Original:** Grupo (com os níveis "Controle" e "Suplemento")
- **Variável Resposta (Y):** Peso_Final (em gramas)

Rato	Grupo	Peso_Final (g)
1	Controle	205
2	Controle	215
3	Suplemento	230
4	Suplemento	240
5	Controle	210

Rato	Grupo	Peso_Final (g)	Suplemento_dummy
1	Controle	205	0
2	Controle	215	0
3	Suplemento	230	1
4	Suplemento	240	1
5	Controle	210	0

Dummy para um fator com 2 níveis



A mágica na interpretação do modelo

- Agora, vamos ver o que acontece quando colocamos essa variável dummy na nossa equação:

$$Peso\_Final_i = \alpha + \beta * (Suplemento\_dummy_i)$$

- **Para um rato do grupo Controle (Suplemento_dummy=0):** A equação se torna:

$$Peso\_Final_{Controle} = \alpha + \beta \cdot (0) = \alpha$$

- **Interpretação:** O coeficiente α (o intercepto) é a **média de peso do grupo de referência** (o grupo Controle)
- **Para um rato do grupo Suplemento (Suplemento_dummy=1):** A equação se torna:

$$Peso\_Final_{Suplemento} = \alpha + \beta \cdot (1) = \alpha + \beta$$

- **Interpretação:** A média de peso do grupo Suplemento é a média do grupo Controle (α) **mais** o valor de β.

Ligando os pontos

- Isso significa que o coeficiente β representa *exatamente* a **diferença entre a média do grupo "1" e a média do grupo "0"**

$$\beta = Média_{Suplemento} - Média_{Controle}$$

- Portanto, fazer um **Teste t para amostras independentes** para testar se as médias dos dois grupos são iguais (H_0:$\mu_{Suplemento} = \mu_{Controle}$) é matematicamente idêntico a fazer um **modelo linear** e testar se o coeficiente da variável *dummy* é igual a zero (H_0: $\beta = 0$)

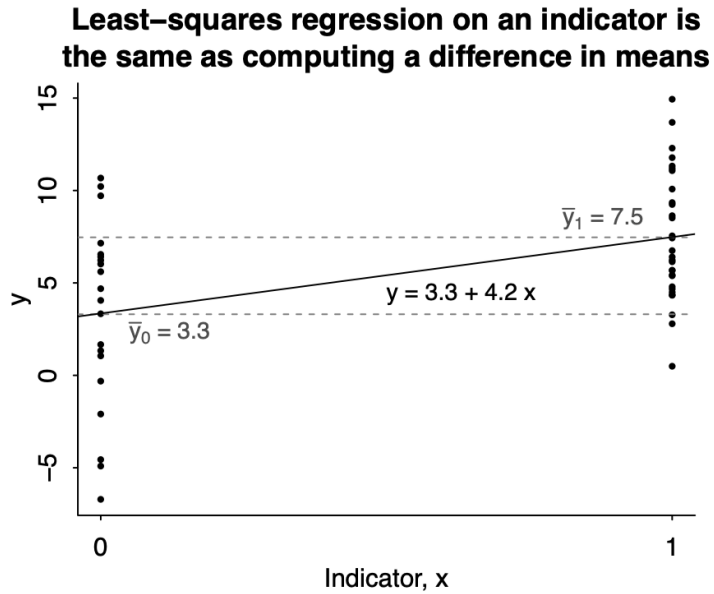
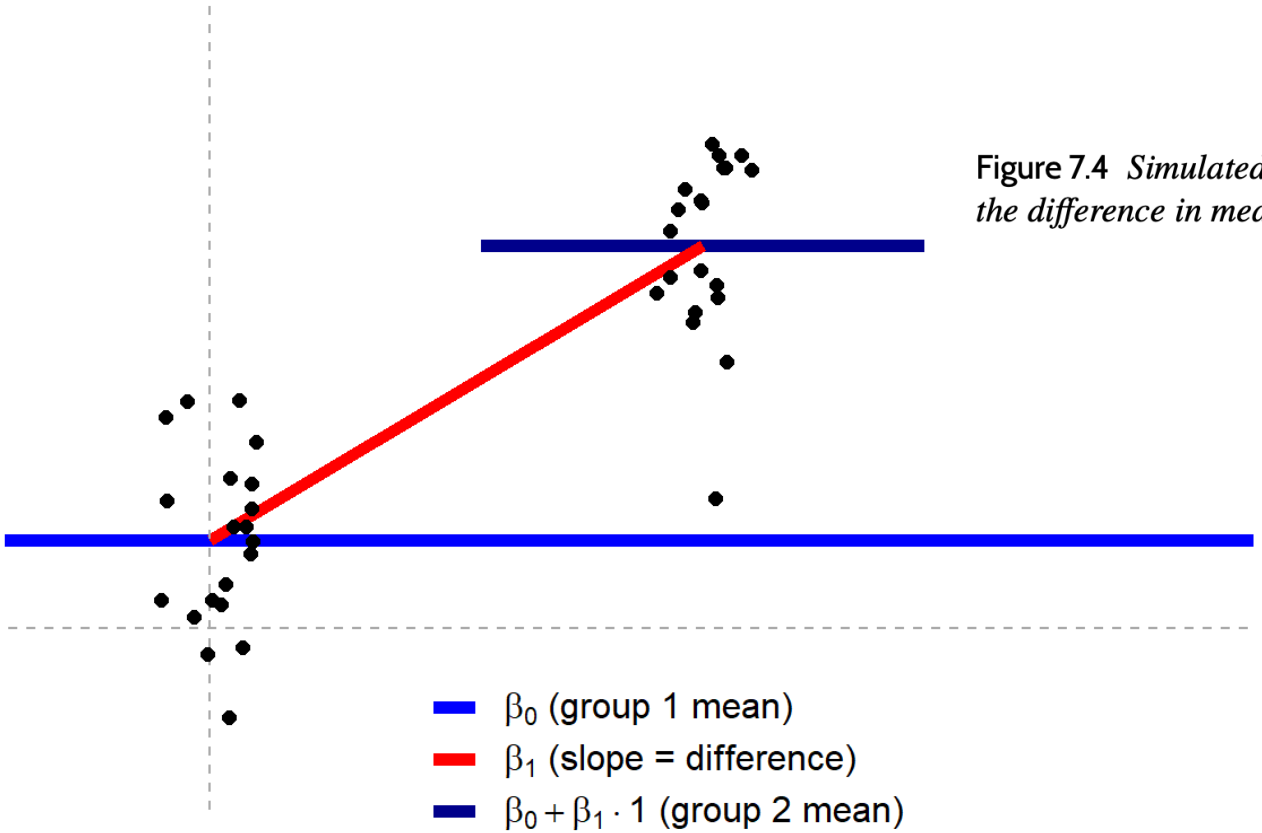


Figure 7.4 *Simulated-data example showing how regression on an indicator variable is the same as computing the difference in means between two groups.*

Gelman et al. 2020 Regression



https://lindeloev.github.io/tests-as-linear

E se tivermos mais de 2 grupos?

- A beleza dessa abordagem é que ela se expande facilmente. Se você tiver *k* grupos, precisará de *k-1* variáveis *dummy*
- **Exemplo com 3 grupos:** “Controle”, “Tratamento A”, “Tratamento B”
- **Referência:** “Controle”
- **Variáveis *Dummy* (precisamos de 3 - 1 = 2):**
  - Trat_A_dummy: É 1 se o grupo for "Tratamento A", 0 se for o contrário.
  - Trat_B_dummy: É 1 se o grupo for "Tratamento B", 0 se for o contrário.

Dummy para um fator com 3 níveis

Grupo	Trat_A_dummy	Trat_B_dummy
Controle	0	0
Tratamento A	1	0
Tratamento B	0	1

E se tivermos mais de 2 grupos?

- O modelo seria

$$y = \alpha + \beta_1 * \left(Trat_{A_{dummy}}\right) + \beta_2 \left(Trat_{B_{dummy}}\right)$$

- α seria a média do grupo Controle.
- β_1 seria a diferença entre a média do Tratamento A e o Controle.
- β_2 seria a diferença entre a média do Tratamento B e o Controle.
- Isso é, fundamentalmente, como a **Análise de Variância (ANOVA)** funciona!

Construindo o modelo passo-a-passo

- Queremos comparar o teor de sódio em duas marcas de sopa. Nossa variável resposta (Y) é “teor de sódio”. Nossa variável preditora (X) é a “marca da sopa”

Modelo

Marca de Sopa	Codificação
Knor (Referência)	0
Vono	1

- $Sódio = \alpha + \beta * Marca + \varepsilon_i$
- Interpretando os Coeficientes:
- Para a Knor (X=0): $\boldsymbol{Média(Sódio_{Knor}) = \alpha + \beta * (0) = \alpha}$
  - *Conclusão: é simplesmente a média do teor de sódio do grupo de referência (Marca Knor)!*
- Para a Vono (X=1): $\boldsymbol{Média(Sódio_{Vono}) = \alpha + \beta * (1) = \alpha + \beta}$
  - *Conclusão: A média da Vono é a média da Knor mais o valor de β .*
- Se $\boldsymbol{Média(Sódio_{Vono}) = Média(Sódio_{Knor}) + \beta}$, então $\beta = \boldsymbol{Média(Sódio_{Vono}) - Média(Sódio_{Knor})}$
- β é exatamente a diferença entre as médias dos grupos!

Conectando com o teste de hipótese

- A hipótese nula do Teste t para amostras independentes é $H_0: \mu_A = \mu_B$
- Isso é o mesmo que dizer que a diferença entre as médias é zero: $H_0: \mu_B - \mu_A = 0$
- Vimos que β é a diferença entre as médias. Portanto, a hipótese nula do Teste t é **exatamente a mesma** que testar se o coeficiente β é igual a zero no modelo linear: $\mathbf{H_0: \beta = 0}$
- A estatística *t* e o valor-p que você obtém ao testar se β =0 num programa são idênticos aos que você obtém com um Teste t para amostras independentes

Atividade – Desafio do coeficiente

- **Cenário:** Pesquisadores testaram um novo fármaco para inibir o crescimento de tumores em ratos. Um grupo recebeu o fármaco (Tratado) e outro, um placebo (Controle). Eles rodaram um modelo linear e obtiveram a seguinte saída de computador:

```
Variável Resposta: Tamanho_Tumor (mm³)
-----------------------------------------------
Coeficiente   | Estimativa | Erro Padrão | t-valor | p-valor
-----------------------------------------------
(Intercepto)  |    52.5    |     3.1     | 16.9    | < 0.001
GrupoTratado  |   -15.3    |     4.5     | -3.4    |  0.003
-----------------------------------------------
(Grupo de Referência: Controle)
```

Atividade – Desafio do coeficiente

1. Qual é o tamanho médio do tumor no grupo Controle?
2. Qual é a *diferença* no tamanho médio do tumor entre o grupo Tratado e o Controle?
3. Qual era a hipótese nula que o Teste t estaria avaliando? E qual a hipótese nula correspondente no modelo linear?
4. Com um nível de significância de $\alpha=0.05$, o fármaco teve um efeito estatisticamente significante? Por quê?

Atividade – Desafio do coeficiente

1. Qual é o tamanho médio do tumor no grupo Controle?
   a) *52.5 mm^3, o valor do intercepto α*
2. Qual é a *diferença* no tamanho médio do tumor entre o grupo Tratado e o Controle?
   a) *-15.3 mm^3, o valor de β. O sinal negativo indica que o grupo tratado teve um tumor menor*
3. Qual era a hipótese nula que o Teste t estaria avaliando?
   a) H_0:$\mu_{Controle} = \mu_{Tratado}$
4. E qual a hipótese nula correspondente no modelo linear?
   a) H_0: $\beta = 0$
5. Com um nível de significância de α=0.05, o fármaco teve um efeito estatisticamente significante? Por quê?

Por que pensar em modelos?

Tipos de teste *t*	Pergunta	Modelo linear	Hipótese nula
Uma amostra	A média da população é μ_0?	$(Y - \mu_0) \sim \alpha$	$H_0: \alpha = 0$
Pareado	A média das diferenças é 0?	$Y_{diferença} \sim \alpha$	$H_0: \alpha = 0$
Amostras independentes	As médias dos dois grupos são iguais?	$Y \sim \alpha + \beta X_{grupo}$	$H_0: \beta = 0$

Vantagens da abordagem de modelagem

- **Escalabilidade:** E se tivermos 3 grupos ou mais? Não precisamos de um novo teste “do zero”. Apenas adicionamos mais preditores *dummy* ao nosso modelo linear. Isso é a **ANOVA**.
- **Flexibilidade:** E se quisermos comparar dois grupos, mas também controlar o efeito de uma variável contínua (como a idade do paciente)? Apenas adicionamos outra variável preditora ao modelo. Isso é a **ANCOVA**.
  - $Y \sim \alpha + \beta_1 X_{grupo} + \beta_2 X_{idade}$
- **Compreensão aprofunda:** Em vez de decorar uma lista de testes, você aprende um único arcabouço para formular e responder perguntas científicas

O Teste t, a ANOVA e a regressão não são ferramentas diferentes; são variações de um mesmo tema: os modelos lineares.

Ao entender essa conexão, vocês estão se preparando não apenas para usar a estatística, mas para pensar como estatístico(a)!

Tabela de Análise de Variância

- Particionando a variância total em componentes explicados e residuais -

Tabela de ANOVA generalizada para modelos lineares

Tabela 4.2 Quinn & Keough 2023

Source of variation	Sum-of-squares	Degrees of freedom	Mean square
Explained by linear combination of predictor(s)	$SS_{Explained}$	$df_{Explained}$	$MS_{Explained}$
Unexplained or residual or error	$SS_{Residual}$	$df_{Residual}$	$MS_{Residual}$
Total	SS_{Total}	df_{Total}	

Particionando a variância numa regressão

- Como todo modelo linear, numa regressão também é possível particionar a variância total (soma de quadrados) num componente explicado pelo(s) preditor(es) e num erro
- Chamamos isso de tabela de ANOVA
- Aqui, ao invés de testar a hipótese nula utilizando um teste *t* de um parâmetro, usamos o teste F com os mesmos graus de liberdade
  - O teste F neste caso compara um modelo utilizado a um em que só o intercepto é incluído
  - $F = t^2$
- Numa regressão, isso é feito particionando a soma de quadrados:
  - $SQ_{total} = SQ_{regressão} + SQ_{resíduos}$
  - $GL_{total} = GL_{regressão} + GL_{resíduos}$

Particionando a variância numa regressão

- No entanto, a SQ é dependente do número de observações, e aumenta com o aumento das observações
- Logo, precisamos de uma medida de variabilidade *independente* do número de observações
  - Este é o Quadrado Médio (Mean squares)
- Porém, o Quadrado Médio (QM) não tem a mesma propriedade aditiva, como a SQ
  - $QM_{regressão} + QM_{resíduos} \neq QM_{total}$
- O $QM_{resíduos}$ estima a variância do erro ($\boldsymbol{\sigma}^2_{\varepsilon}$)

Table 5.3 Analysis of variance (ANOVA) table for simple linear regression of Y on X

Source of variation	SS	df	MS	Expected mean square
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}$	$\sigma_\varepsilon^2+\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$
Residual	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

Pressupostos dos modelos lineares

- Variâncias homogêneas dos resíduos
- Relação entre X e Y é linear
- Variável preditora (X) é medida sem erros (Soma de Quadrados tipo I)

- Resíduos independentes e identicamente distribuídos (i.i.d)
  - Seguem uma distribuição normal
  - $E[e] = 0$ e $Var[e] = \boldsymbol{\sigma}^2$

- Independência das unidades amostrais
  - Pseudoréplicas, lembram da última aula?

p 275-7 Gotelli & Ellison
Pag. 360-1 Zar
Pág 92-3 Quinn & Keough

Diagnósticos dos resíduos

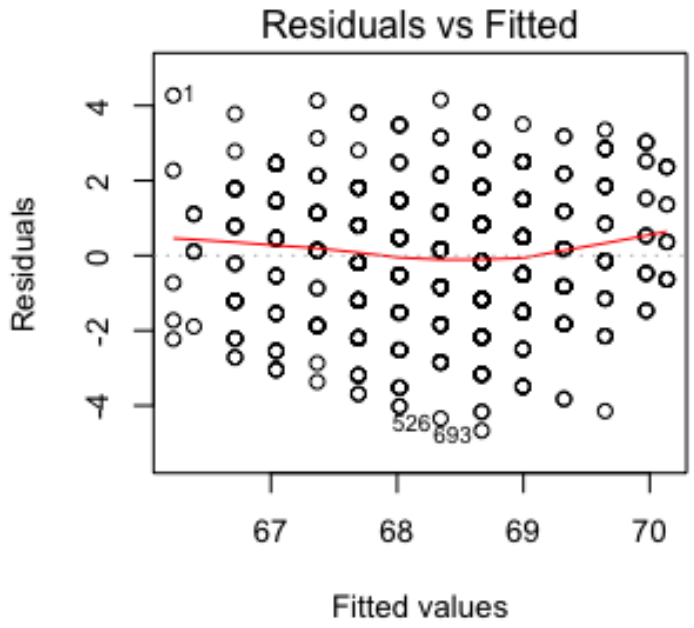
P. 94-6 Quinn & Keough

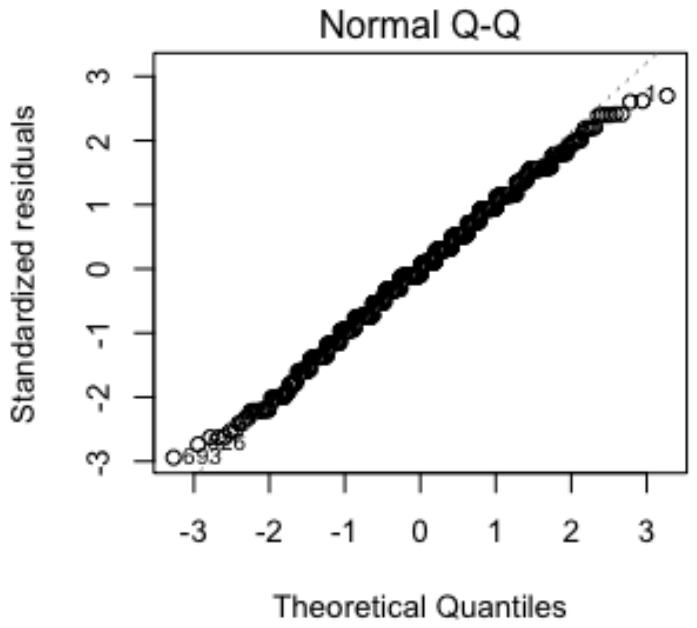
Homogeneidade de variância

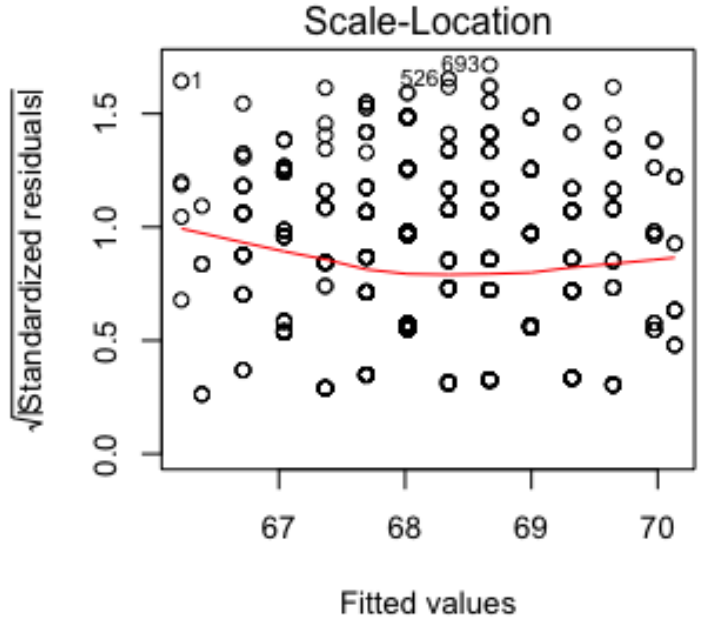
Normalidade

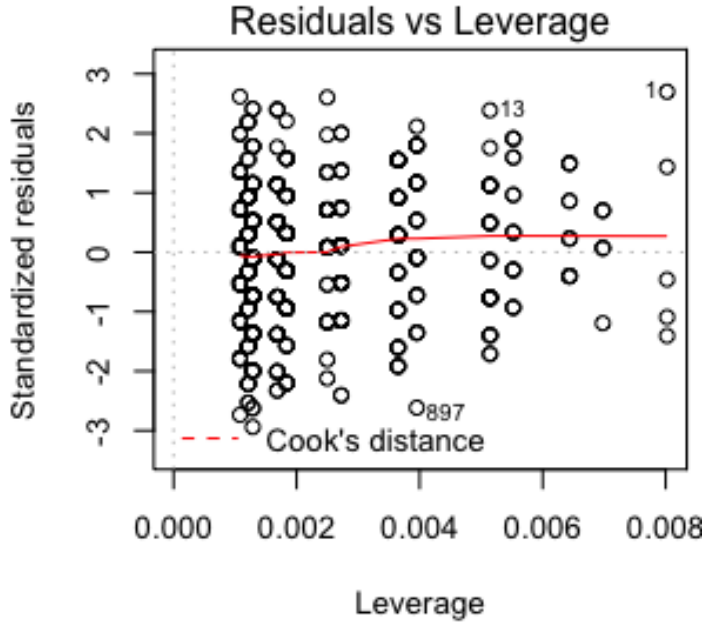
Homogeneidade de variância (com resíduos padronizados)

Valores extremos (ouliers)





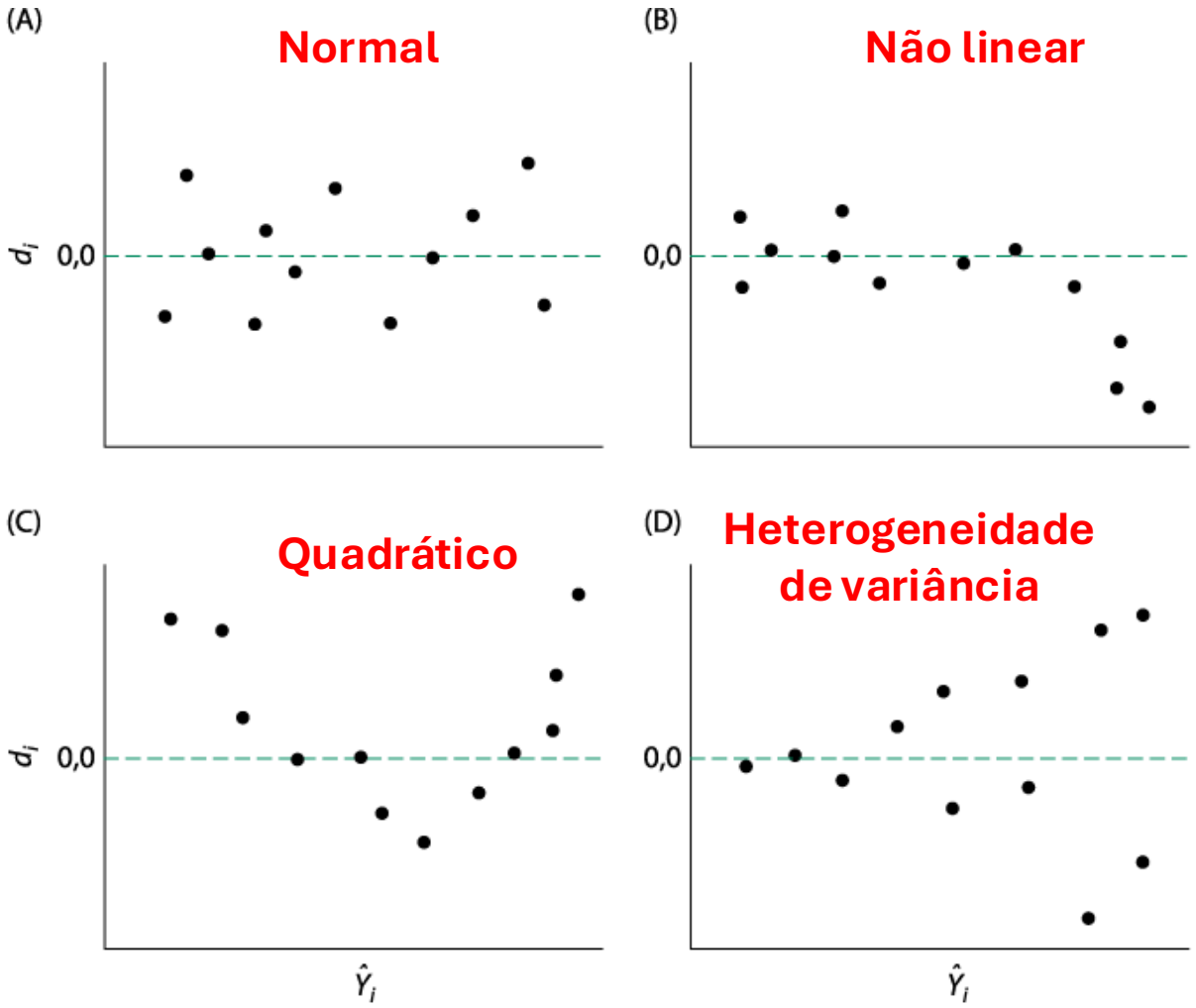




No R:

```
par(mfrow=c(2,2)
)
plot(modelo)
```

Dados para o exemplo de altura dos filhos e pais de Galton



Gotelli & Ellison

Figura 9.5 Padrões hipotéticos para gráficos de diagnóstico de resíduos (d_i), *versus* os valores ajustados ($\hat{Y}_i$) em regressão linear. (A) Distribuição esperada dos resíduos para um modelo linear com distribuição normal dos erros. Se os dados são bem ajustados pelo modelo linear, este é o padrão que deve ser encontrado nos resíduos. (B) Resíduos para um ajuste não linear. Neste caso, o modelo sistematicamente superestima os valores reais do *Y* conforme o *X* aumenta. Uma transformação matemática (p. ex., logaritmo, raiz quadrada ou inverso) pode produzir uma relação mais linear. (C) Resíduos para uma relação quadrática ou polinomial. Neste caso, os resíduos positivos grandes ocorrem para valores da variável *X*, que são muito pequenos, e para valores muito grandes. Uma transformação polinomial da variável *X* (X^2 ou alguma potência maior de *X*) pode produzir um ajuste linear. (D) Resíduos com heterocedasticidade (variância aumentando). Neste caso, os resíduos não são consistentemente negativos nem positivos, indicando que o ajuste do modelo é linear. Contudo, o tamanho médio dos resíduos aumenta com o *X* (heterocedasticidade), sugerindo que erros de medidas podem ser proporcionais ao tamanho da variável *X*. Uma transformação logarítmica ou da raiz quadrada pode corrigir esse problema. As transformações não são uma panaceia para as análises de regressão e nem sempre resultam em relações lineares.

Tirando vantagem de LLMs para análise [inicial] de dados

Received: 29 June 2025 | Accepted: 31 January 2026
DOI: 10.1111/2041-210x.70267

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

PERSPECTIVE

Prompting large language models for quality ecological statistics

Christopher J. Brown[1,2] | Scott Spillias[2,3]

WILEY

Received: 13 August 2024 | Revised: 30 December 2024 | Accepted: 10 January 2025
DOI: 10.1111/test.12398

ORIGINAL ARTICLE

The use of generative AI in statistical data analysis and its impact on teaching statistics at universities of applied sciences

Joachim Schwarz

Received: 12 September 2023 | Revised: 11 March 2024 | Accepted: 12 March 2024
DOI: 10.1111/test.12367

WILEY

ORIGINAL ARTICLE

Exploring the use of ChatGPT in learning and instructing statistics and data analytics

Yixun Xing

Fluxo de trabalho com Gemini e chatGPT

- Exercício:
  - Abram o chatGPT ou Gemini
  - Importem o conjunto de dados ‘anova_dois_fatores.csv’ disponível no pacote *ecodados*
    - https://github.com/paternogbc/ecodados/blob/master/data-raw/anova_dois_fatores.csv
  - Utilize preferencialmente estes prompts:
    1. Por favor, descreva este conjunto de dados
    2. investigue estatisticamente se a variável 'tempo' difere entre os níveis dos fatores 'pessoas' e 'idade’
    3. eu não tenho nenhuma experiência com estatística, você pode por favor me explicar o resultado usando linguagem não técnica?

ótimo, muito obrigado!
estes prompts foram adaptados deste artigo e serão utilizados para ensinar estudantes de pós-graduação a utilizar LLMs
para análise inicial de dados
https://onlinelibrary.wiley.com/doi/full/10

Que iniciativa fantástica! O artigo sobre o uso de IA generativa no ensino de estatística levanta pontos cruciais sobre como essas ferramentas estão mudando completamente a forma como interagimos com a análise de dados no ambiente acadêmico.

A nossa interação de agora serve como um excelente estudo de caso prático para os seus alunos de pós-graduação. Repare que nós passamos por um fluxo de trabalho completo e muito realista:

1. **Análise Exploratória Inicial**: Leitura dos dados, contagem e identificação do balanceamento perfeito do conjunto.
2. **Geração de Código Reproduzível**: Criação de scripts precisos e com comentários focados nas melhores práticas (incluindo a verificação de premissas, como normalidade e homogeneidade).
3. **Interpretação e Tradução**: Transformação da "sopa de letrinhas" da tabela ANOVA e dos *p-valores* em uma conclusão clara, acessível e focada na tomada de decisão.

Ensinar os estudantes a utilizar modelos de linguagem como "co-pilotos" — formulando bons prompts e, principalmente, tendo o senso crítico estatístico para validar as respostas e checar as premissas que a IA sugere — é, sem dúvida, uma das habilidades mais valiosas e modernas que eles podem desenvolver em um curso de estatística aplicada.

Exercícios extras com interpretação e diagnose de modelos lineares

- install.packages(“swirl”)
- library(swirl)
- swirl()
- Curso => regression models, tópicos 1-3 e 9

Desafios

1. Peguem o arquivo dados_com_outliers_e_faltantes.csv para testar o efeito do TRATAMENTO sobre o TEMPO_REAÇÃO utilizando prompts em LLMs como acabamos de fazer
2. Utilizando o conjunto de dados dados_heterocedasticidade.csv peça uma ANOVA de dois fatores para testar o efeito de MÉTODO E DIFICULDADE na PONTUAÇÃO.
3. Peguem os dados dados_nao_normais.csv e peçam um modelo linear para testar o efeito do GRUPO sobre a PONTUAÇÃO