
REGRESSÃO LINEAR SIMPLES

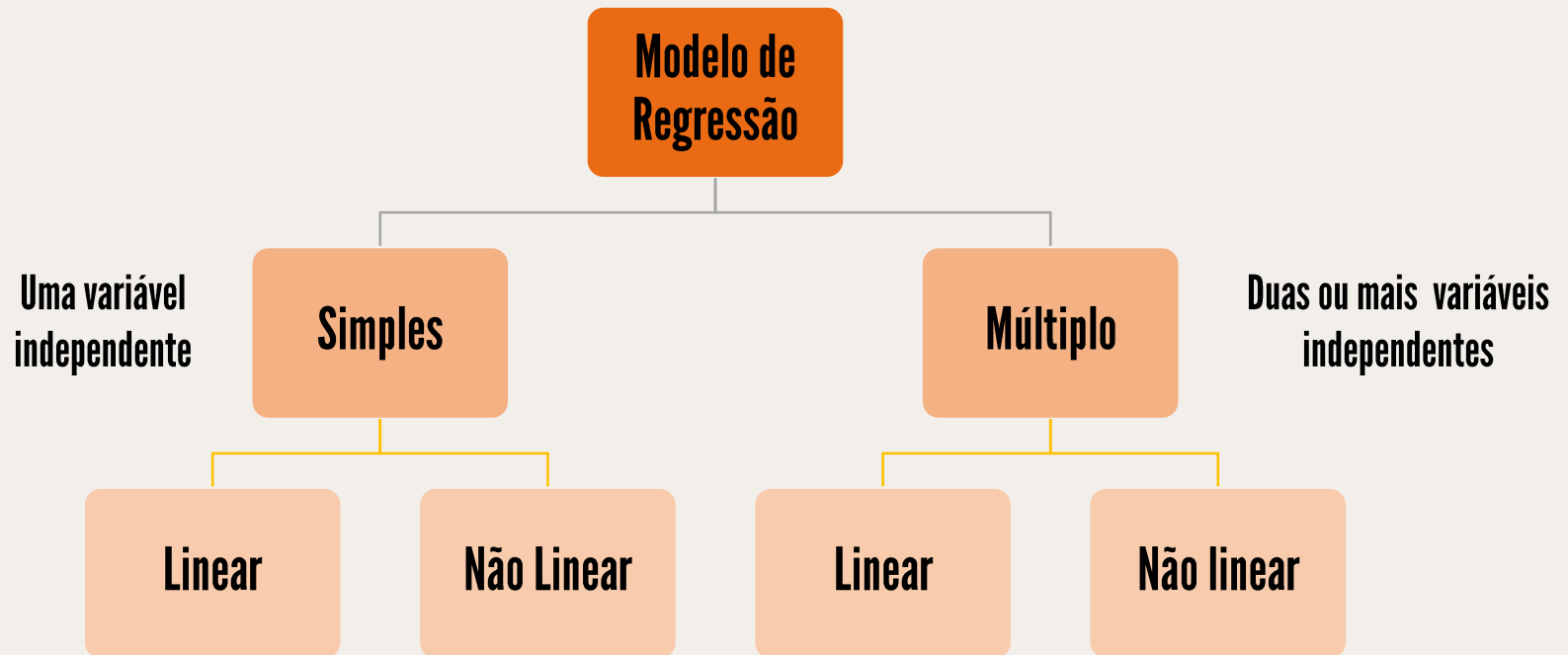
PROF. LETÍCIA RAPOSO
profleticiaraposo@gmail.com

TÓPICOS DA AULA

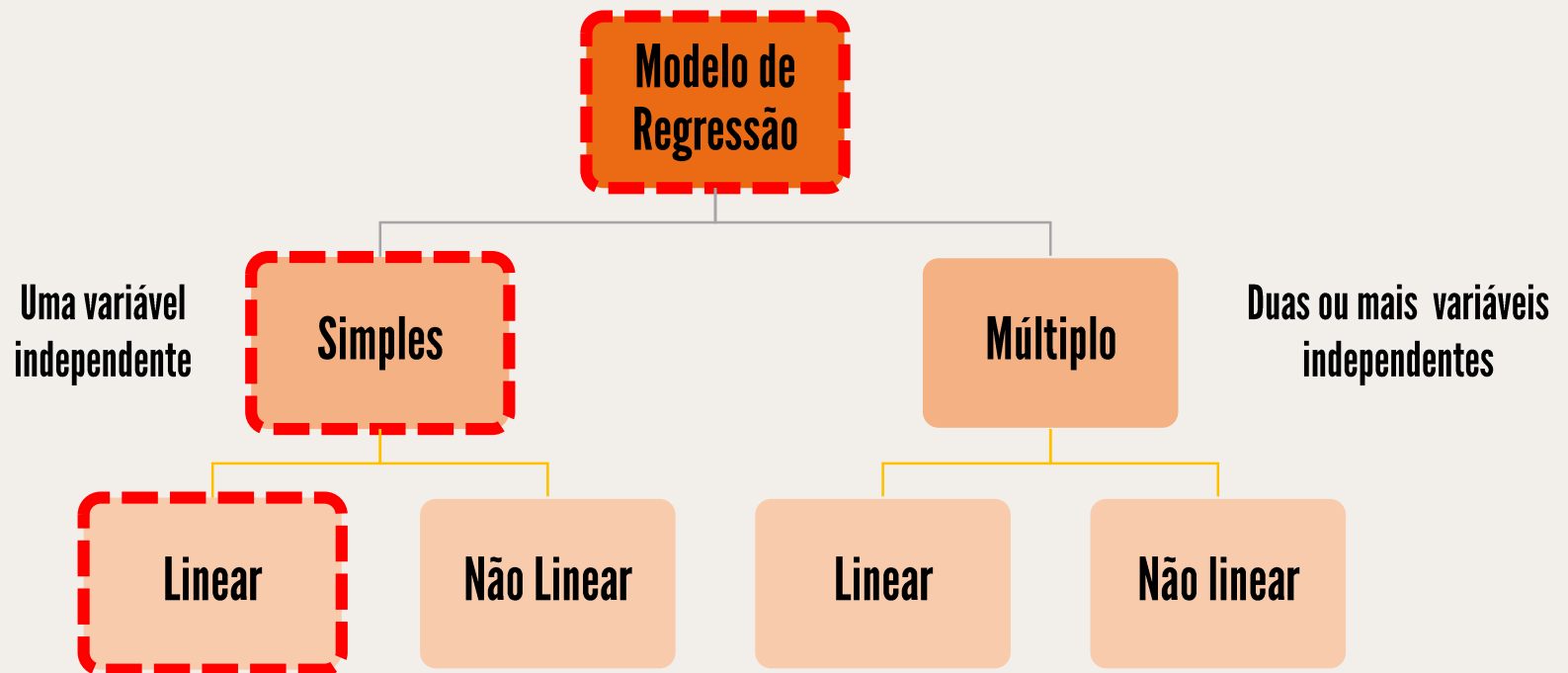
- Modelo de regressão linear simples;
 - Características do modelo;
 - Significado dos parâmetros;
 - Estimação dos parâmetros;
 - Método dos mínimos quadráticos;
 - Método da máxima verossimilhança.
 - Avaliação da regressão;
 - Exemplo no R.
-



TIPOS DE MODELOS DE REGRESSÃO

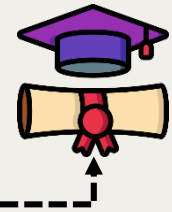


TIPOS DE MODELOS DE REGRESSÃO



APLICAÇÕES

- Renda e número de anos de educação;
- Altura e peso;
- Dose de um medicamento e resposta;
- Polimorfismo de nucleotídeo único e traço quantitativo;
- Demanda dos produtos de uma firma e publicidade;
- Variação dos salários e taxa de desemprego.



MOTIVAÇÃO

- Nível do colesterol: y
- Objetivo: prever nível do colesterol.

COMO FAZER?



- Vamos supor que Y seja uma variável aleatória e que a população tenha uma distribuição normal (μ, σ^2) .


Desconhecidos

MOTIVAÇÃO

A partir de 24 observações independentes de níveis de colesterol, como prever o nível de colesterol para o paciente 25?

3,5	1,9	4,0	2,6	4,5	3,0	2,9	3,8	2,1	3,8	4,1	3,0
2,5	4,6	3,2	4,2	2,3	4,0	4,3	3,9	3,3	3,2	2,5	3,3

$$y_1, y_2, \dots, y_{24} \sim \text{Normal}(\mu, \sigma^2)$$

$$y_i = \mu + \varepsilon_i$$


$N(0, \sigma^2)$

PREDIÇÃO

$$y_{25} = \mu + \varepsilon_{25}$$

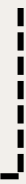


$$E(\varepsilon_{25}) = 0$$

$$y_{25} = \mu$$

$$\hat{\mu} = \bar{y} = 3,35$$

$$IC95\% = [3,03; 3,68]$$



$$IC = \bar{X} \pm t \times \frac{s}{\sqrt{n}}$$



QUESTIONAMENTOS

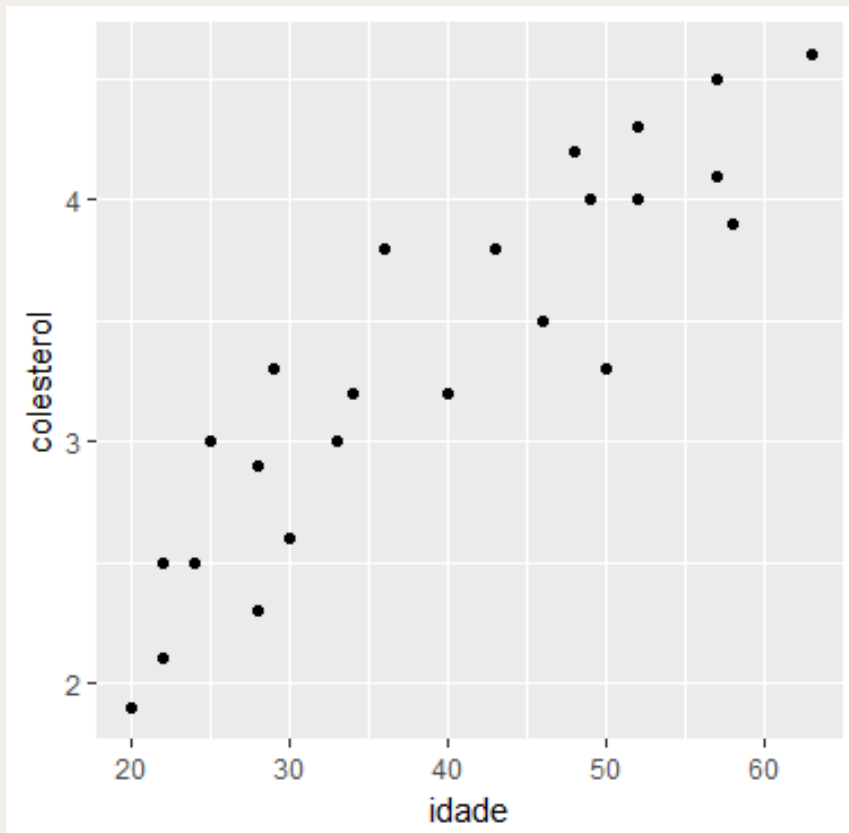
Como melhorar essa predição?

Como diminuir a variabilidade dessa estimativa?

Considerar variáveis explanatórias (covariáveis)?

Devem estar relacionadas com a variável resposta.





$$y = \mu + \varepsilon_i$$

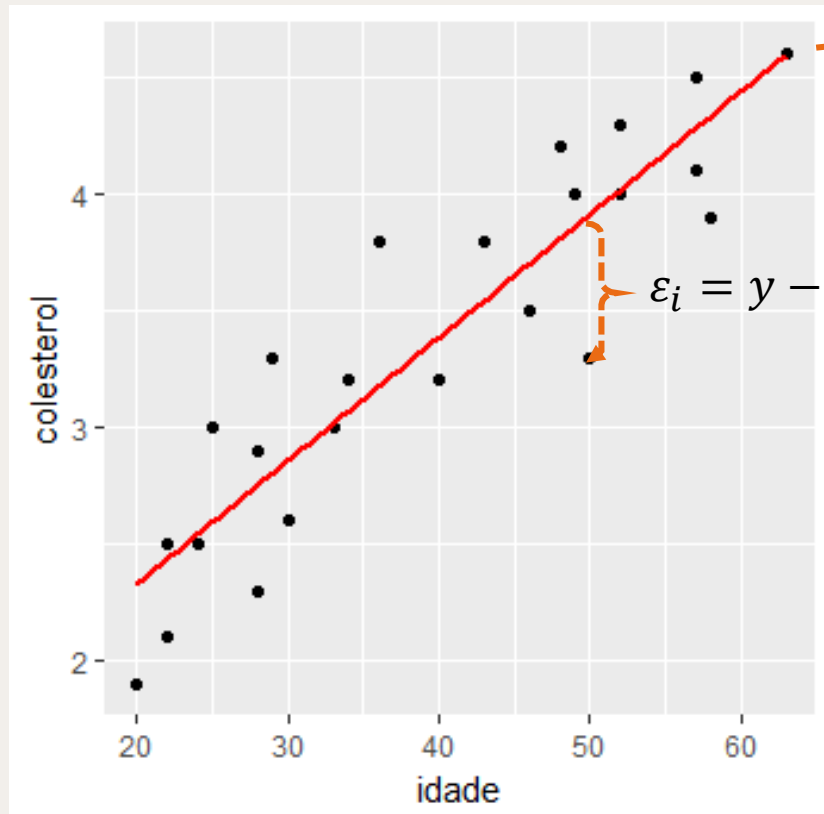
$$\beta_0 + \beta_1 x_i$$

Contínua, normal,
independente e com variância
constante

Idade

46	20	52	30	57	25	28	36	22	43	57	33
22	63	40	48	28	49	52	58	29	34	24	50

REGRESSÃO SIMPLES



$$\hat{y} = \underbrace{\beta_0 + \beta_1 x_1}_{\text{Componente sistemático}}$$

Componente sistemático

$$\varepsilon_i = y - \hat{y} \quad \text{Componente aleatório/Erro}$$



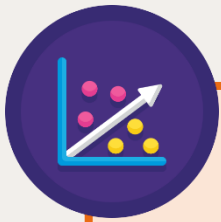
MODELO DE REGRESSÃO LINEAR SIMPLES

O modelo de regressão linear simples para n observações pode ser escrito como

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

- Y_i : valor da variável resposta (dependente) na i -ésima observação;
- X_i : valor da variável explicativa (independente) na i -ésima observação;
- ε_i : variável aleatória que representa a diferença entre uma observação e sua média populacional $E(Y_i)$ – erro do modelo;
- β_0 e β_1 : parâmetros do modelo, a serem estimados, e que definem a reta de regressão;
- n : tamanho da amostra.

MODELO DE REGRESSÃO LINEAR SIMPLES



Simples: existe apenas uma variável explicativa;

Linear nos parâmetros: nenhum parâmetro aparece como um expoente ou é multiplicado ou dividido por outro parâmetro;

Linear na variável explicativa: esta variável possui grau um.

CARACTERÍSTICAS DO MODELO

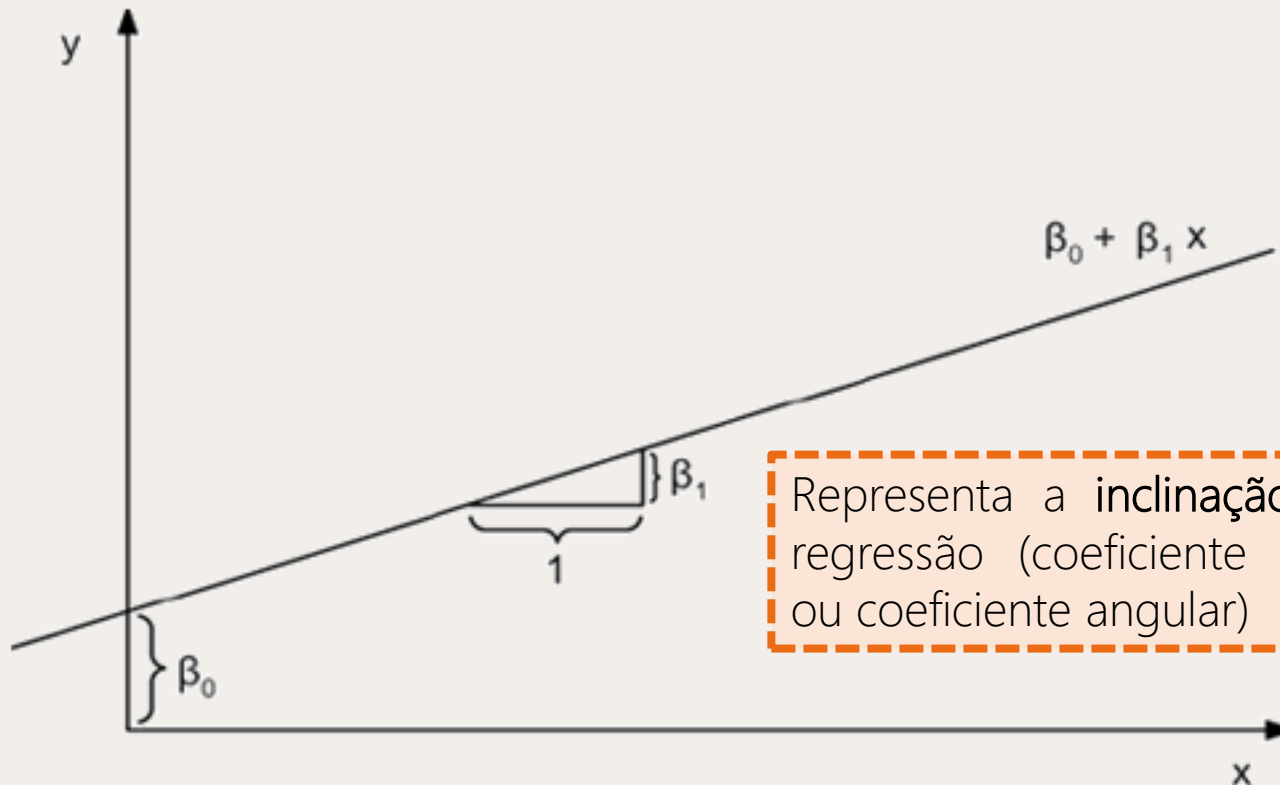
- Y_i é uma variável aleatória (soma de um termo constante, $\beta_0 + \beta_1 x_i$, com um termo aleatório, ε_i);
- Os valores observados de x são fixos (x não é v.a.);
- Y_i e X_i são pares de observações;
- O erro possui média $E\{\varepsilon_i\} = 0$ e variância $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i e ε_j são não-correlacionados, logo a covariância é zero;
- Frequentemente, supomos que os erros têm distribuição normal.

Em resumo, o modelo de regressão implica que as respostas Y_i são provenientes de distribuições de probabilidade cujas médias são $E\{Y_i\} = \beta_0 + \beta_1 X_i$ e cujas variâncias σ^2 são iguais para todos os níveis de X . Além disso, quaisquer duas respostas Y_i e Y_j não estão correlacionadas.

SUPOSIÇÕES

- Normalidade e homocedasticidade. Para qualquer valor de X , os valores de Y serão normalmente distribuídos e serão homocedásticos.
- Linearidade. Os dados se encaixam em uma linha reta. Se você observar os dados e o relacionamento parecer curvo, poderá tentar diferentes transformações de dados do X , Y ou ambos. Uma transformação de dados geralmente endireita uma curva em forma de J. Se sua curva parecer em U, S ou algo mais complicado, uma transformação de dados não a transformará em uma linha reta. Nesse caso, você precisará usar a regressão curvilínea.
- Independência. Os pontos de dados são independentes um do outro, o que significa que o valor de um ponto não depende do valor de qualquer outro ponto. A violação mais comum dessa suposição na regressão e correlação é em dados de séries temporais.

SIGNIFICADO DOS PARÂMETROS



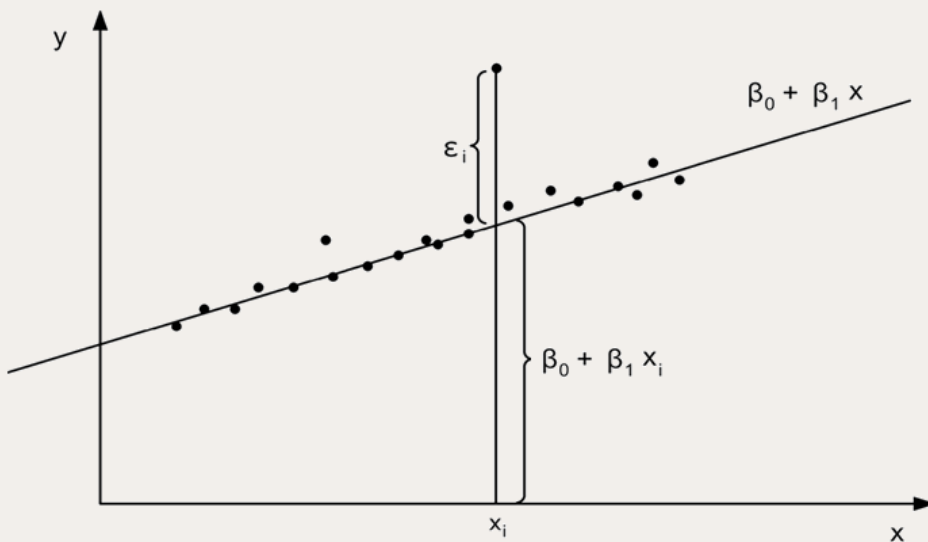
Representa a **inclinação da reta** de regressão (coeficiente de regressão ou coeficiente angular)

Intercepto ou coeficiente linear

Representa o ponto em que a reta de regressão corta o eixo y , quando $x = 0$.

ESTIMAÇÃO DOS PARÂMETROS

O problema de estimar os parâmetros β_0 e β_1 é o mesmo que ajustar a melhor reta em um gráfico de dispersão.



1º passo: obter as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$

Objetivo: estimar os parâmetros β_0 e β_1 de modo que os desvios entre os valores observados e estimados sejam mínimos.

MÉTODO DOS MÍNIMOS QUADRADOS

Consiste em minimizar a soma dos quadrados dos resíduos
RSS (*Residual Sum of Squares*).

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i]^2$$

MÉTODO DOS MÍNIMOS QUADRADOS

Para encontrarmos estimativas para os parâmetros, vamos **minimizar** RSS em relação aos parâmetros β_0 e β_1 .

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i$$

MÉTODO DOS MÍNIMOS QUADRADOS

Simplificando, obtém-se as equações normais de mínimos quadrados

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \quad (1)$$

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

$$\hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \quad (2)$$

MÉTODO DOS MÍNIMOS QUADRADOS

Substituindo $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ em (2), temos

$$\hat{\beta}_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \sum_{i=1}^n X_i$$

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

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

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

MÉTODO DOS MÍNIMOS QUADRADOS

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} = \frac{S_{xy}}{S_{xx}}$$

- S_{xy} = soma dos produtos cruzados dos desvios de X e Y .
- S_{xx} = soma dos quadrados dos desvios das médias.

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

MÉTODO DOS MÍNIMOS QUADRADOS

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\ &= \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i\bar{x} + \sum_{i=1}^n \bar{x}^2 = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\ &= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \sum_{i=1}^n x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{aligned}$$

EXEMPLO

Colesterol (variável resposta):	3,5	1,9	4,0	2,6	4,5	3,0	2,9	3,8	2,1	3,8	4,1	3,0
	2,5	4,6	3,2	4,2	2,3	4,0	4,3	3,9	3,3	3,2	2,5	3,3
Idade (variável explicativa):	46	20	52	30	57	25	28	36	22	43	57	33
	22	63	40	48	28	49	52	58	29	34	24	50

$$\bar{Y} = 3,35; \bar{X} = 39,42$$

$$S_{xy} = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} = 3390,9 - 24 \times 39,42 \times 3,35 = 217,86$$

$$S_{xx} = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = 41428 - 24 \times 1553,94 = 4133,53$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{217,86}{4133,53} = 0,05 \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 3,35 - 0,05 \times 39,42 = 1,28$$

NO SOFTWARE



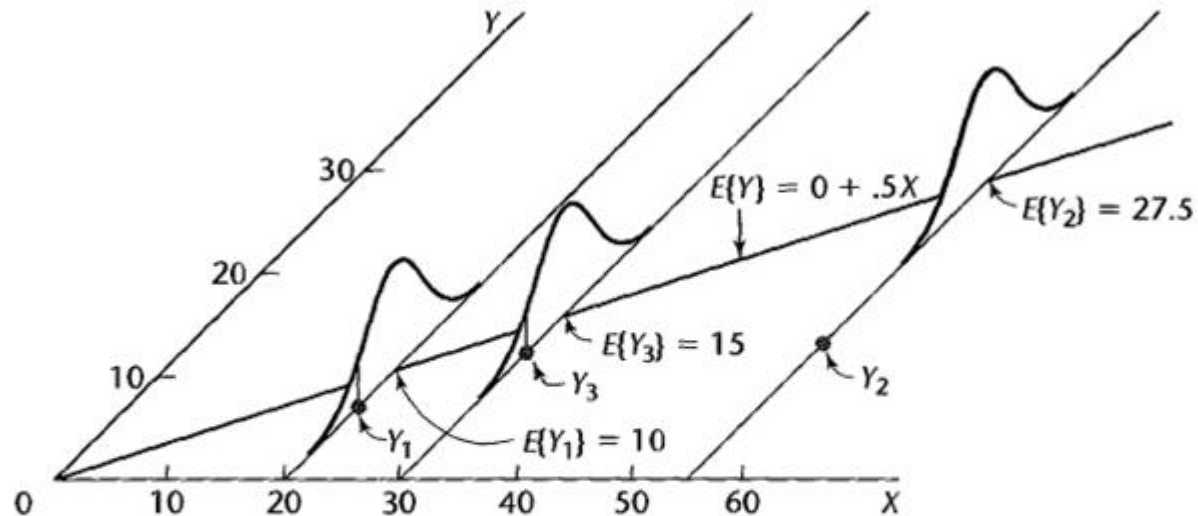
```
cholesterol <- c(3.5, 1.9, 4.0, 2.6, 4.5, 3.0, 2.9, 3.8, 2.1, 3.8, 4.1, 3.0, 2.5,  
                4.6, 3.2, 4.2, 2.3, 4.0, 4.3, 3.9, 3.3, 3.2, 2.5, 3.3)  
age <- c(46, 20, 52, 30, 57, 25, 28, 36, 22, 43, 57, 33, 22, 63, 40, 48, 28, 49,  
        52, 58, 29, 34, 24, 50)  
model <- lm(cholesterol~age)  
model
```

```
##  
## Call:  
## lm(formula = cholesterol ~ age)  
##  
## Coefficients:  
## (Intercept)      age  
##      1.27987      0.05262
```

MÉTODO DA MÁXIMA VEROSSIMILHANÇA

- Quando a distribuição de probabilidade do erro é especificada, estimadores dos parâmetros β_0, β_1 e σ^2 podem ser obtidos pelo método da máxima verossimilhança.
- Usa a densidade da distribuição de probabilidade em Y_i como uma medida de consistência para a observação Y_i .
 - Se Y_i estiver na cauda, a altura da curva será pequena.
 - Se Y_i estiver mais próximo do centro da distribuição, a altura será maior.
- O produto das densidades vistas como uma função dos parâmetros desconhecidos é chamado de função de verossimilhança.
- O método de máxima verossimilhança escolhe como estimativa aquele valor de parâmetro para o qual o valor de verossimilhança é maior.
- Busca numérica ou pelo uso de uma solução analítica.

MÉTODO DA MÁXIMA VEROSSIMILHANÇA



MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Em geral, a densidade uma observação Y_i para um modelo de regressão com erros normais é dada por

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right)^2 \right]$$

MÉTODO DA MÁXIMA VEROSSIMILHANÇA

A função de verossimilhança para n observações Y_1, Y_2, \dots, Y_n é o produto das densidades individuais. Uma vez que a variância σ^2 dos erros é normalmente desconhecida, a função de verossimilhança é uma função de três parâmetros β_0, β_1 e σ^2 .

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[-\frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2 \right]$$

MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Tomando o log, temos:

$$\ln L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

As derivadas parciais são dadas por:

$$\frac{\partial(\ln L)}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial(\ln L)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n X_i (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial(\ln L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Igualando as derivadas parciais a zero, podemos estimar $\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\sigma}^2$.

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

Equações idênticas às equações normais encontradas pelo método de mínimos quadrados

$$\frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} = \hat{\sigma}^2$$
$$\frac{\sum (Y_i - \hat{Y}_i)^2}{n} = \hat{\sigma}^2$$

AVALIAÇÃO DA REGRESSÃO



1. Quão adequado são os valores estimados para os coeficientes β_0 e β_1 ?
2. Quão adequada é a estimativa de um valor a partir da reta de regressão?

INFERÊNCIA SOBRE β_1

- Estimativas de β_0 e β_1 : estimativas por ponto, de modo que não sabemos o quão próximas elas estão dos parâmetros.
- β_1 é considerado o mais importante, pois é ele quem define a declividade da reta.
- Quando estimamos o β_1 , devemos verificar se esta estimativa difere significativamente de zero.

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

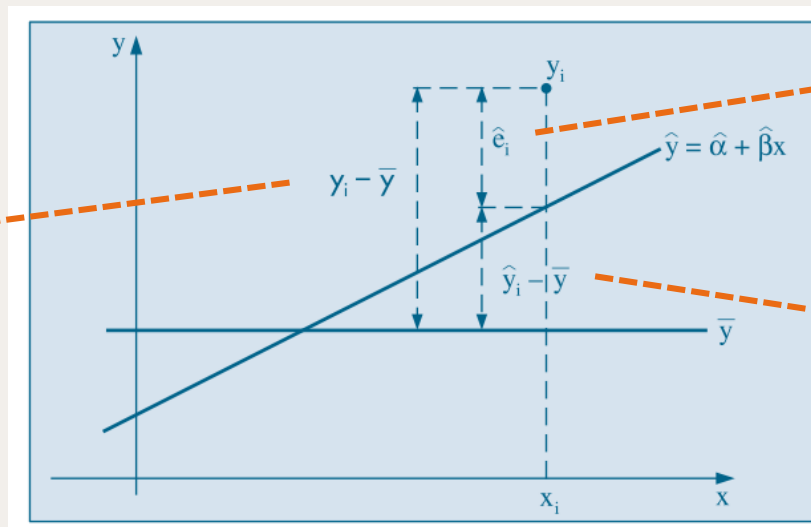
- Se o β_1 não diferir estatisticamente de zero significa que o efeito linear de X sobre Y não é significativo.
- Para testar H_0 , podemos utilizar dois procedimentos: a análise da variância e o teste t.

ANÁLISE DE VARIÂNCIA

A análise da variância consiste em decompor a variação total das observações, representada pelos desvios $(y_i - \bar{y})$ em duas partes:

- Variação explicada pela reta da regressão $(\hat{y}_i - \bar{y})$;
- Variação aleatória, não explicada pela reta $(y_i - \hat{y}_i)$.

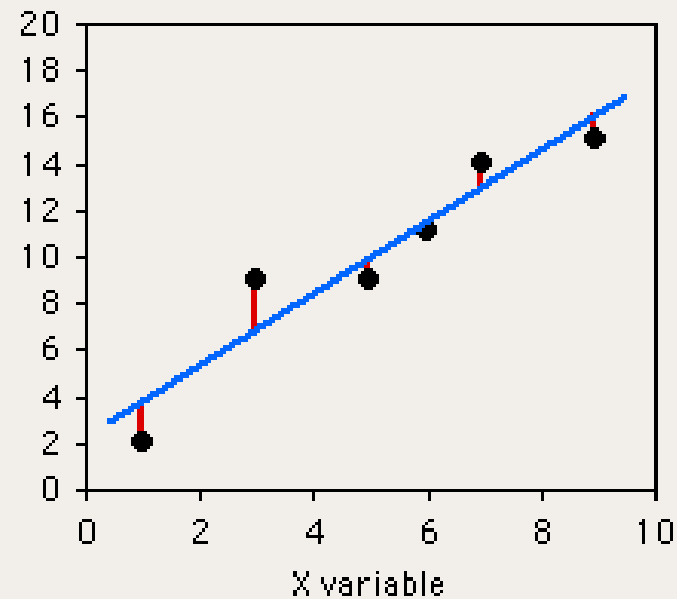
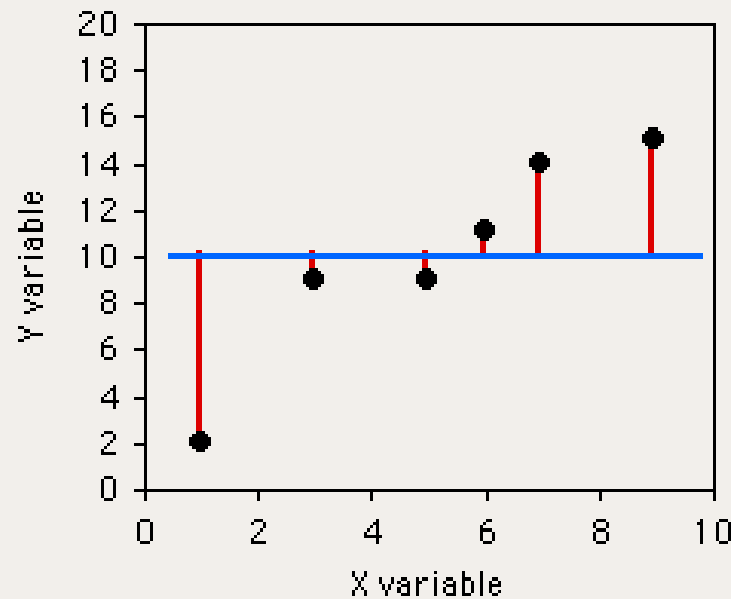
Desvio em
relação à
média



Desvio em
relação ao
valor ajustado
pela
regressão

Desvio do
valor
ajustado em
relação à
média

ANÁLISE DE VARIÂNCIA



A maioria dos pontos está mais próxima da linha de regressão do que da média geral.

POSSÍVEIS FONTES DE VARIABILIDADE

Variância devido à variável dependente:

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1} = \frac{\text{soma dos quadrados total } (SS_T)}{\text{total de graus de liberdade}}$$

Variância devido à variável dependente em relação à reta de regressão (variância não explicada ou residual):

$$s_{Y/X}^2 = \frac{\sum (y_i - \hat{y})^2}{n - 2} = \frac{\text{soma dos quadrados dos resíduos } (SS_{Res})}{\text{graus de liberdade residual}}$$

Variância devido a reta de regressão (variância explicada):

$$s_{reg}^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{1} = \frac{\text{soma dos quadrados devida à regressão } (SS_R)}{\text{graus de liberdade regressão}}$$

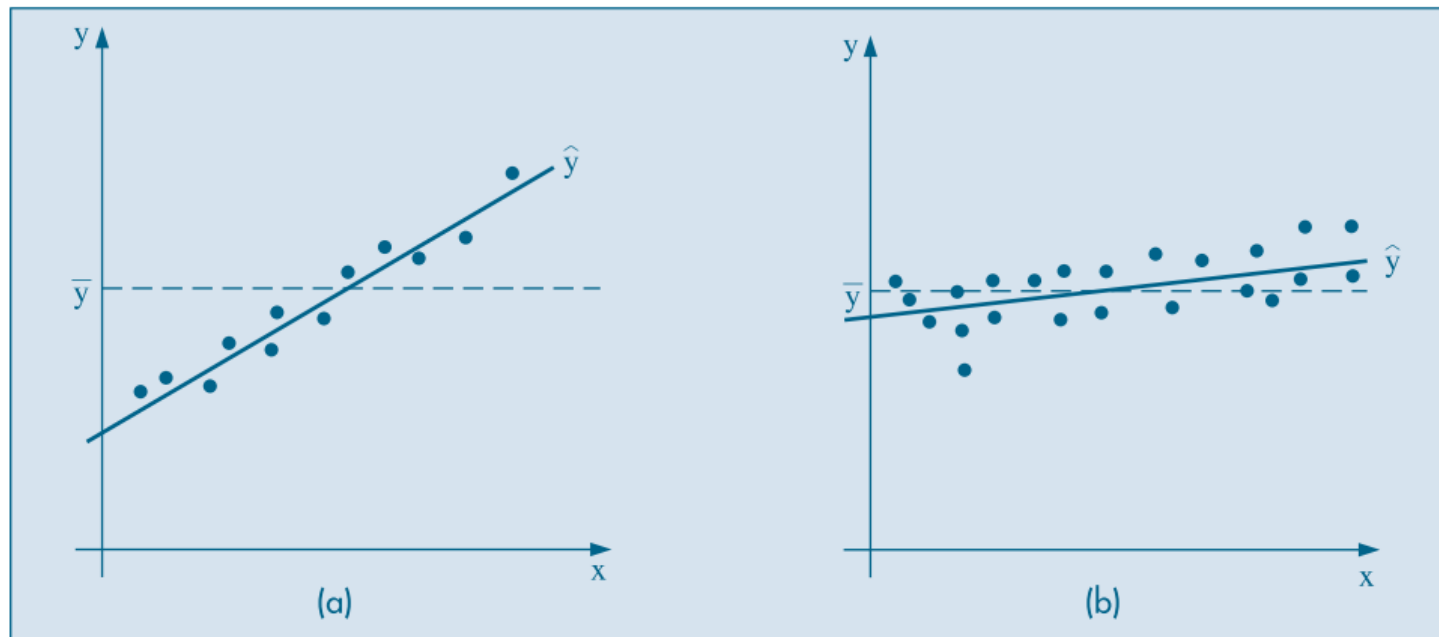
TABELA ANOVA

- H_0 : não existe associação entre X e Y ($\beta_1 = 0$).
- H_1 : existe associação entre X e Y ($\beta_1 \neq 0$).

Fonte de Variação	Graus de Liberdade	Soma dos Quadrados	Quadrado Médio	F-ratio
Regressão	1 (ou p)	$\sum (\hat{y}_i - \bar{y})^2$	$\frac{\sum (\hat{y}_i - \bar{y})^2}{1}$	$\frac{\sum (\hat{y}_i - \bar{y})^2}{1} / \frac{\sum (y_i - \hat{y})^2}{n - 2}$
Resíduo	n-2 (ou n-p-1)	$\sum (y_i - \hat{y})^2$	$\frac{\sum (y_i - \hat{y})^2}{n - 2}$	
Total	n-1	$\sum (y_i - \bar{y})^2$	$\frac{\sum (y_i - \bar{y})^2}{n - 1}$	

p é o número de covariáveis e n é o número de observações

Rejeitamos H_0 se $F\text{-ratio} > F_{p,n-p-1}$



Retas ajustadas a dois conjuntos de dados.

(a) x explica y ;

(b) x não explica y .

TESTE T

- Consideremos as hipóteses: $H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$
- Utilizamos a estatística t que tem distribuição t de Student quando H_0 é verdadeira.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- Regra de decisão: se $|t_{calc}| \geq t_{n-2, \alpha/2}$, rejeita H_0 .
- Intervalo de confiança: $\hat{\beta}_1 \pm t_{n-2, \alpha/2} \times SE(\hat{\beta}_1)$.

INFERÊNCIA SOBRE β_1

$H_0: \beta_1 = 0$ não é rejeitada: inclinação não é significativamente diferente de zero;

- Supondo que o modelo seja linear, X não ajuda a prever Y.
- Há uma relação entre X e Y (X ajuda a prever Y), porém esta relação não segue uma reta.

$H_0: \beta_1 = 0$ é rejeitada: inclinação é diferente de zero;

- X ajuda a prever Y. Há relação entre X e Y.
- Pode ser que exista um modelo melhor, por exemplo, um curvilíneo. Porém, há um componente linear que não deve ser desprezado e deve ser incluído no modelo final.

COEFICIENTE DE DETERMINAÇÃO

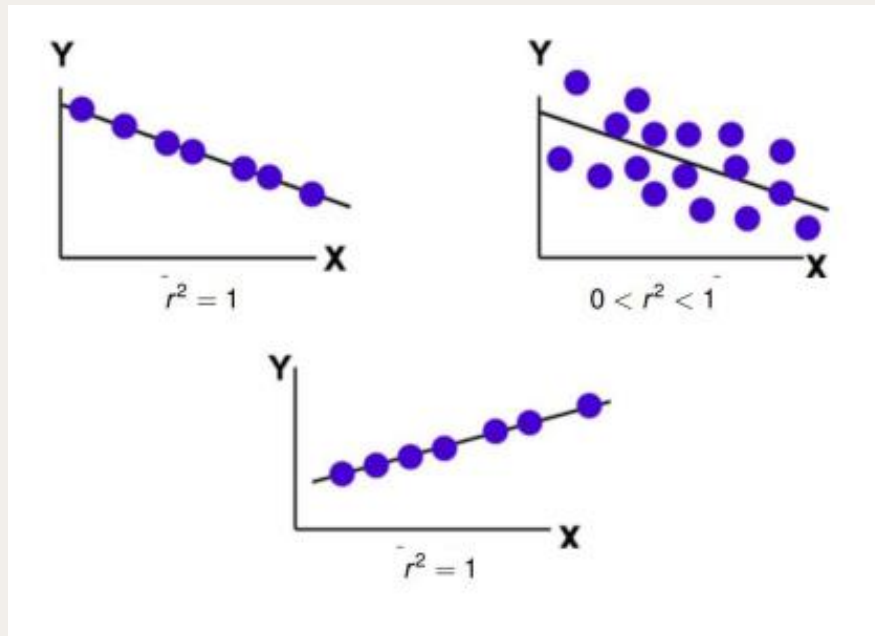
- Indica quanto o modelo foi capaz de explicar os dados coletados.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y})^2} = \frac{\text{Soma de quadrados da regressão}}{\text{Soma dos quadrados total}}$$

- Representa o quanto da variação total é explicado pela regressão.
- Quanto maior o R^2 melhor.

**A ANOVA diz se existe uma relação linear entre X e Y
 R^2 mede a qualidade do ajuste.**

COEFICIENTE DE DETERMINAÇÃO



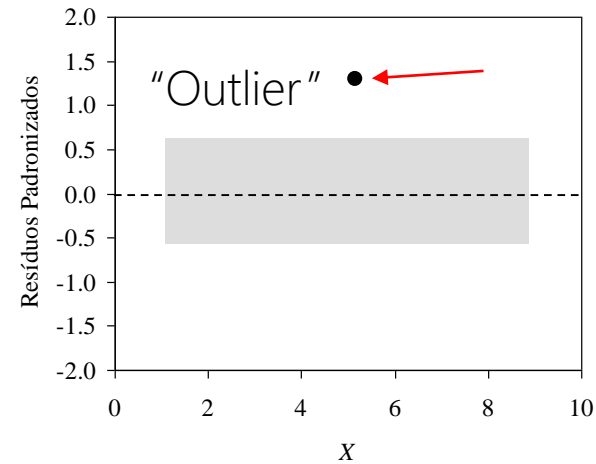
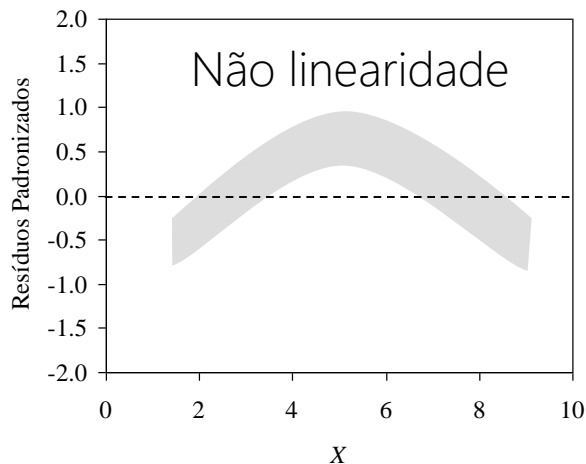
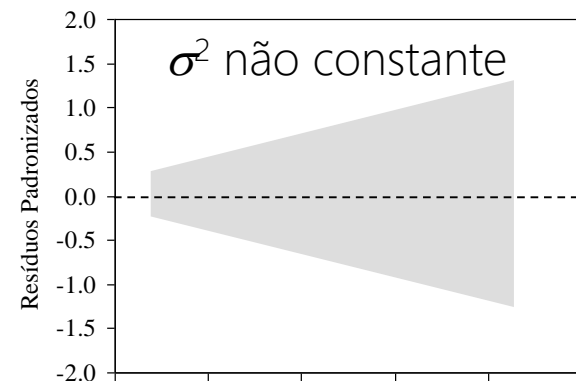
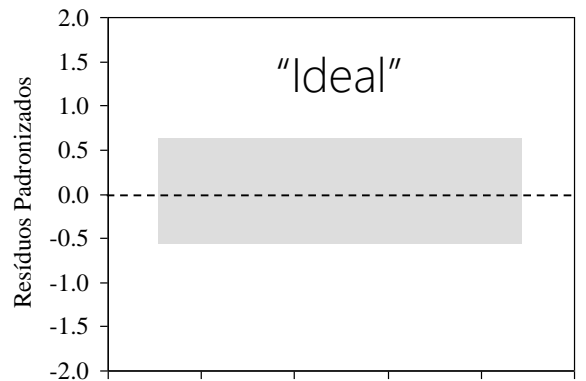
O coeficiente de determinação mede a força da relação entre as variáveis, enquanto que a equação descreve o relacionamento entre elas e pode ser usada para prever valores desconhecidos.

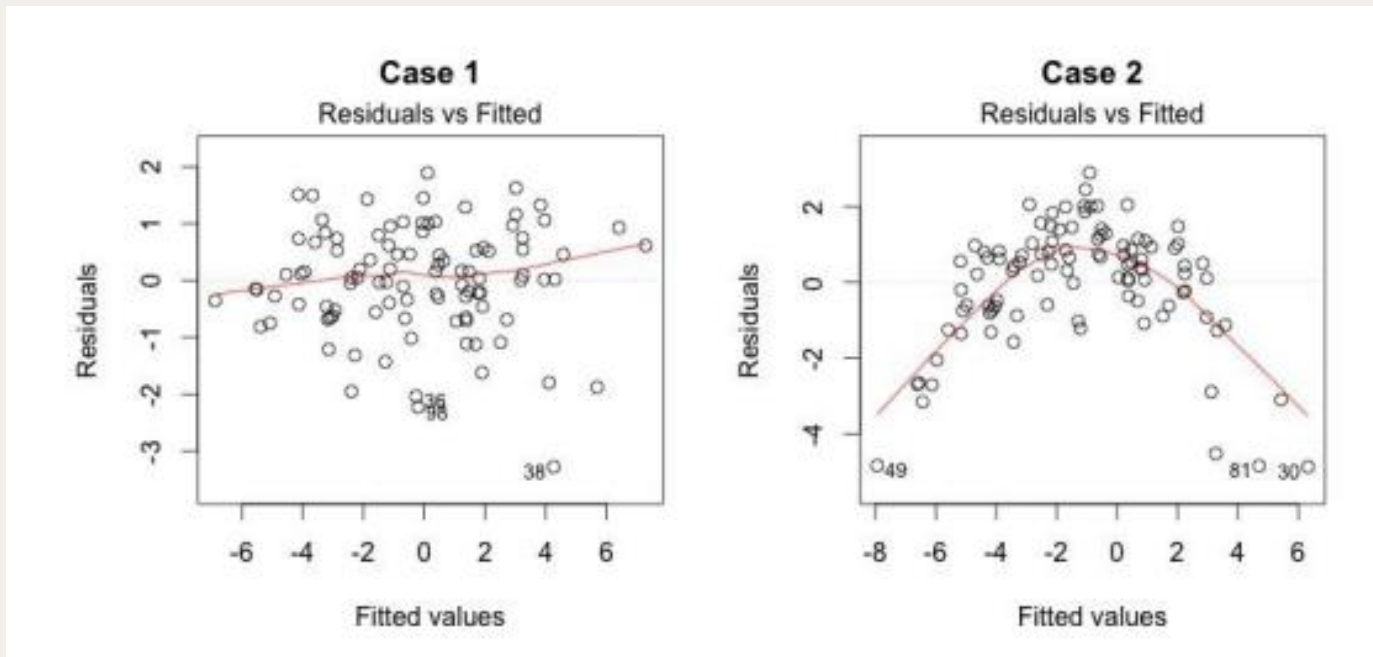
ANÁLISE DOS RESÍDUOS

- Para verificar se um modelo é adequado, temos que investigar se as suposições feitas para o desenvolvimento do modelo estão satisfeitas. Para tanto, estudamos o comportamento do modelo usando o conjunto de dados observados, notadamente as discrepâncias entre os valores observados e os valores ajustados pelo modelo, ou seja, fazemos uma análise dos resíduos.
- O i -ésimo resíduo é a diferença entre o valor observado Y_i e o valor estimado correspondente \hat{Y}_i .
- O resíduo é denotado por e_i e é definido como:

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

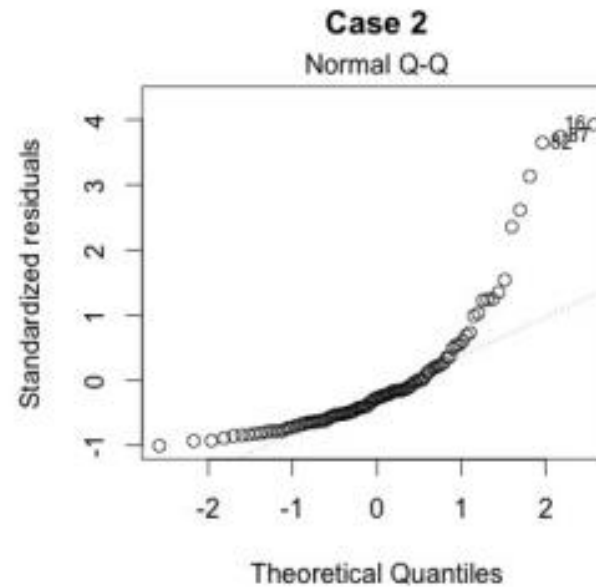
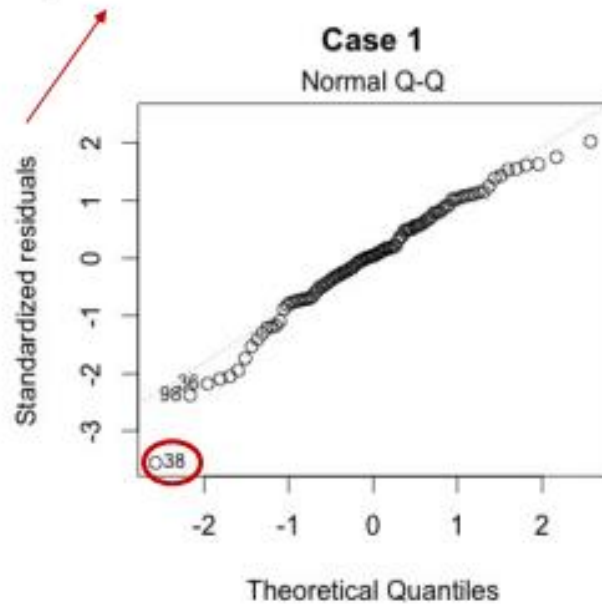
ANÁLISE DOS RESÍDUOS



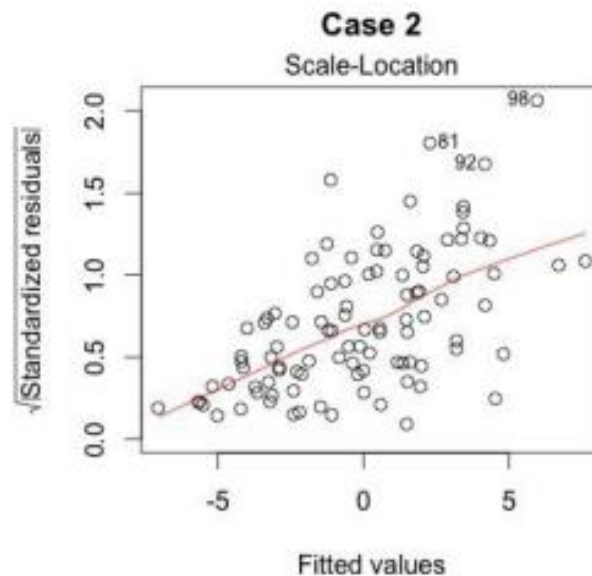
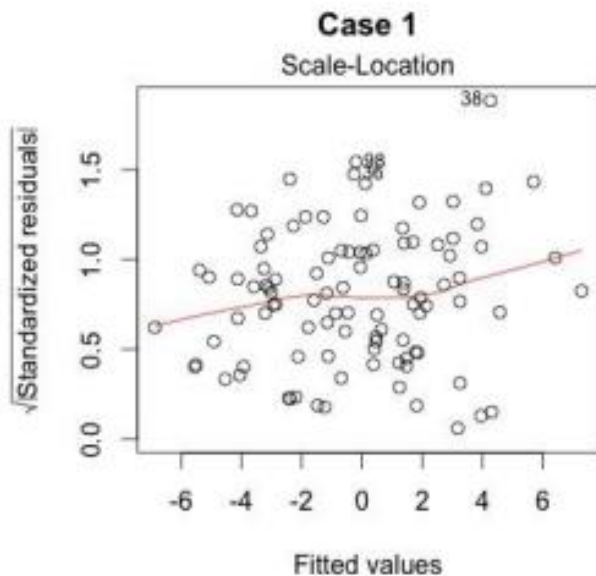


- Este gráfico mostra se os resíduos têm padrões não lineares. Pode haver uma relação não linear entre as variáveis explicativas e a variável resposta, e o padrão pode aparecer nesse gráfico se o modelo não capturar o relacionamento não linear.
- Resíduos igualmente espalhados ao redor de uma linha horizontal sem padrões distintos é uma boa indicação de que não há relacionamentos não lineares.

Resíduo
Desvio padrão dos resíduos

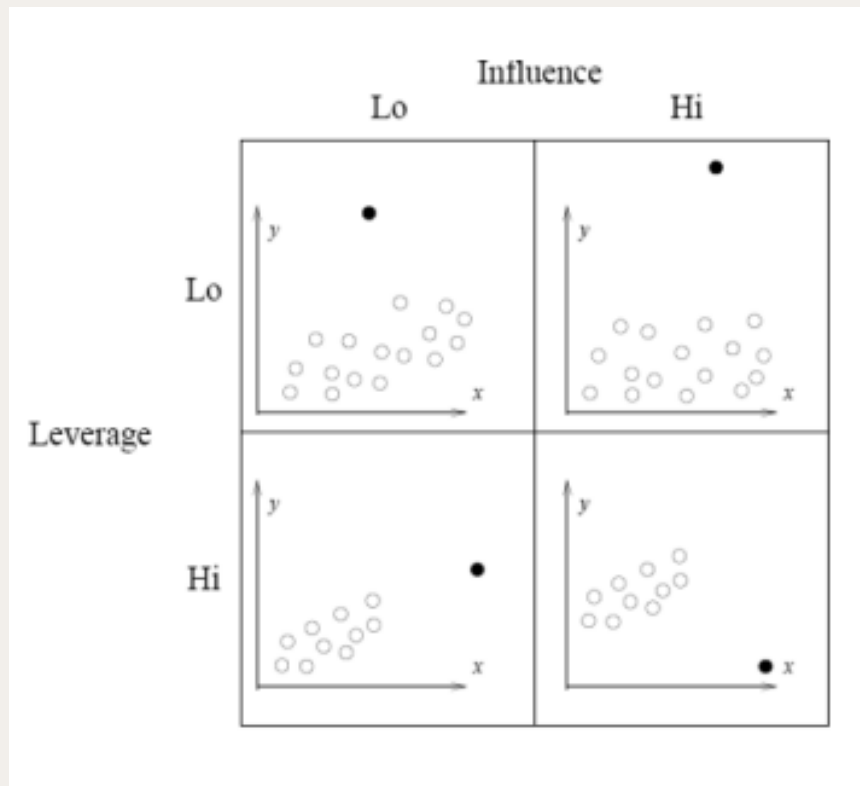


- Este gráfico mostra se os resíduos são normalmente distribuídos.
- Os resíduos seguem bem uma linha reta ou desviam-se severamente?
- O ideal é que os resíduos estejam bem alinhados na linha tracejada.



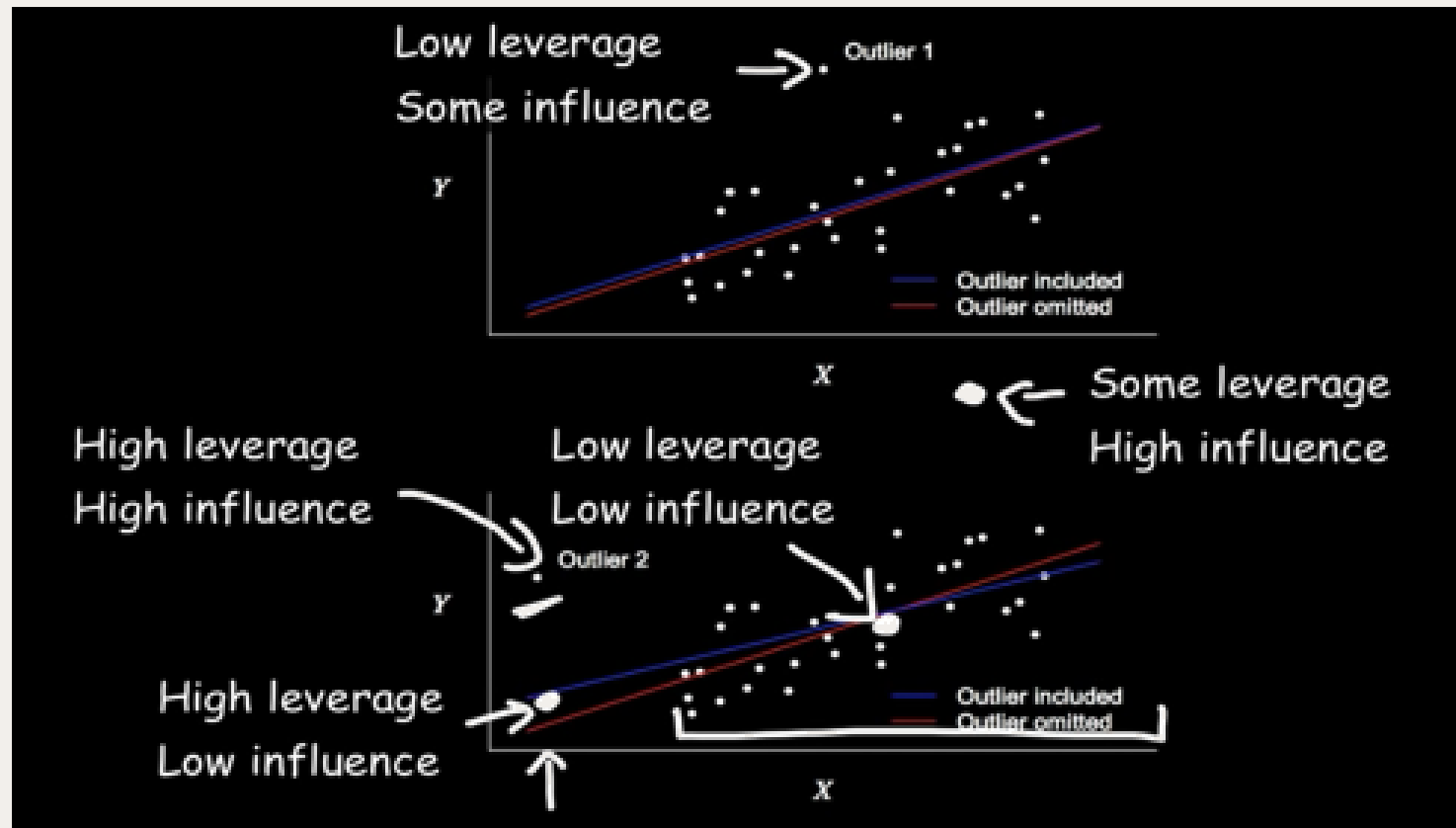
- Este gráfico mostra se os resíduos (com média zero e variância um) são distribuídos igualmente ao longo dos intervalos da variável explicativa.
- Este gráfico elimina o sinal dos resíduos. Grandes resíduos (ambos positivos e negativos) são plotados no topo e pequenos resíduos na parte inferior.
- Útil para verificar a suposição homocedasticidade.
- O ideal é encontrar uma linha horizontal com pontos aleatoriamente distribuídos.

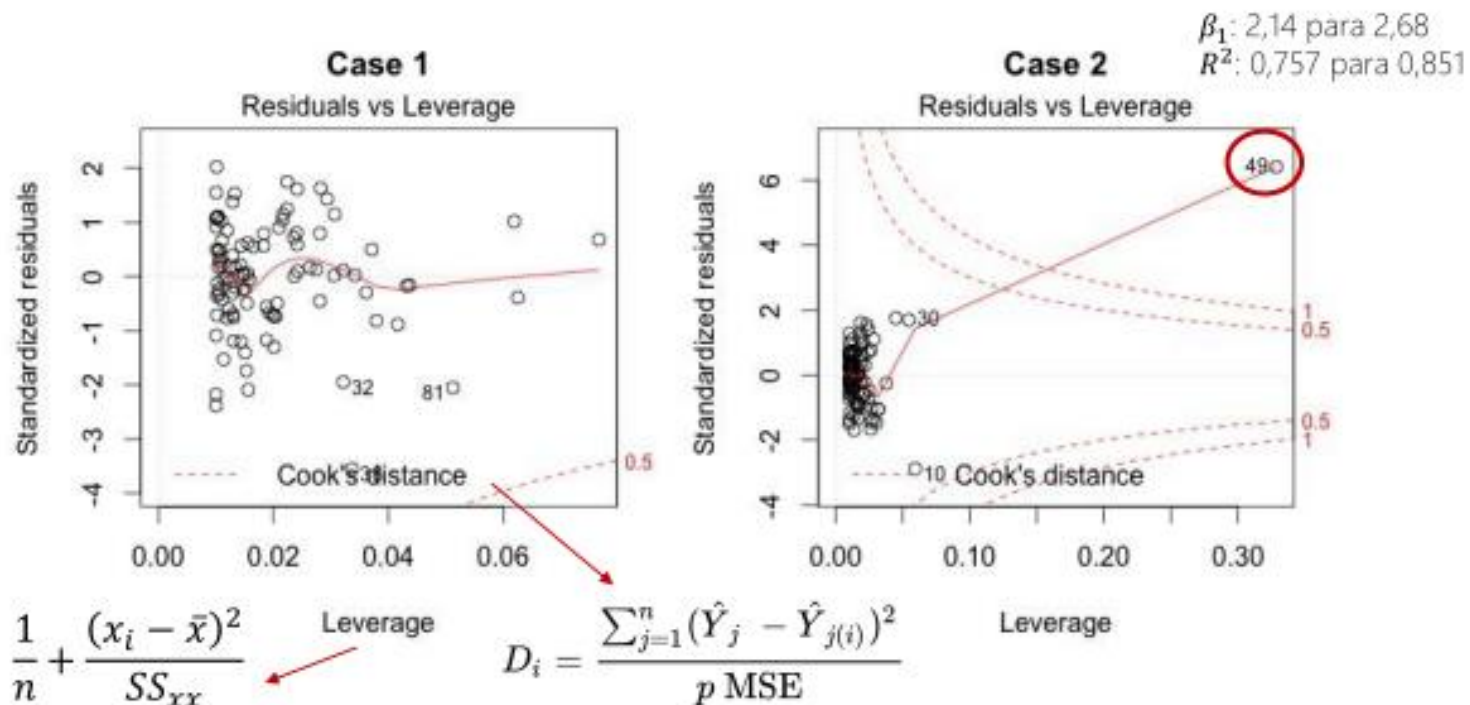
PONTOS DE ALAVANCAGEM E INFLUÊNCIA



- Outlier: observação que tem um grande resíduo.
- Pontos de alavancagem: observação que tem um valor de x longe da média de x .
- Observações influentes: observação que altera a inclinação da linha.
- Assim, pontos influentes têm uma grande influência no ajuste do modelo. Um método para encontrar pontos influentes é comparar o ajuste do modelo com e sem cada observação.

PONTOS DE ALAVANGAGEM E INFLUÊNCIA





- Este gráfico nos ajuda a encontrar casos influentes, se houver.
- Observamos os valores periféricos no canto superior direito ou no canto inferior direito. Esses pontos são os lugares onde os casos podem influenciar uma linha de regressão.
- Quando os casos estão fora da distância de Cook, os casos são influentes para os resultados da regressão. Os resultados da regressão serão alterados se excluirmos esses casos.

VOLTANDO AO NOSSO EXEMPLO NO



```
cholesterol <- c(3.5, 1.9, 4.0, 2.6, 4.5, 3.0, 2.9, 3.8, 2.1, 3.8, 4.1, 3.0, 2.5,
                4.6, 3.2, 4.2, 2.3, 4.0, 4.3, 3.9, 3.3, 3.2, 2.5, 3.3)
age <- c(46, 20, 52, 30, 57, 25, 28, 36, 22, 43, 57, 33, 22, 63, 40, 48, 28, 49,
        52, 58, 29, 34, 24, 50)
data <- as.data.frame(cbind(age, cholesterol))
model <- lm(cholesterol~age)
summary(model)
```

```
##
## Call:
## lm(formula = cholesterol ~ age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.6111 -0.2151 -0.0058  0.2297  0.6256
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.279868   0.215699   5.934 5.69e-06 ***
## age          0.052625   0.005192  10.136 9.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.334 on 22 degrees of freedom
## Multiple R-squared:  0.8236, Adjusted R-squared:  0.8156
## F-statistic: 102.7 on 1 and 22 DF, p-value: 9.428e-10
```

R^2 ajustado: comparar modelos que têm diferentes números de preditores. Ele incorpora o número de preditores no modelo.

VOLTANDO AO NOSSO EXEMPLO NO



```
cholesterol <- c(3.5, 1.9, 4.0, 2.6, 4.5, 3.0, 2.9, 3.8, 2.1, 3.8, 4.1, 3.0, 2.5,  
                4.6, 3.2, 4.2, 2.3, 4.0, 4.3, 3.9, 3.3, 3.2, 2.5, 3.3)  
age <- c(46, 20, 52, 30, 57, 25, 28, 36, 22, 43, 57, 33, 22, 63, 40, 48, 28, 49,  
        52, 58, 29, 34, 24, 50)  
data <- as.data.frame(cbind(age, cholesterol))  
model <- lm(cholesterol ~ age)  
summary(model)
```

```
##  
## Call:  
## lm(formula = cholesterol ~ age)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -0.6111 -0.2151 -0.0058  0.2297  0.6256   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  1.279868   0.215699   5.934 5.69e-06 ***  
## age          0.052625   0.005192  10.136 9.43e-10 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.334 on 22 degrees of freedom  
## Multiple R-squared:  0.8236, Adjusted R-squared:  0.8156   
## F-statistic: 102.7 on 1 and 22 DF, p-value: 9.428e-10
```

```
anova(model)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: cholesterol
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)      
## age         1 11.4648  11.4648   102.75 9.428e-10 ***
```

```
## Residuals  22  2.4548   0.1116
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

UMA DICA!

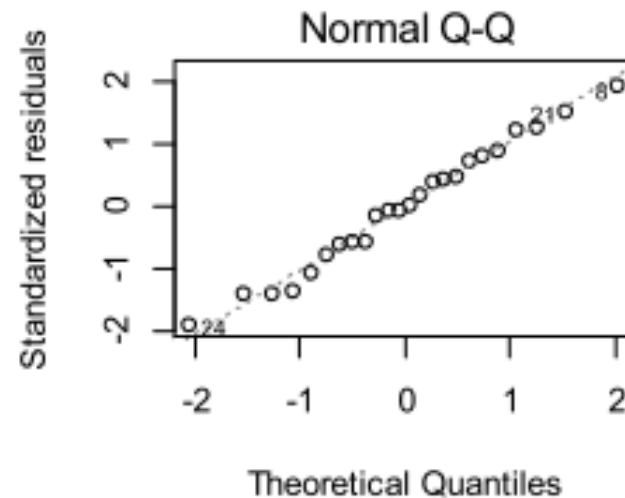
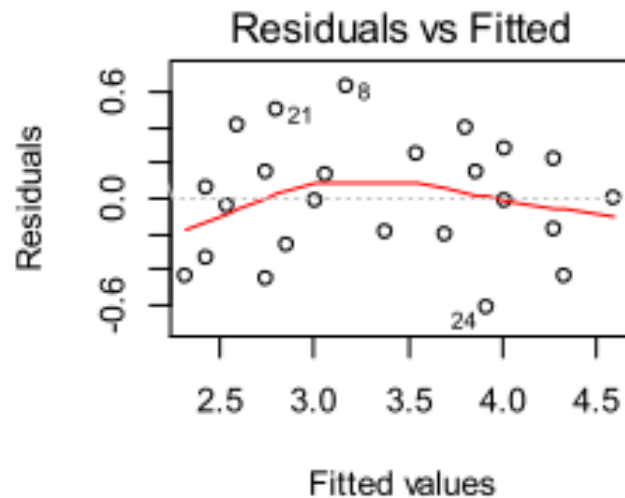


Como o valor-p é uma função do coeficiente de determinação e do tamanho da amostra, você não deve usar o valor-p como uma medida da força da associação.

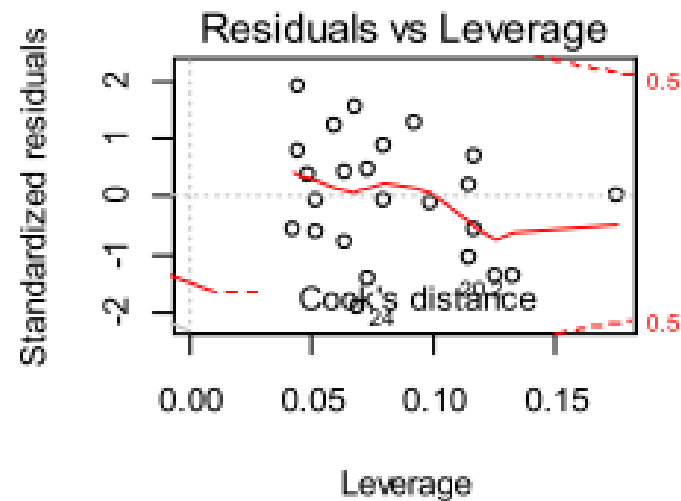
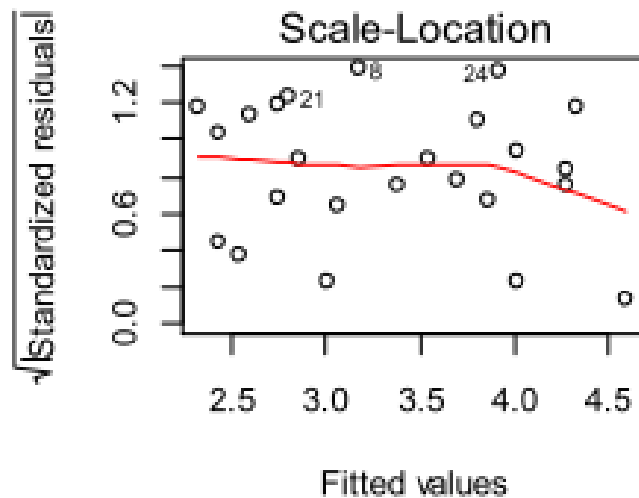
Se a correlação de A e B tem um valor-p menor que a correlação de A e C, isso não significa necessariamente que A e B tenham uma associação mais forte; pode ser que o conjunto de dados para o experimento A-B seja maior. Se você deseja comparar a força da associação de diferentes conjuntos de dados, use o coeficiente de correlação ou de determinação.

ANÁLISE DE RESÍDUOS

lm(cholesterol ~ age)



ANÁLISE DE RESÍDUOS



RESUMINDO!

- Quando queremos verificar se duas variáveis de medição estão correlacionadas uma à outra (se uma variável aumenta, a outra tende a aumentar (ou diminuir)): teste de correlação com o valor-p.
- Quando queremos estimar a força do relacionamento entre duas variáveis: coeficiente de correlação.
- Quando queremos encontrar a equação de uma linha que se encaixa na nuvem de pontos: equação da regressão linear. Você pode usar esta equação para previsão.

E LEMBREM-SE: CORRELAÇÃO NÃO IMPLICA CAUSALIDADE, MAS INDICA QUE ALGO INTERESSANTE ESTÁ ACONTECENDO.

<https://www.tylervigen.com/spurious-correlations>

REFERÊNCIAS



- RAWLINGS, John O.; PANTULA, Sastry G.; DICKEY, David A. **Applied regression analysis: a research tool**. Springer Science & Business Media, 2001.
- MORETTIN, Pedro; BUSSAD, Wilton. **Estatística Básica**, 6. ed.
- MCDONALD, John H. **Handbook of biological statistics**. Baltimore, MD: sparky house publishing, 2009.
- <https://onlinecourses.science.psu.edu/stat501>, acessado em 05 de abril de 2020.