Métodos Estatísticos Avançados em Epidemiologia

Regressão Linear Simples e Múltipla

Referência: Caps. 17-19 - Pagano e Gauvreau (2004)

Enrico A. Colosimo/UFMG

Depto. Estatística - ICEx - UFMG
http://www.est.ufmg.br/~enricoc

Introdução

- Avaliar uma possível **associação/correlação** entre características quantitativas é objetivo de vários estudos em biologia e ciências da saúde.

- Em geral, para estudar a associação entre duas variáveis, uma *amostra aleatória* é selecionada e **as duas variáveis são observadas simultaneamente** para cada paciente/indivíduo.

- **Exemplo:** Vinte e cinco pacientes são atendidos em uma clínica oftalmológica e as medidas são registradas para *pressão intra-ocular* (PIO) e *idade*.

Correlação - Estudo Transversal

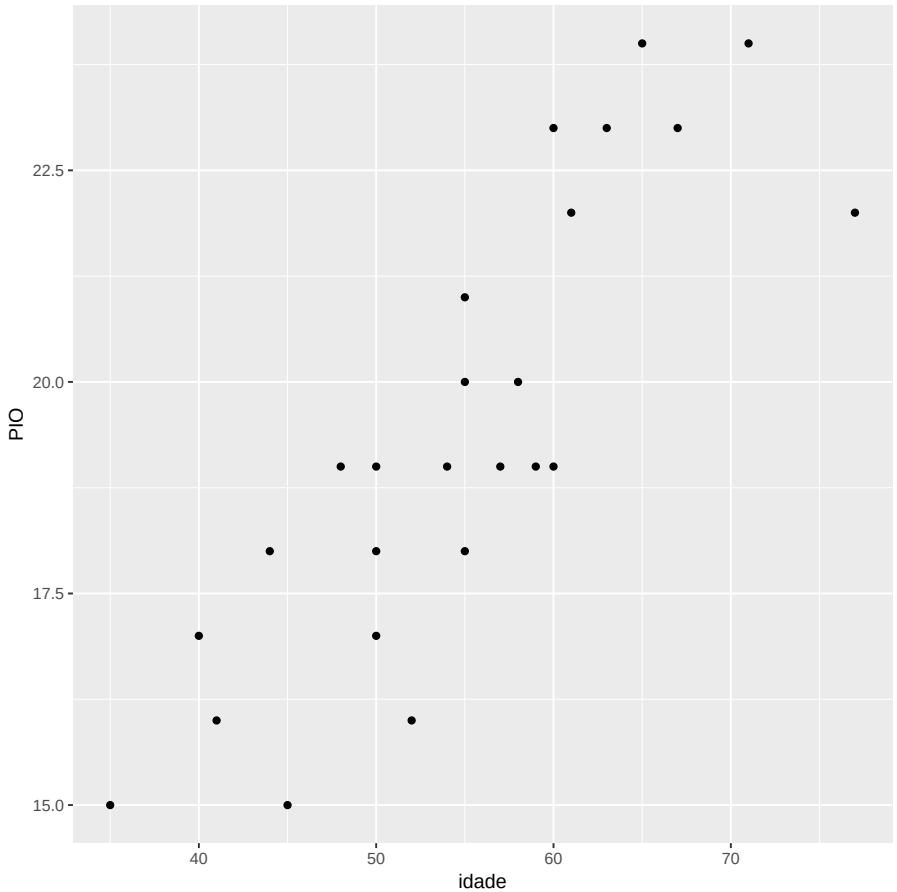
Exemplo - Tabela de dados brutos

Indivíduo	Idade	PIO	Indivíduo	Idade	PIO
1	35	15	14	55	20
2	40	17	15	57	19
3	41	16	16	58	20
4	44	18	17	59	19
5	45	15	18	60	23
6	48	19	19	60	19
7	50	19	20	61	22
8	50	18	21	63	23
9	50	17	22	65	24
10	52	16	23	67	23
11	54	19	24	71	24
12	55	18	25	77	22
13	55	21			

Correlação

Exemplo - Gráfico de dispersão

- Uma maneira de descrever os dados conjuntamente é através do *gráfico de dispersão* (também conhecido como *scatter plot*).



Correlação
Exemplo

- O gráfico de dispersão indica uma possível correlação positiva entre as variáveis.
- Ou seja, os mais velhos estão associados com valores de pressão intra-ocular mais altos.
- **Perguntas de interesse:**
  - Esta relação é real (significativa)?
  - Qual o tipo/forma da relação entre as variáveis *Idade* e *PIO*?
  - Qual a força desta relação?

Correlação

Tipos de relação

- A relação entre duas variáveis quantitativas pode ser descrito como *linear* (*positiva* ou *negativa*) e *não-linear*.

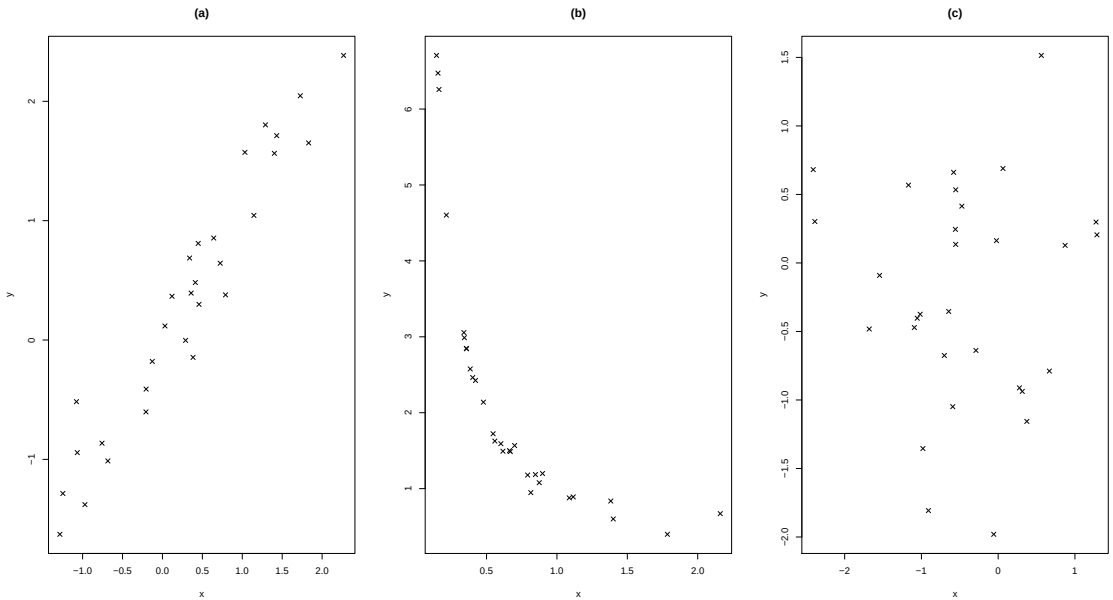


Figura: Tipos de relação entre x e y: (a) relação linear (*linha reta*) positiva; (b) relação não-linear negativa; (c) sem relação.

Correlação
Força da relação

- A “força" da relação se refere à proximidade dos pontos e a curva ou a linha reta subjacente.

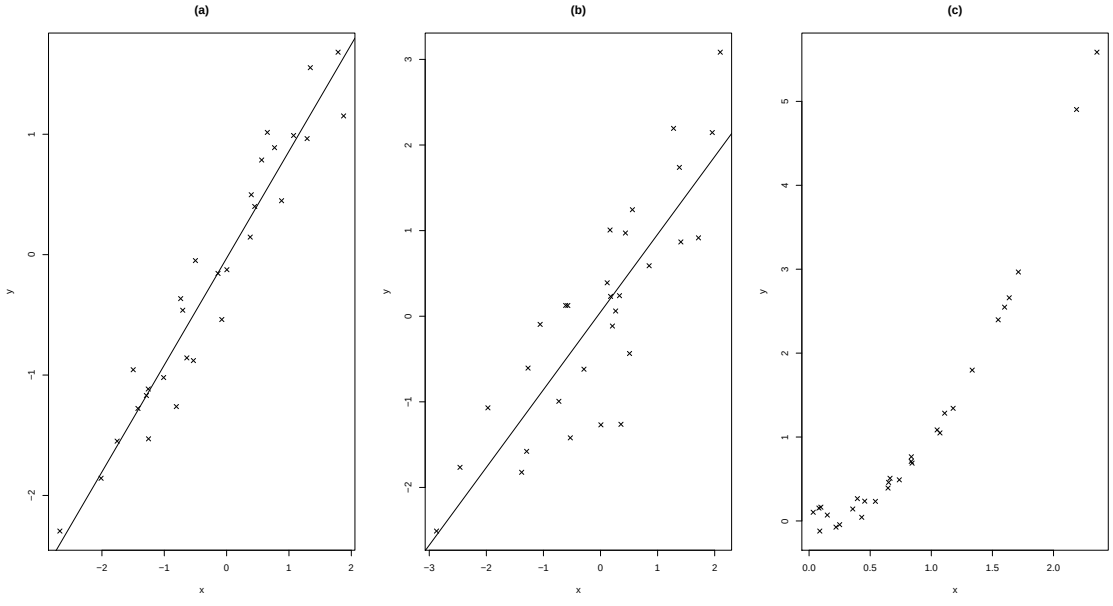


Figura: (a) relação linear positiva forte; (b) relação linear positiva fraca; (c) relação não-linear forte.

Coeficiente de correlação

- A força de uma associação (linear) pode ser medida por um **coeficiente de correlação**.
- O *coeficiente de correlação de Pearson* é o mais utilizado para medir a intensidade da associação entre duas variáveis quantitativas.
- O coeficiente de correlação de Pearson é representado por ρ e sua estimativa por r.
- O coeficiente de correlação de Spearman é uma solução alternativa (não-paramétrica).

Coeficiente de correlação

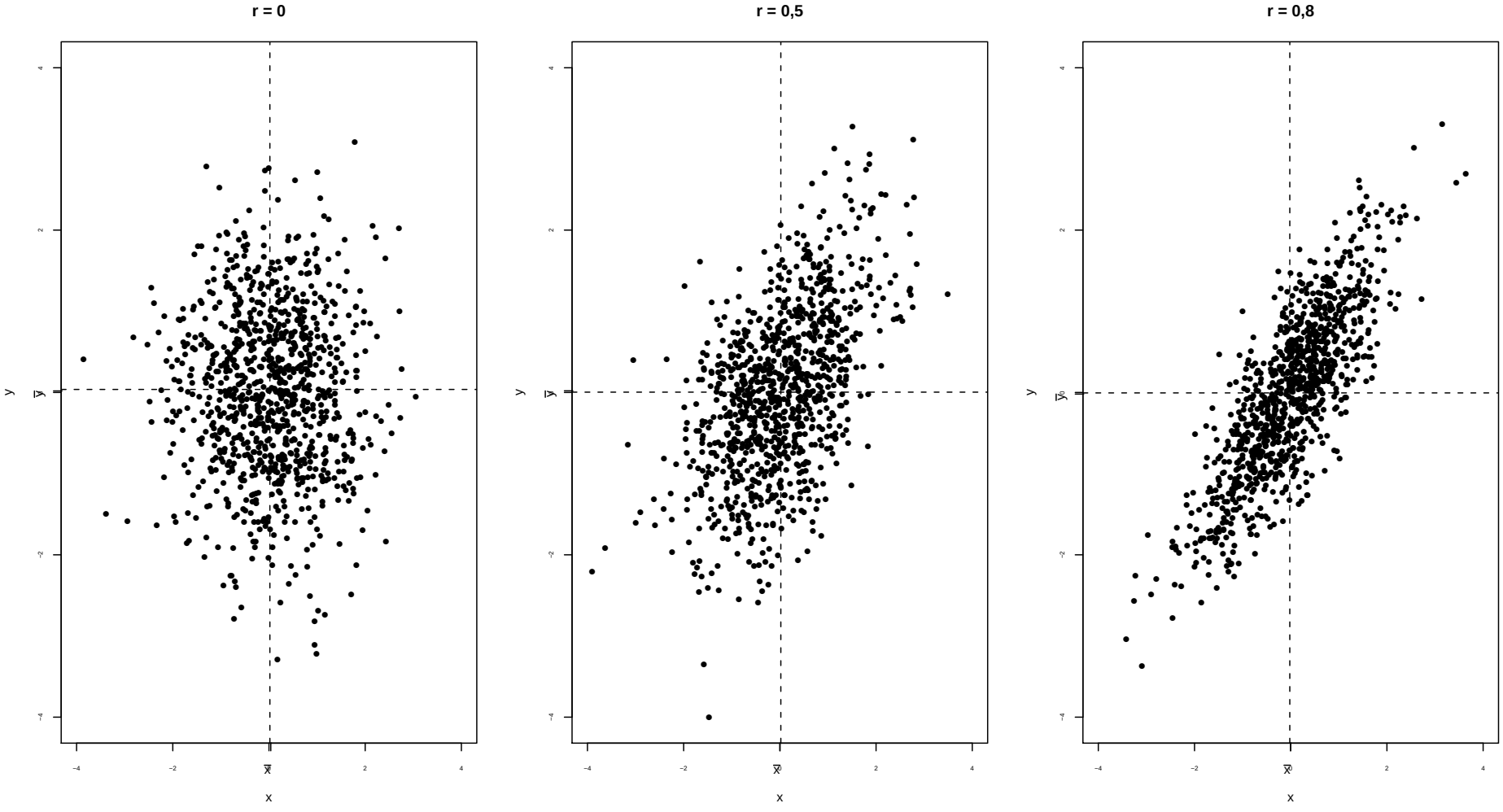
- O coeficiente de correlação pode variar entre -1 e 1.
- Valores negativos indicam uma correlação do tipo inversa (*negativa*):
  - Quando x aumenta, y diminui (ou vice-versa).
- Valores positivos ocorrem quando a correlação é direta (*positiva*):
  - x e y variam no mesmo sentido.
- No exemplo da clínica oftamológica, idade e pressão intra-ocular indicam uma possível correlação positiva (ainda não medimos a força desta correlação).

Coeficiente de correlação

- O valor máximo (tanto $\rho = +1$ como $\rho = -1$) é obtido quando todos os pontos do diagrama de dispersão estão em uma linha reta (correlação perfeita).
- Quando não existe correlação entre x e y, os pontos se distribuem em nuvens circulares ($\rho = 0$).
- Quando os pontos formam uma nuvem cujo eixo principal não é uma reta (não-linear), o coeficiente de correlação não mede adequadamente a associação entre as variáveis.

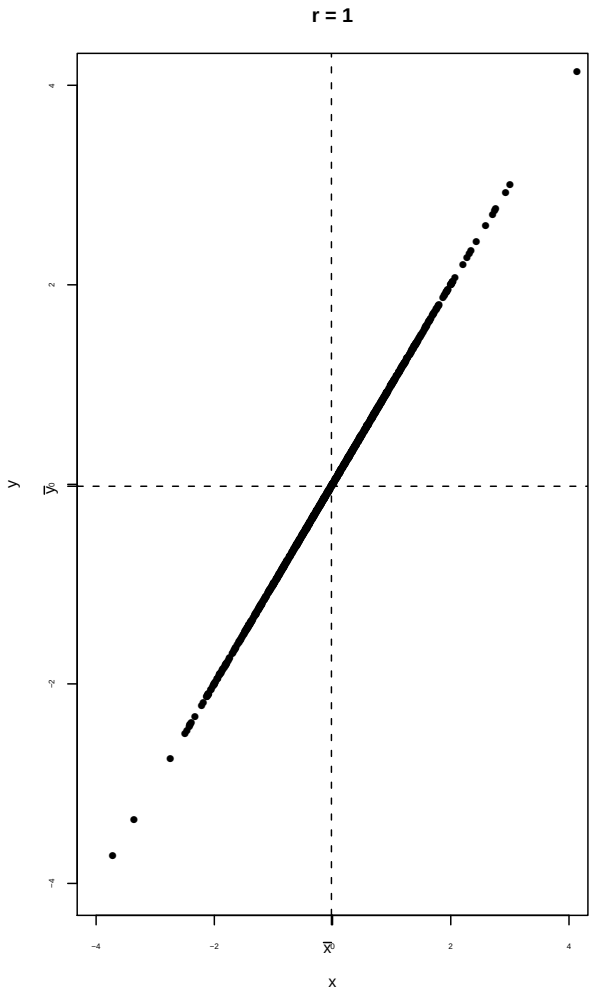
Coeficiente de correlação

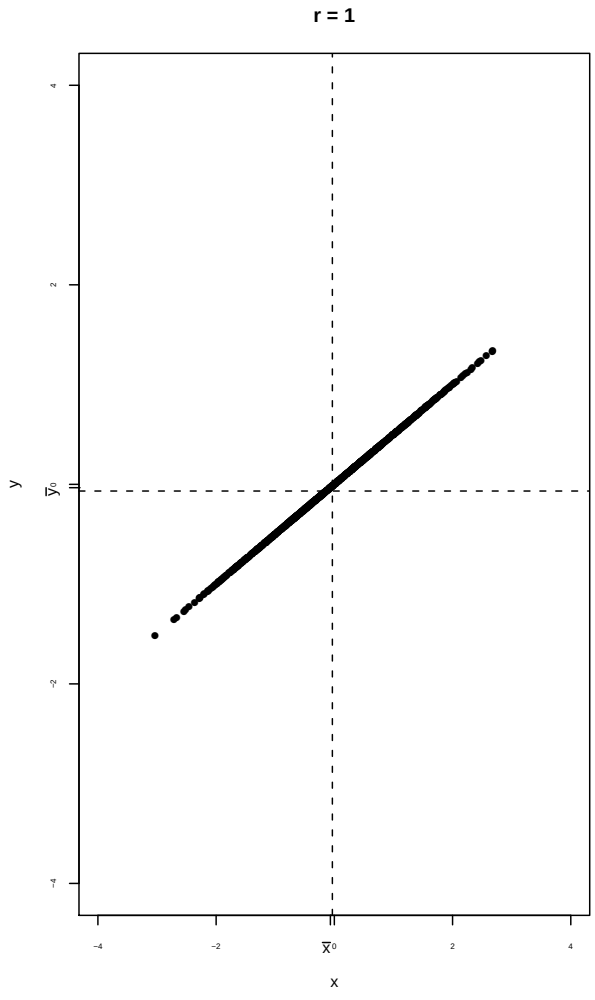
Exemplos

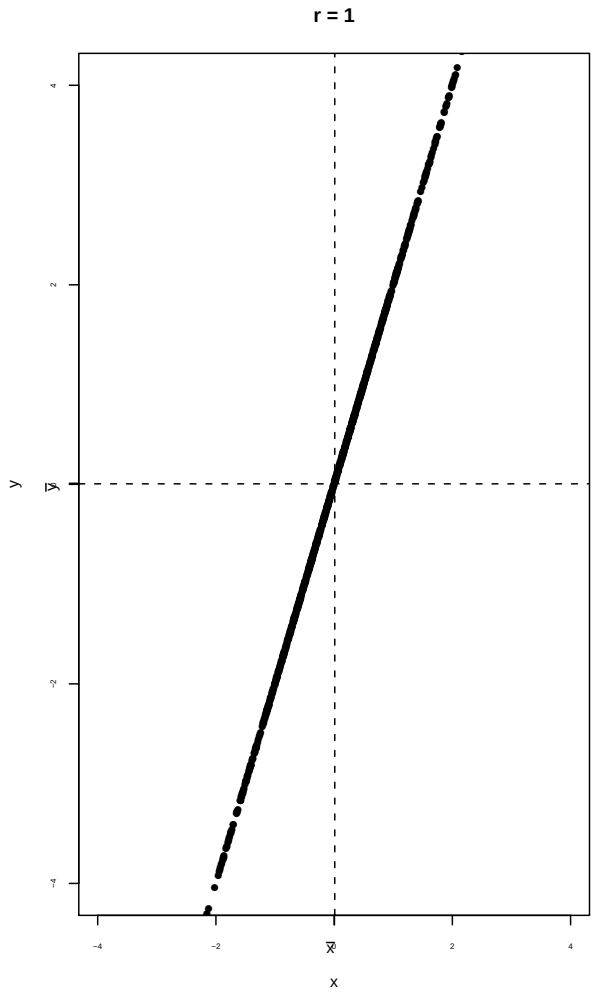


Coeficiente de correlação

Exemplos: Correlação Perfeita

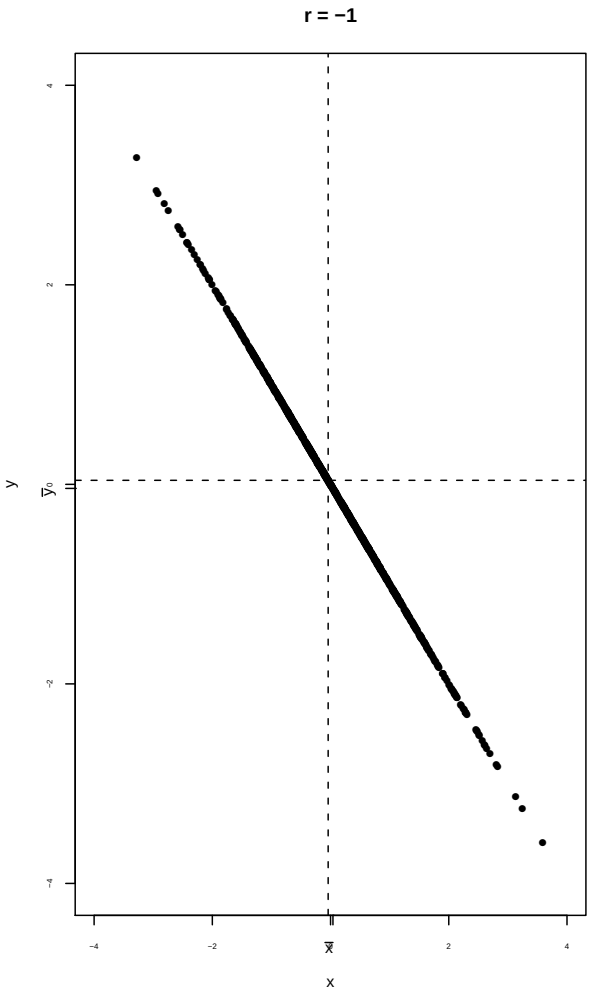


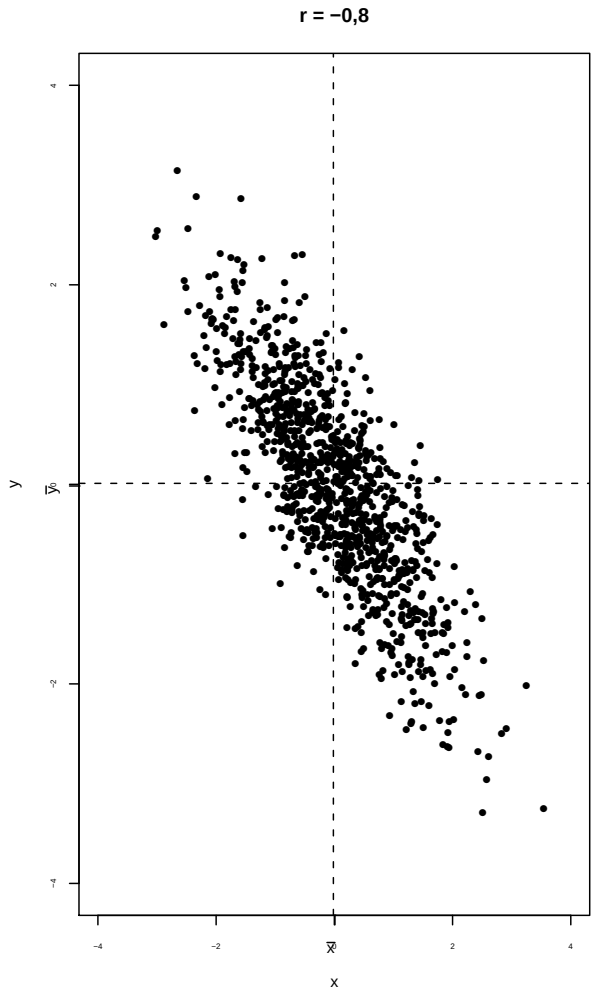


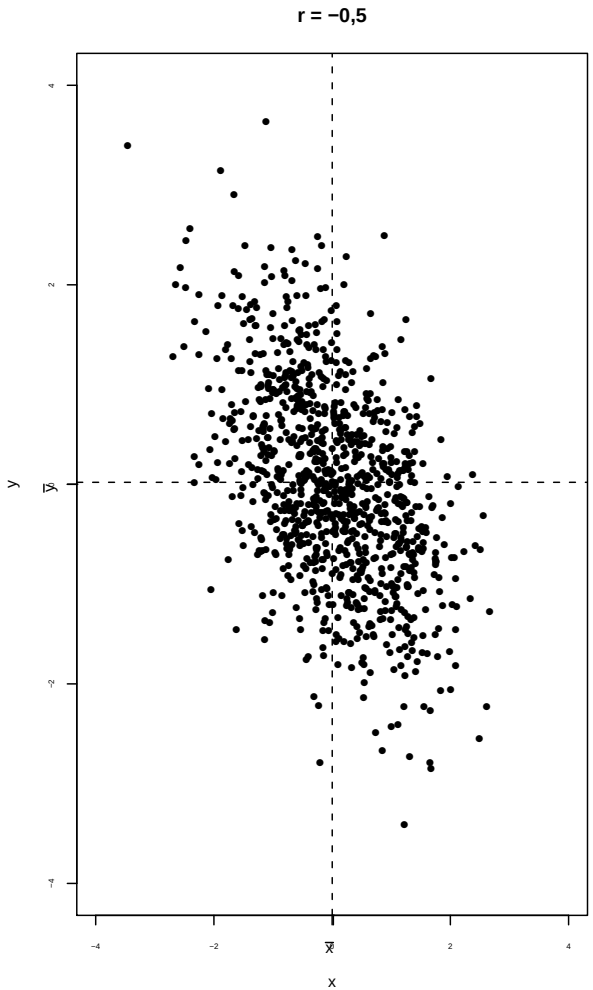


Coeficiente de correlação

Exemplos







Coeficiente de correlação

Cálculo do coeficiente de correlação

- O estimador do coeficiente de correlação de Pearson para uma amostra de tamanho n é

$$r = \frac{cov_{xy}}{s_x \times s_y},$$

- $cov_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/(n-1)$ é a covariância de x e y,
-

$$cov_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x} \times \overline{y}}{n-1}$$

- s_x e s_y são os desvios-padrão de x e y, respectivamente

Coeficiente de correlação

Cálculo do coeficiente de correlação

- No exemplo da clínica oftamológica:
  - X: idade, $\overline{x} = 54,9$ anos e $s_x = 9,9$ anos.
  - Y: pio, $\overline{y} = 19,4$ mmHg e $s_x = 2,7$ mmHg.
  - $\sum_{i=1}^{n} x_i y_i = 27217$
- $$r = \frac{(27217 - (25 \times 54,9 \times 19,4))/24}{9,9 \times 2,7} = 0,84.$$

Coeficiente de correlação

Teste de significância

- A estimativa r é um valor para uma certa amostra.
- É necessário ter a significância do coeficiente de correlação. Ou seja, testar a hipótese: $H_0 : \rho = 0$.

Estatística de teste

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

sob H_0, tem distribuição t, com $n-2$ graus de liberdade (sob certas suposições).

- **Suposições** para realizar o teste de hipóteses:
  - Tanto a variável *x* quanto a variável *y* têm distribuição normal.
  - A relação entre *x* e *y* é linear.
- **No exemplo:** ao nível de 5% de significância, $t_{0,05;23} = 2,069$

$$t = 0,84\sqrt{\frac{25-2}{1-0,84^2}} = 7,57$$

valor-p $<$ 0,001.

- Ao nível de 5% há fortes evidências de que ρ é diferente de zero.

Coeficiente de correlação

Intervalo de Confiança para ρ

O IC para ρ é baseado na transformação de Fisher.

- A Transformação de Fisher é a seguinte:

$$\eta = 0,5 \log\left(\frac{1+\rho}{1-\rho}\right)$$

  η é um número real.
- Construímos o Int. de 95% Confiança para η

$$\widehat{\eta} \pm 1,96 \times \sqrt{1/(n-3)}$$

- Finalmente, tomamos a transformação inversa para obter o IC para ρ.

$$\rho = \frac{e^{2\eta}-1}{e^{2\eta}+1}$$

- O Intervalo de 95% Confiança para ρ é $(0,68; 0,93)$.

Coeficiente de correlação

Força da correlação

- Uma vez determinada a existência de associação entre as variáveis, pode-se avaliá-la qualitativamente quanto à intensidade. Uma proposta de avaliação é a seguinte.

$\lvert\rho\rvert$	A correlação é dita
0	Nula
$0 \mid -0,3$	Fraca
$0,3 \mid -0,6$	Regular
$0,6 \mid -0,9$	Forte
$0,9 \mid -1$	Muito forte
1	Plena ou perfeita

Coeficiente de correlação

Cuidado: Associações Espúrias

- Podemos obter estimativas para a correlação para qualquer par de variáveis, mas sempre devemos ter o devido cuidado em sua interpretação.
- Isto é, obtemos uma medida de associação que não é necessariamente uma relação de causa e efeito.
- Por exemplo, existe uma forte relação positiva linear entre o número de médicos e o número de mortes por cidade/ano!
- À primeira vista, podemos ser tentados a concluir que ter mais médicos conduz a mais mortes!!!!
- Ou seja, podemos ter associações espúrias.

Regressão linear simples

- O estudo da **regressão** aplica-se àquelas situações em que há razões para supor uma relação assimétrica entre duas variáveis.
- Regressão linear simples é similar a correlação, em que deseja-se expressar matematicamente a relação entre as variáveis.
- Geralmente chama-se a *variável resposta* (desfecho) de Y e a *covariável* (variável explicativa) de x.
- O termo **regressão** deve-se a **Sir Francis Galton**, que publicou, em 1886, um artigo no qual tentou explicar por que pais de alta estatura tinham filhos com estatura em média mais baixa do que a deles e pais de baixa estatura tinham filhos em média mais altos.
- Galton chamou esse fenômeno de “regressão à média".

Regressão linear simples

- Os objetivos da análise de regressão são:
  1. Avaliar uma possível associação/dependência de Y em relação a x.
  2. Expressar matematicamente esta relação por meio de uma equação linear.
- Vária são as relações de associação que podem ser descritas ou aproximadas por linhas retas.
- A análise de regressão linear simples é um procedimento que fornece equações de linhas retas (por isso, o termo “linear"), que descrevem fenômenos em que há apenas uma variável explicativa (por isso, “simples").

- O modelo de regressão linear pode ser escrito como:

$$Y = \alpha + \beta x + \epsilon,$$

em que

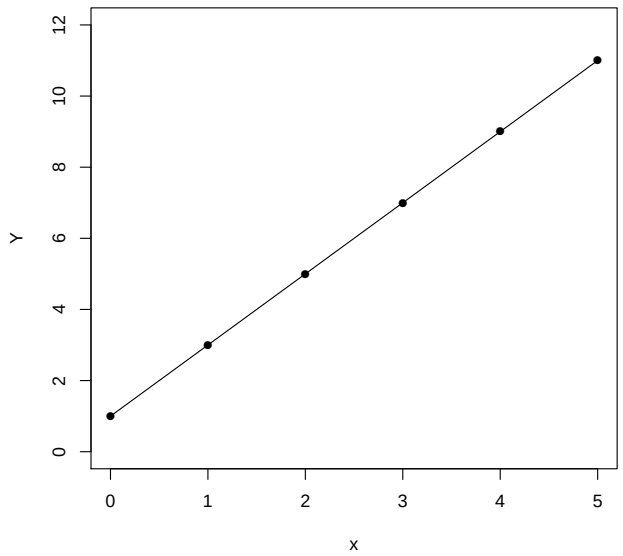
  - Y: variável resposta (desfecho);
  - α: coeficiente linear ou intercepto;
  - β: coeficiente angular ou inclinação da reta (acréscimo ou decréscimo em Y para cada acréscimo de uma unidade em x)
  - x: variável explicativa ou covariável.
  - ϵ: erro aleatório.

- Outra representação:

$$\text{média}\,(Y|X = x) = \alpha + \beta x$$

- **Exemplo:** Suponha que o número de funcionários (Y) em uma empresa é completamente determinado pelo número de clientes (x), da seguinte forma:

$$Y = 1 + 2x,$$



Isto significa que: (1) independente do número de clientes a empresa precisa de um funcionário; (2) taxa de crescimento do número de funcionários é de 2 para cada cliente.

- Na prática não temos uma reta "perfeita". Isto significa, que não existe uma reta que passa "exatamente"sobre todos os valores amostrais.
- Ou seja, teremos parte da dispersão da resposta medida pela reta de regressão e o restante irá para o erro aleatório.
- A partir de uma amostra de dados, devemos obter estimativas $\hat{\alpha}$ e $\hat{\beta}$ para α e β.
- O *Método dos Mínimos Quadrados* é o mais utilizado para obter as estimativas. Este método obtem a reta com as menores distâncias entre os valores observados (y) e a própria reta.

Para uma amostra de tamanho n, encontrar os valores de α e β que minimizam a distância entre Y (valor observado) e $\alpha + \beta x$ (valor predito pela reta de regressão). Ou seja, minimizar a seguinte função:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

- O coeficiente β é estimado da seguinte maneira:

$$\hat{\beta} = \frac{cov_{xy}}{s_x^2}, \quad \text{ou} \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- O coeficiente α é estimado por:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- $\bar{y}$ e $\bar{x}$ são as médias amostrais de y e x.

- A reta estimada de regressão é

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x.$$

- **Exemplo:** (Clínica Oftalmológica) Acredita-se que a *pressão intra-ocular* é explicada (depende) da *idade*. Relembrando:

Indivíduo	Idade	PIO	Indivíduo	Idade	PIO
1	35	15	14	55	20
2	40	17	15	57	19
3	41	16	16	58	20
4	44	18	17	59	19
5	45	15	18	60	23
6	48	19	19	60	19
7	50	19	20	61	22
8	50	18	21	63	23
9	50	17	22	65	24
10	52	16	23	67	23
11	54	19	24	71	24
12	55	18	25	77	22
13	55	21			

Regressão linear simples

Obtendo a reta de regressão - Exemplo (Clínica Oftalmológica)

- A seguinte equação é proposta:

$$PIO = \alpha + \beta \times idade.$$

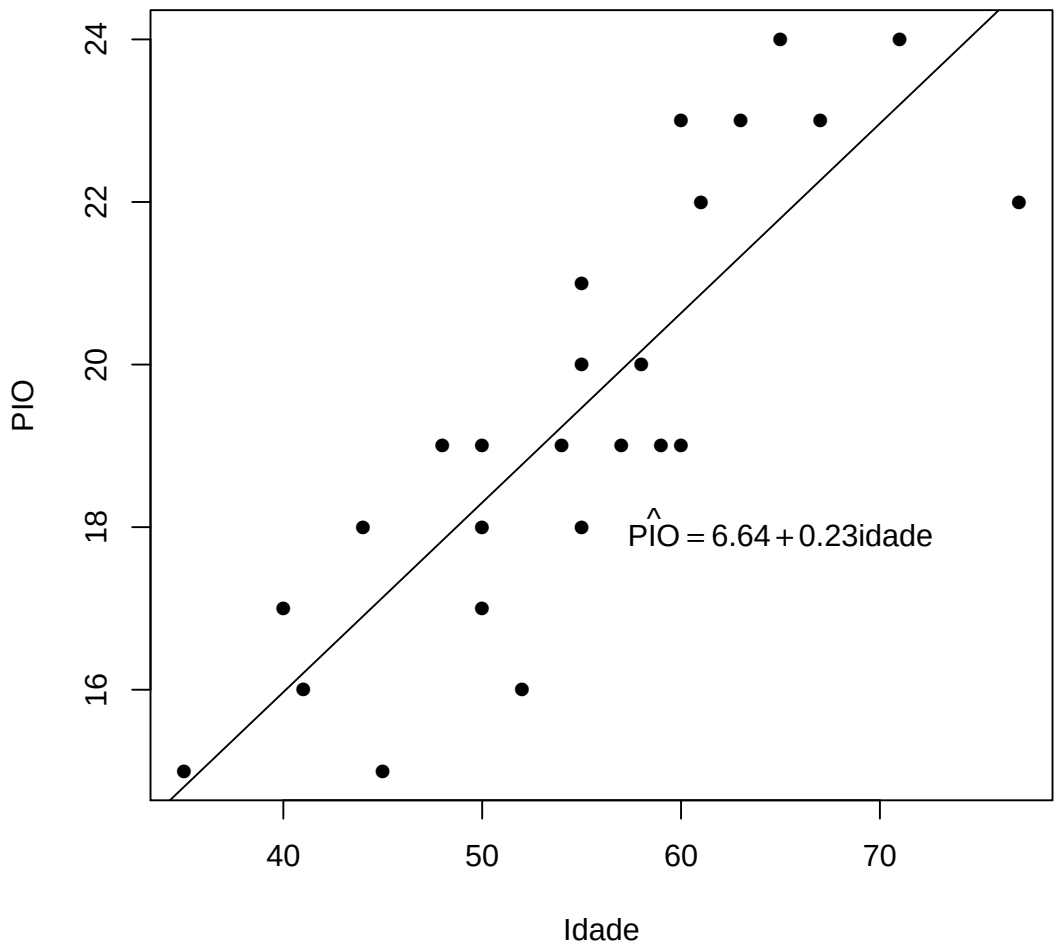
- $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{27217 - (1372)(486)/25}{77634 - 1882384/25} = 0,23$$

- $$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 19,44 - 0,2318 \times 54,88 = 6,64$$

- $$\widehat{PIO} = 6,64 + 0,23 \times idade.$$

Regressão linear simples

Obtendo a reta de regressão - Exemplo (Clínica Oftalmológica)



- A **dependência** de Y em relação a x é representada pelo coeficiente β.
- Para afirmar que $\hat{\beta}$ representa uma associação verdadeira entre Y e x, deve-se realizar um teste de hipóteses.
- O teste de interesse é verificar se x está associado a resposta, o que é equivalente a testar

$$H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0.$$

- Este teste é, exatamente, equivalente ao teste:

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0.$$

Estatística de teste

Para testar $H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0$ a estatística de teste é dada por

$$t = \frac{\hat{\beta}}{\hat{ep}(\hat{\beta})},$$

em que $\hat{ep}(\hat{\beta})$ é uma estimativa para o *erro padrão* do estimador $\hat{\beta}$ e é dado por

$$\hat{ep}_{\hat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{(n-2)\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

t, sob H_0, tem distribuição t de *Student* com $n-2$ graus de liberdade (sob certas suposições). Exatamente o mesmo teste para $H_0 : \rho = 0$.

Regressão linear simples

Utilizando a reta de regressão

A reta de regressão permite:

1. Representar a dependência de uma variável quantitativa (desfecho) em relação à outra (explicativa), que pode ser quantitativa ou qualitativa, por meio de uma equação linear.

2. Predizer valores para a variável resposta Y a partir de valores especificados (inclusive não-observados) da variável x.

    - Isto é permitido dentro da faixa de valores estudados para x.

3. Uma forma de avaliar a qualidade do ajuste é através do coeficiente de determinação $R^2 = r^2$ (quadrado do coeficiente de correlação de Pearson).
    - Exemplo (IOC): $R^2 = 0,845^2 = 0,71$. Isto significa que 71% da variação dos dados está sendo explicada pela reta de regressão.

Regressão linear simples

Saída do R

```
>summary(mod1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7685 -1.3021  0.1956  1.0970  2.3661

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.64318    1.71597   3.871 0.000774 ***
x            0.23318    0.03079   7.572 1.09e-07 ***
---

Residual standard error: 1.489 on 23 degrees of freedom
Multiple R-squared: 0.7137,     Adjusted R-squared: 0.7013
F-statistic: 57.34 on 1 and 23 DF,  p-value: 1.086e-07
```

Regressão linear simples

Requisitos ao uso da regressão linear

Certas exigências devem ser satisfeitas para se realizarem inferências válidas sobre α e β.

1. A variação de Y deve ser a mesma em cada valor de x (homocedasticidade).
2. Os pontos no gráfico de dispersão devem apresentar uma tendência linear (linearidade).
3. Os valores de Y foram obtidos ao acaso da população e são independentes uns dos outros. (independência)
4. A variável Y deve ter distribuição normal (ou aproximadamente normal).
   - Esta exigência não é necessária para x.
5. A variável x deve ser medida sem erro.

Se (1), (2) e (3) forem violadas, transformação em Y é uma forma de contornar o problema.
Estas suposições são verificadas utilizando os resíduos.

- Os resíduos são muito úteis para este propósito.

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i); \quad i = 1, \ldots, n.$$

- Os gráficos apresentados no contexto de ANOVA também são úteis para verificar as suposições básicas do modelo de regressão.

Os gráficos são os seguintes:

1. Homocedasticidade e Linearidade: Gráfico de resíduos (e_i) vs preditos ($\hat{y}_i$);
   - Os pontos devem ficar aleatoriamente distribuídos em torno de uma linha horizontal paralela ao eixo x na altura do resíduo zero.
   - A violação do pressuposto de homocedasticidade produz faixas em forma de cone, enquanto violações do pressuposto de linearidade produz faixas não-lineares.
2. Homocedasticidade (somente): Gráfico de $\sqrt{\text{abs(resíduos)}}$ ($\sqrt{|e_i|}$) vs preditos ($\hat{y}_i$);
   - Os pontos devem ficar aleatoriamente distribuídos em torno de uma linha horizontal paralela ao eixo x mas não do valor zero.

Os gráficos são os seguintes:

3 Normalidade: gráfico de probabilidade dos resíduos.

- Os pontos deve ficar distribuídos ao longo da reta $y = x$

4 Detecção de pontos atípicos: distância de Cook: valores grandes são identificados como pontos atípicos.

1. Testes de Significância:
   - Shapiro-Wilks para os resíduos testa normalidade (H_0: erros normais)
   - Teste de Faraway para homocedasticidade: baseado no gráfico de ($\sqrt{|e_i|}$) vs preditos ($\hat{y}_i$) - realiza uma regressão linear simples destas quantidades. (H_0: erros homocedásticos).

Regressão linear simples

Análise de resíduos

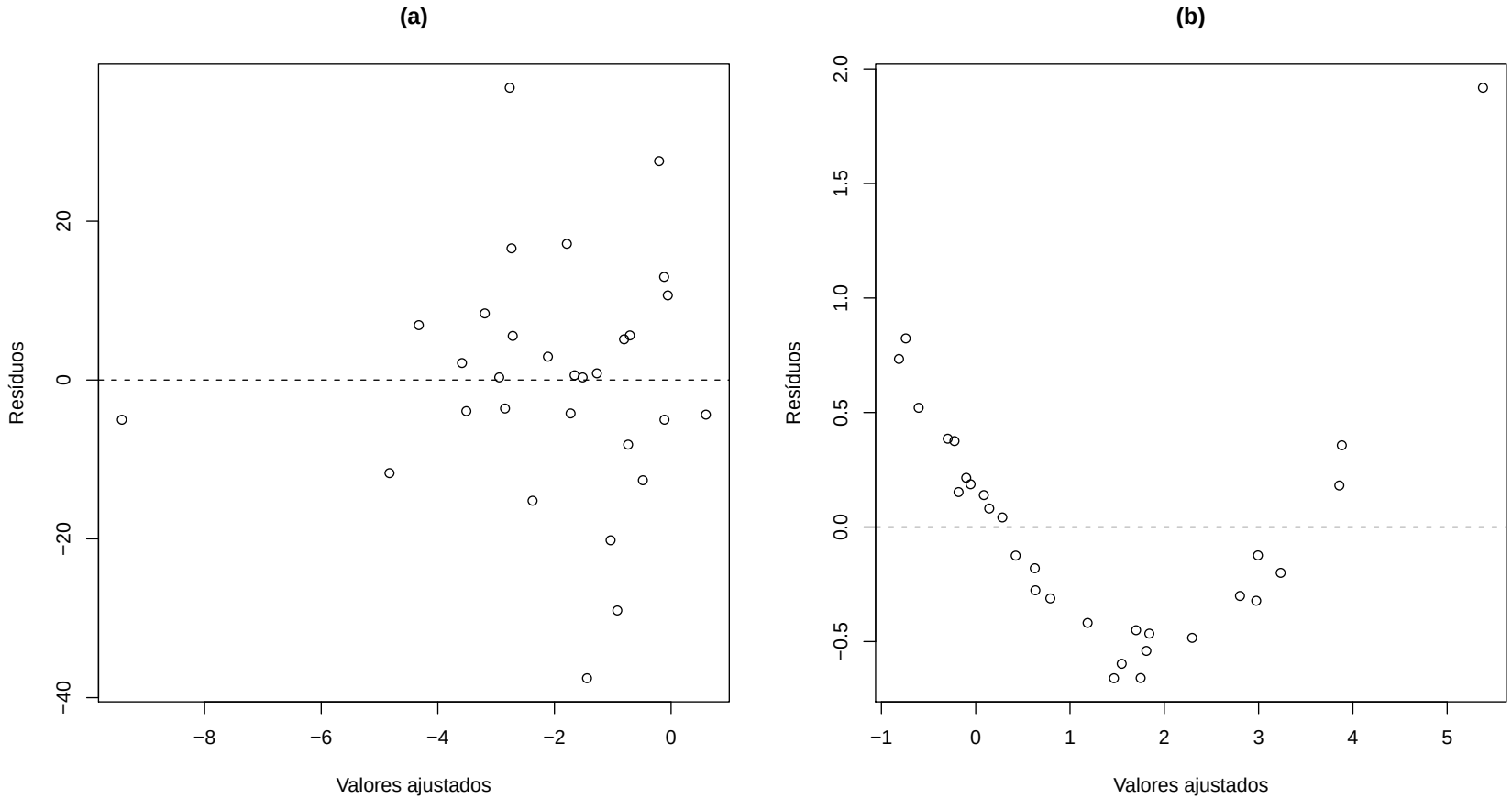
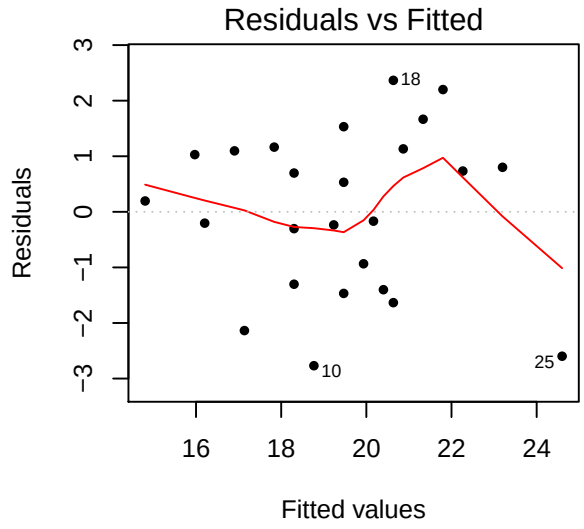
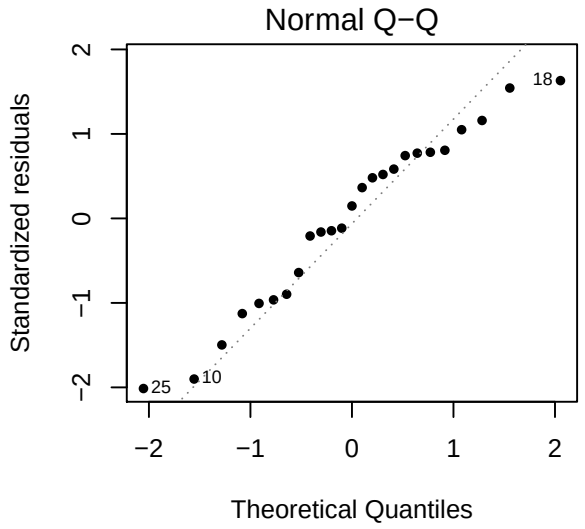


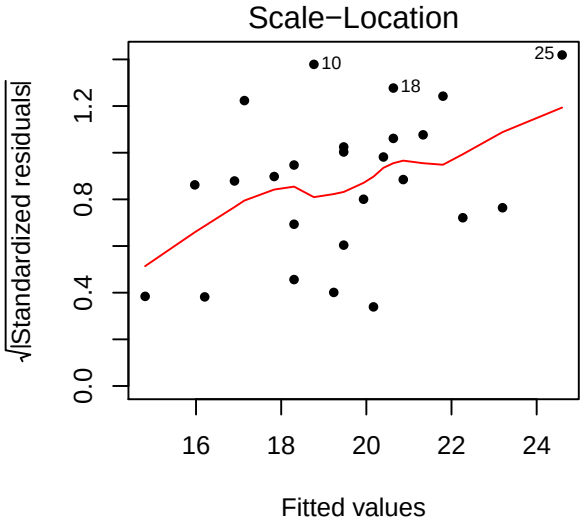
Figura: (a) violação da homocedasticidade; (b) a reta não é a equação adequada para descrever o fenômeno.

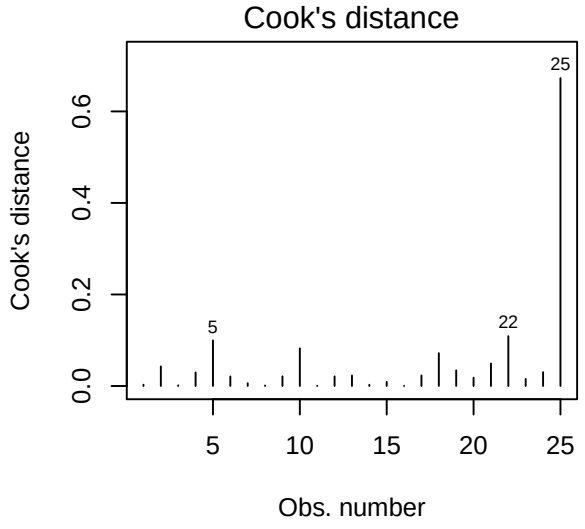
Exemplo: PIO

Análise de resíduos









Exemplo: PIO

Análise de resíduos

- Shapiro-Wilks: $p - value = 0,3994$.
- Faraway: $p - value : 0,0641$.

Conclusão: existe uma leve evidência de heterocedasticidade mas não foi significativa. Uma possível razão desta violação é a presença de um valor atípico (indivíduo 25 do banco de dados, 77 anos e 22 mmHg de PIO).

Possíveis melhorias no ajuste: (1) excluir o indivíduo 25 (somente se houver suspeitas reais de erro dos valores); (2) modelar a dispersão dos dados (modelo heterocedástico).

Exemplo: PIO
Conclusão Final

- Existe uma forte evidência de associação (**linear**) entre idade e PIO.
- Coeficiente de correlação de Pearson (força da associação) foi de 0,84 (IC; 95% 0,68; 0,93) e o $R^2 = 71\%$.
- O aumento em ano na idade, aumenta a PIO média em 0,23 (IC; 95% 0,17; 0,29) mmHg. Ou
- O aumento em quatro anos na idade, aumenta a PIO média em cerca de 1 mmHg (IC; 95% 0,68; 1,16) mmHg.

Regressão linear múltipla

- É comum querermos explicar a variável resposta/desfecho utilizando mais de uma variável explicativa.
- No exemplo, a PIO pode mudar entre homens e mulheres. Ou seja, parte da variação da PIO que não foi explicado pela idade, pode ser devido ao gênero.
- Em termos de modelo isto significa que:

$$PIO = \alpha + \beta_1 idade + \beta_2 sexo + \epsilon$$

Regressão linear múltipla

PIO vs Idade e Gênero

- **Exemplo:** (Clínica Oftalmológica) Acredita-se que a *pressão intra-ocular* é explicada (depende) pela *idade e também pelo gênero*. Relembrando:

Indivíduo	Idade	Sexo	PIO	Indivíduo	Idade	Sexo	PIO
1	35	M	15	14	55	F	20
2	40	M	17	15	57	F	19
3	41	M	16	16	58	F	20
4	44	M	18	17	59	F	19
5	45	M	15	18	60	F	23
6	48	M	19	19	60	F	19
7	50	M	19	20	61	F	22
8	50	M	18	21	63	F	23
9	50	M	17	22	65	F	24
10	52	M	16	23	67	F	23
11	54	M	19	24	71	F	24
12	55	F	18	25	77	F	22
13	55	F	21				

A partir de um banco de dados em que a i-ésima observação é o vetor $(y_i, x_{i1}, \ldots, x_{ip})$ queremos estimar $\alpha, \beta_1 \ldots, \beta_p$ pelo método de mínimos quadrados. Ou seja, queremos encontrar os valores que minimizam a seguinte função:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2$$

em que p é o número de variáveis explicativas.

Estatística de teste

Para testar $H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$ a estatística de teste é dada por

$$t = \frac{\hat{\beta}_1}{\hat{ep}(\hat{\beta}_1)},$$

em que $\hat{ep}(\hat{\beta}_i)$ é uma estimativa para o *erro padrão* do estimador $\hat{\beta}_1$. Sob H_0, t tem distribuição t de *Student* com $n - p - 1$ graus de liberdade (sob certas suposições).

De forma análoga, temos os testes para as demais variáveis.

Regressão linear múltipla

Tabela de Análise de Variância

- $SQ\text{total} = \sum_{i=1}^{n}(y_i - \overline{y})^2$,
- $SQ\text{res} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$
- $SQ\text{reg} = SQ\text{total} - SQ\text{res}$.

Termo	SQ	GL	QM	F	valor-p
Regressão	$SQ\text{reg}$	p	$QM\text{reg} = SQ\text{reg}/(p-1)$	$QM\text{reg}/QM\text{res}$	
Resíduo	$SQ\text{res}$	$n-p-1$	$QM\text{res} = SQ\text{res}/(n-p)$		
Total	$SQ\text{total}$	$n-1$			

A estatística F está testando a hipótese $H_0 : \beta_1 = \ldots = \beta_p = 0$. Ou seja, nenhuma variável explicativa está associada com a resposta.

- Avaliação sobre a qualidade do ajuste.

$$R^2 = \frac{SQ\text{regressão}}{SQ\text{total}} 100\%$$

- As variáveis explicativas podem ser quantitativas ou qualitativas.

- Utilizar resíduos para verificar a adequação do modelo ajustado:
  - Homocedasticidade/Linearidade: gráfico de resíduos vs ajustados;
  - Homocedasticidade: gráfico de absoluto(resíduos) vs ajustados e seu respectivo teste;
  - normalidade: gráfico de probabilidade normal e o teste de Shapiro-Wilks.

- Variáveis qualitativas com mais de dois níveis devem ser representadas por variáveis indicadoras.
- Um termo extra pode ser incluído para explicar a interação entre duas variáveis explicativas.
- Seleção de Modelos: Método Stepwise (Backward).
- Usar um termo quadrático, ou similar, se o tipo de relação não for linear.

Exemplo: Raça: branca, parda e preta.

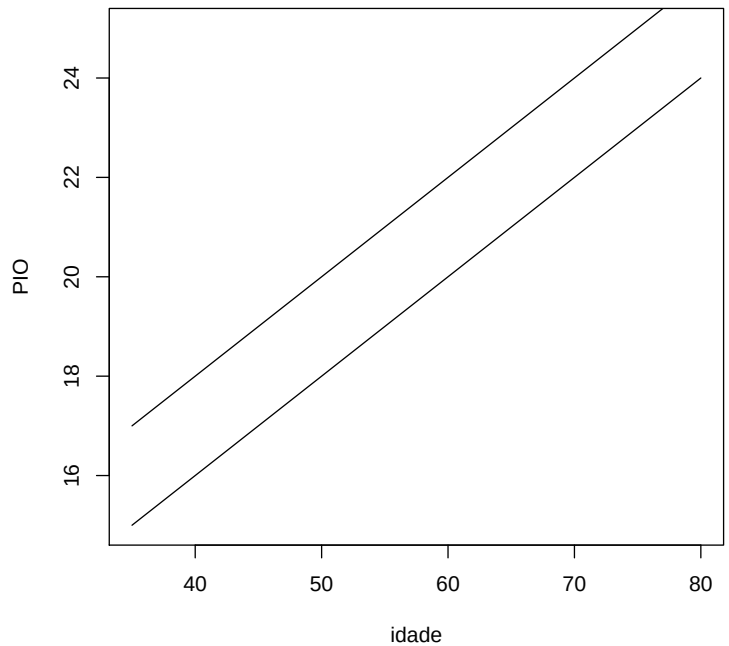
Raça	X_1	X_2
Branca	0	0
Parda	1	0
Preta	0	1

Regressão linear Múltipla

Sem Termo de Interação

$$PIO = \alpha + \beta_1 idade + \beta_2 sexo$$

- Homem: $PIO = \alpha + \beta_1 idade$
- Mulher: $PIO = (\alpha + \beta_2) + \beta_1 idade$



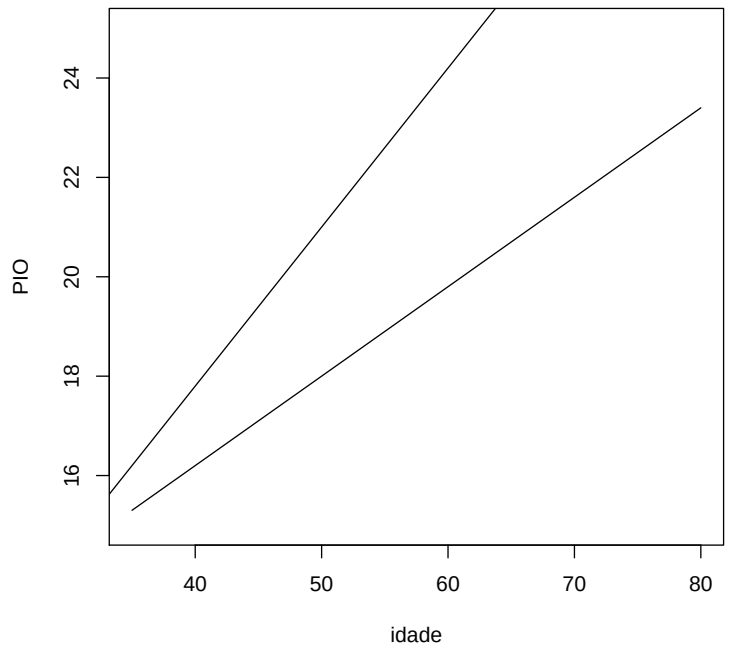
Mesma inclinação para homen e mulher (sem interação).

Regressão linear Múltipla

Com Termo de Interação

$$PIO = \alpha + \beta_1 idade + \beta_2 sexo + \beta_3 (idade * sexo)$$

- Homem: $PIO = \alpha + \beta_1 idade$
- Mulher: $PIO = (\alpha + \beta_2) + (\beta_1 + \beta_3) idade$



Inclinações diferentes para homen e mulher (com interação).

Regressão linear Múltipla

Com Termo de Interação

- Interação significa que idade modifica o efeito de sexo e vice versa.

- No slide anterior mostramos a interpretação de idade, modificada por sexo. Ou seja, a taxa de crescimento da PIO com a idade para os homens (β_1) é diferente da taxa das mulheres ($\beta_1 + \beta_3$).

- Da mesma forma, a comparação entre homens e mulheres é modificada pela idade:

  - Idade=40 anos: $PIO = \alpha + 40\beta_1$ (homem)
    $PIO = (\alpha + 40\beta_1) + \beta_2 + 40\beta_3$ (mulher)
    Ou seja a diferença entre homens e mulheres no PIO médio é de $\beta_2 + 40\beta_3$

  - Idade=60 anos : a diferença entre homens e mulheres no PIO médio é de $\beta_2 + 60\beta_3$

- Os regressores (fatores) podem ser correlacionados.
- Exemplo no Projeto 2: IMC, percentil IMC e escore z IMC.
- Efeito da multicolinearidade é inflacionar os erros-padrão.
- Forma de detecção mais simples: matriz de correlações. Devemos ter cuidado com fatores categóricos.
- O VIF (Variance Inflation Factor) é uma forma mais adequada para esta tarefa. Suspeitar de VIF$>$10.

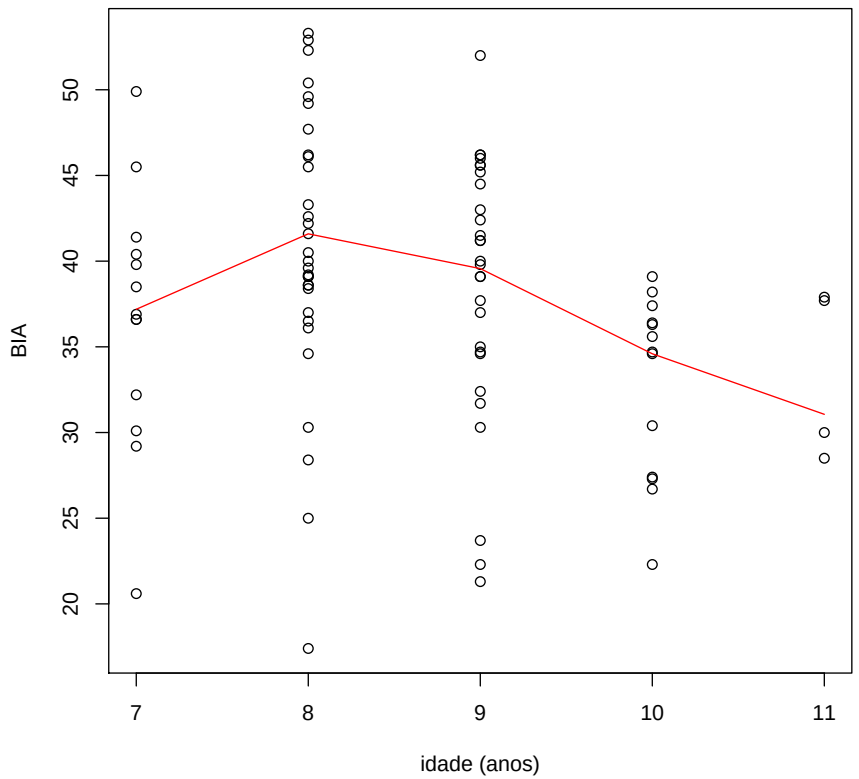
$$VIF(\beta_i) = \frac{1}{1 - R_i^2}.$$

uma regressão do i-ésimo fator contra os demais, e usa-se o R_i^2.

- O pacote car do R obtem o VIF a partir do ajuste do modelo mod (mod$<-lm(y \sim X1 + X2 + \ldots + X_p)$)

Exemplo: Projeto 2

$$BIA = \alpha + \beta_1 idade + \beta_2 (idade^2) + \text{outras} + \epsilon$$



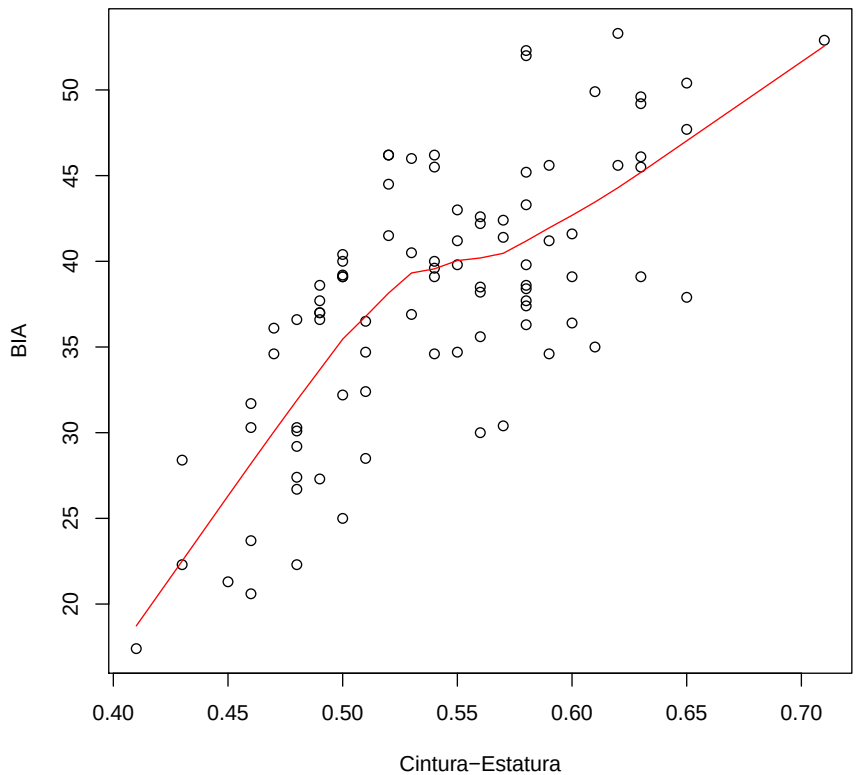
A curva vermelha é um alisamento local (lowess).

Regressão linear Múltipla

Forma Funcional da Covariável

Exemplo: Projeto 2

$$BIA = \alpha + \beta_1 cintura - estatura + \text{outras} + \epsilon$$



A curva vermelha é um alisamento local (lowess).

Exemplo: PIO vs Idade e Gênero

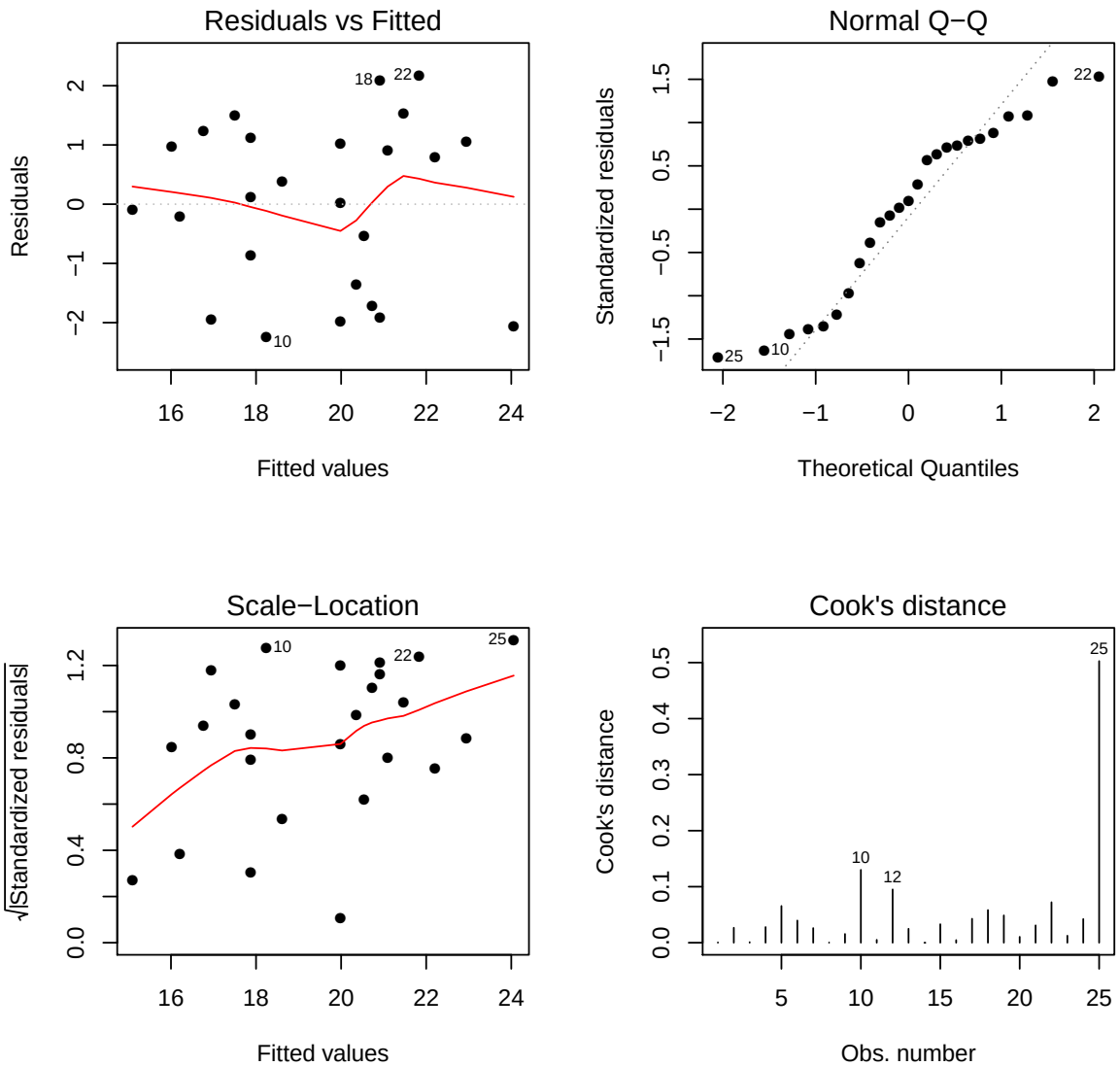
Ajuste de Mínimos Quadrados.

Termo	Estimativa	Erro-Padrão	t	valor-p
Intercepto	8,616	2,336	3,689	0.00128 **
Idade	0,185	0,0496	3,736	0.00115 **
Gênero	1,187	0,9655	1,230	0.23181

- $$R^2 = 0,73$$
- No ajuste do modelo sem idade, o termo para gênero é significativo (valor-p $< 0,001$). O que está acontecendo?

Regressão linear múltipla

Análise de resíduos - similar ao ajuste somente com idade.



Regressão linear múltipla

Testes utilizando os resíduos

- Teste de Faraway (2004) para Homocedasticidade
  valor-p = 0,064.
- Teste de Shapiro-Wilks para Normalidade
  valor-p = 0,055.

UM ROTEIRO PARA A CONSTRUÇÃO DE MODELOS

1. Descrever o estudo e seus objetivos.
2. Desenho/Planejamento do Estudo
   - Identificar o desenho amostral;
   - Identificar possíveis fatores de confusão.
   - Identificar possíveis fontes viéses (informação, seleção, etc).
   - Etc, etc.
3. Preparação/digitação dos dados. Verificação da Consistência da estrutura do banco de dados.

UM ROTEIRO PARA A CONSTRUÇÃO DE MODELOS

4 Análise Descritiva:

- Cada variável separadamente (gráficos, tabelas).
- Variáveis categóricas: tabelas de frequência, gráfico de barras, etc.
- Variáveis numéricas: medidas sumários (medianas, dp, quartis), box-plots, histogramas, etc.

5 Análise Univariada (ou bivariada)

- Análise de cada covariável com a desfecho/resposta.
- Teste t (correlação de Pearson) ou regressão linear simples.
- Verificar forma funcional (linear, quadrática, etc) de covariáveis quantitativas.
- REGRA EMPíRICA:

$$\text{Excluir covariáveis com valor-p} > 0,25$$

UM ROTEIRO PARA A CONSTRUÇÃO DE MODELOS

6 Modelo de Regressão Multivariado

- Utilizar as covariáveis não excluídas no passo anterior;
- Investigar possíveis associações entre as covariáveis (colinearidade);
- Obter um "Modelo Final"utilizando algum método de construção de modelos.

7 Verificar a adequação do modelo ajustado.

8 Incluir possíveis termos de interação.

9 Interpretar o modelo final apresentando intervalos de confiança para as quantidades de interesse.

10 Escrever o Relatório.

Regressão linear múltipla

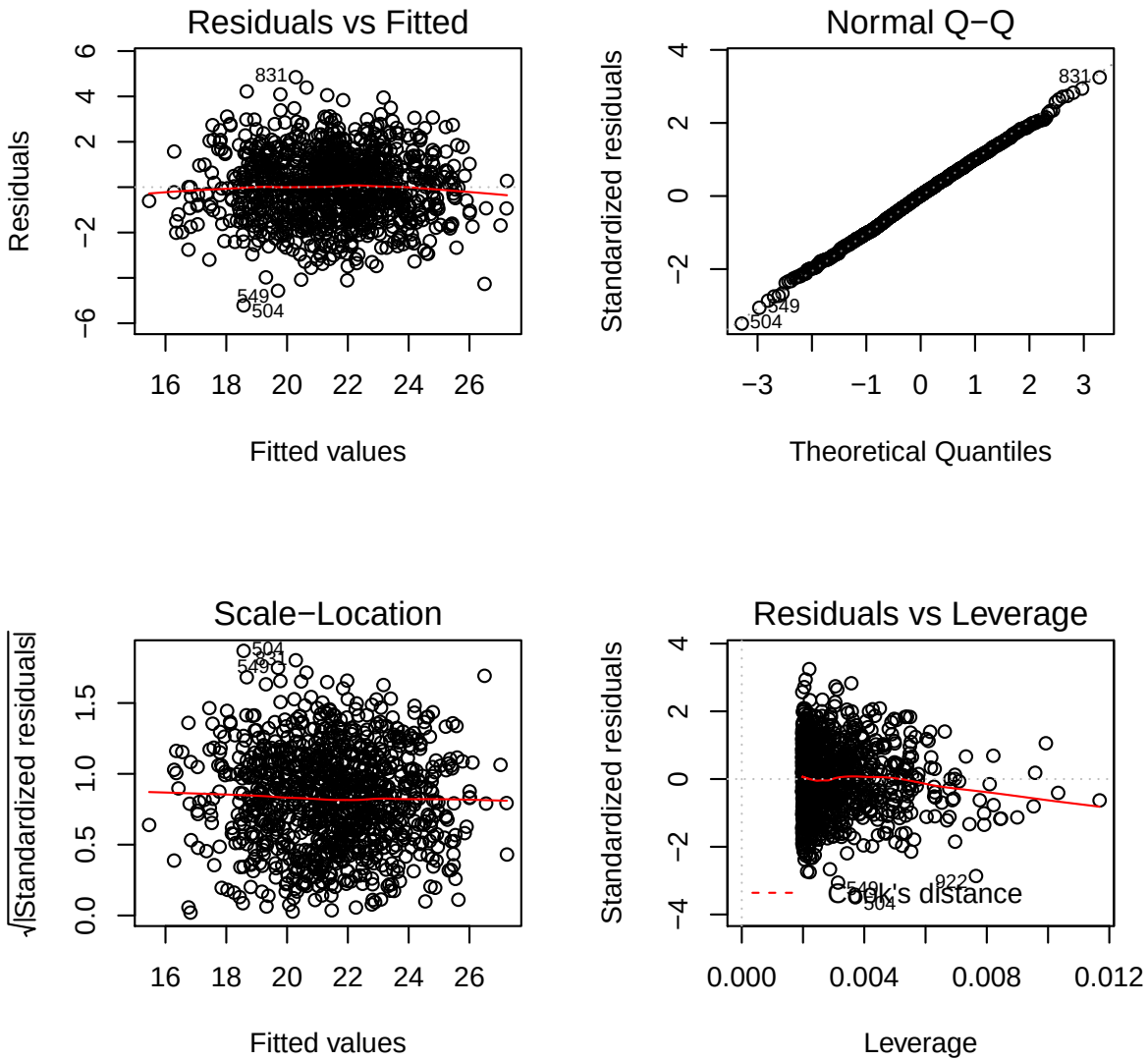
Simulação - Violação das Suposições

- Valores próximos do exemplo de PIO = Idade + Sexo
- $\beta_0 = 10,\ \beta_1 = 0.2, \beta_2 = 1.$
- $n = 1000, \sigma = 1.5$
- Idade $\sim N(55, 10)$ e Sexo $\sim$ Bernoulli (0,5)

Termo	Correto	NL (0.1 quadrático)	Hetero (0.1 idade)	Não-normal (SN 5)
constante	9,9	-286	-0,23	11
idade	0,2	11.3	0,38	0.2
sexo	1,0	-0.02	4,0	1,0

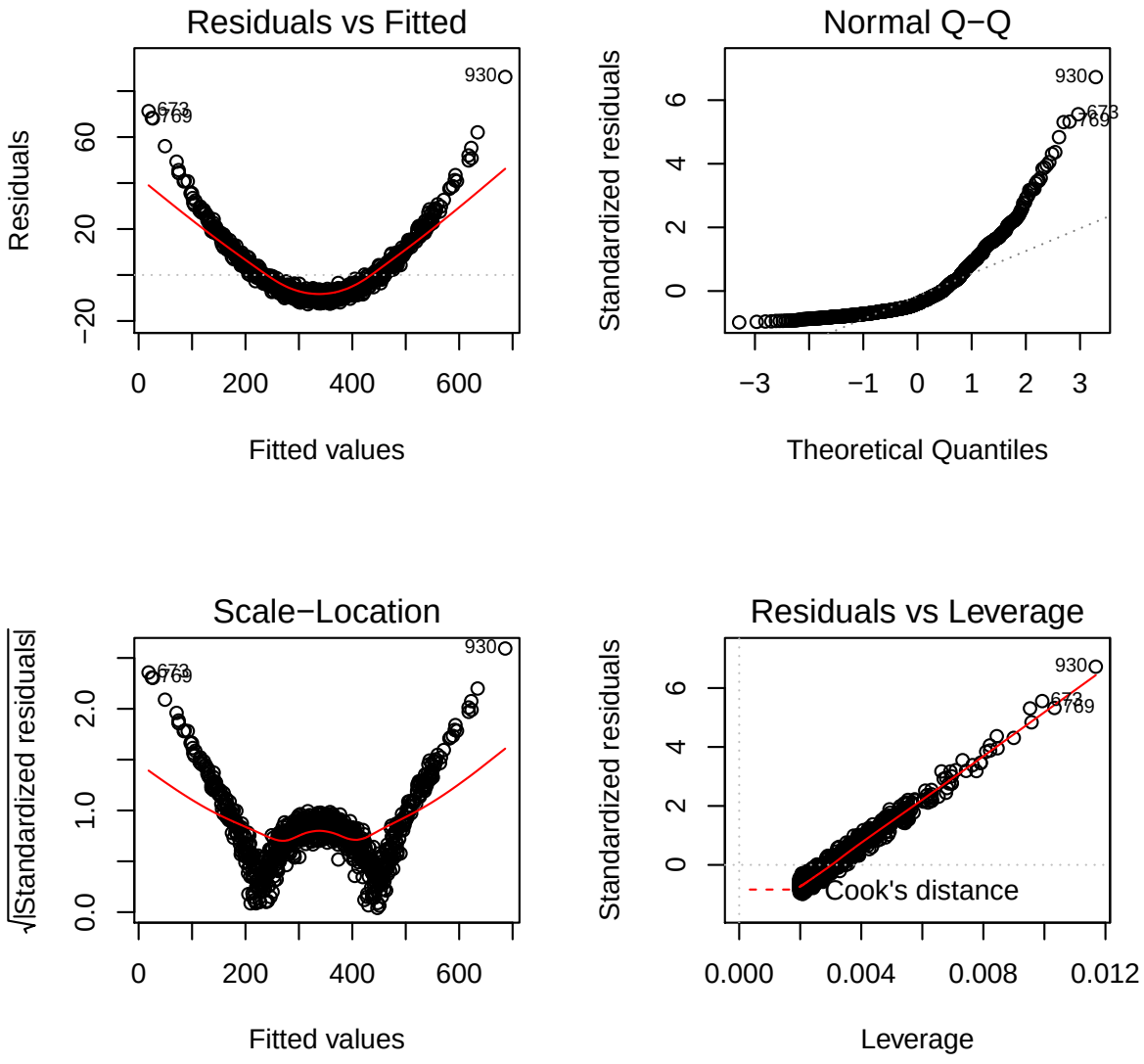
Regressão linear múltipla

Simulação - Gráfico de Resíduos - Modelo Correto



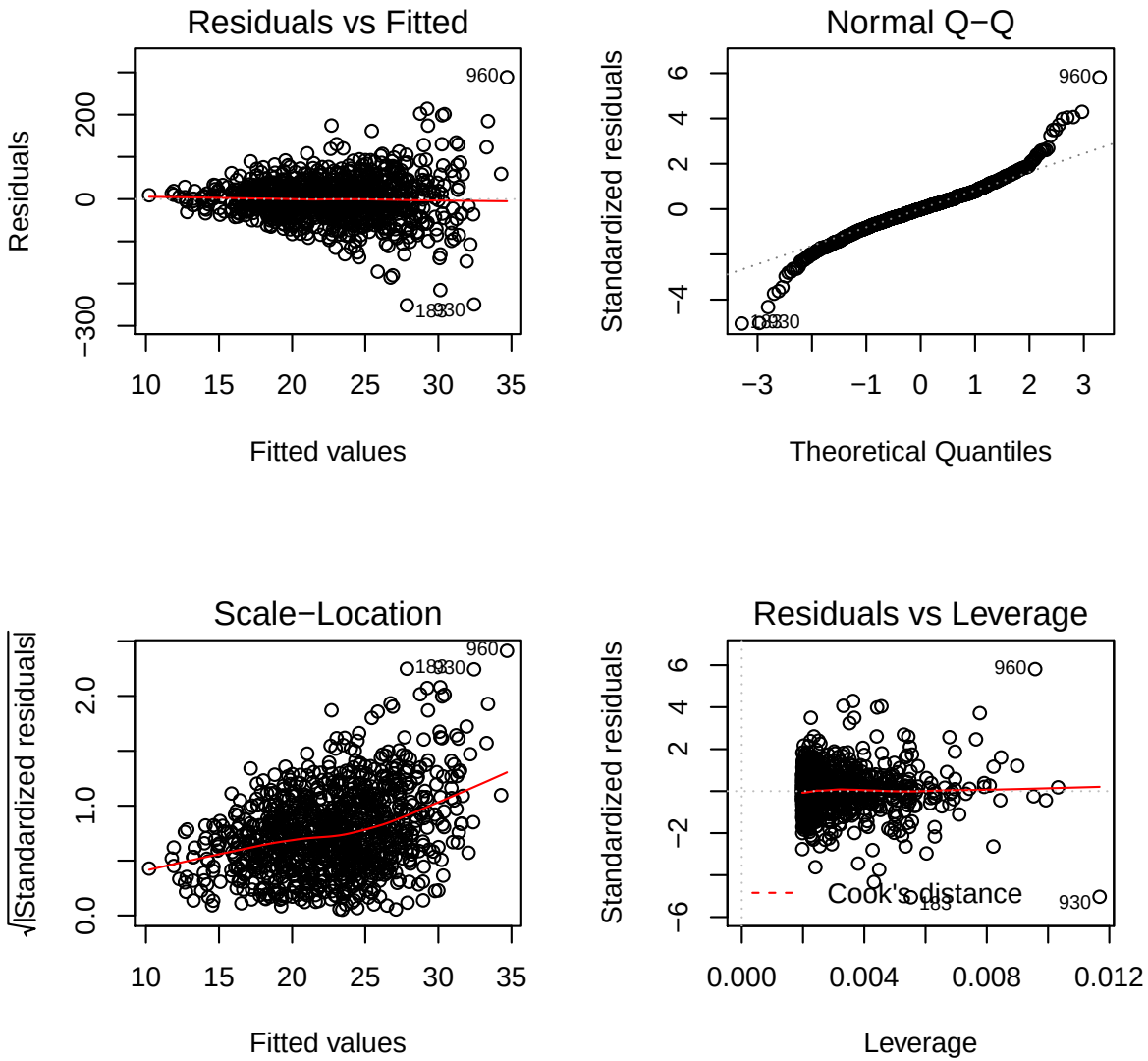
Regressão linear múltipla

Simulação - Gráfico de Resíduos - Modelo Não-Linear



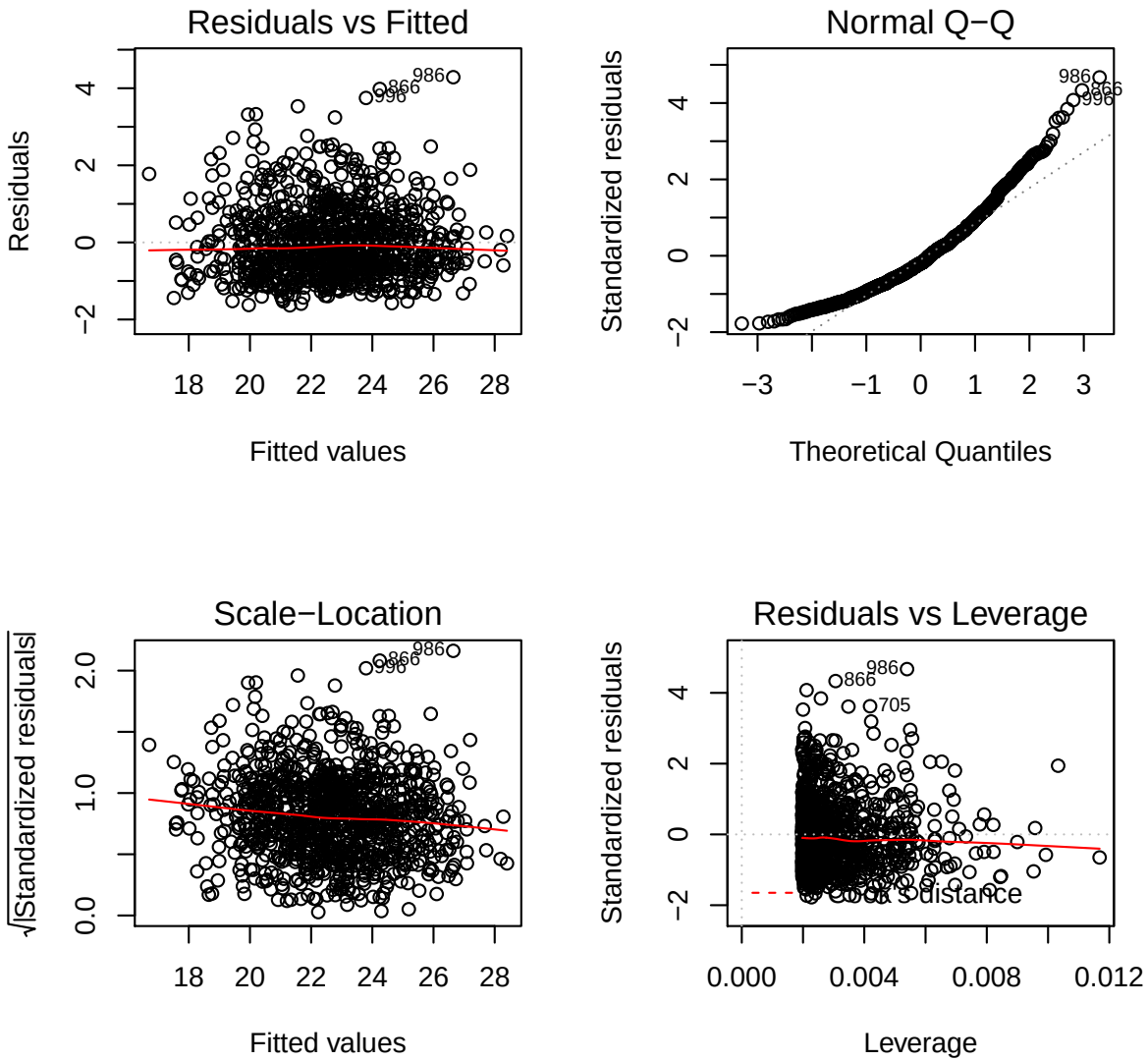
Regressão linear múltipla

Simulação - Gráfico de Resíduos - Modelo Heterocedástico



Regressão linear múltipla

Simulação - Gráfico de Resíduos - Modelo Não-Normal



Regressão linear múltipla

Propostas de Solução sob Violação das Suposições

- Verificar especificidades da própria análise:
  - presença de pontos atípicos (outliers);
  - categorizar (forma funcional) covariáveis quantitativas;
  - ausência de termos importantes (por exemplo, interação);
  - etc, etc e etc.
- Transformar a resposta (proposta antiga mas ainda em vigor).
- Utilizar outro modelo estatístico.
- Modelar o desvio-padrão do erro.
- Utilizar técnicas não-paramétricas (splines)
- Consultar um estatístico.

Regressão linear múltipla

Interpretação dos Coeficientes

Resposta: porcentagem de gordura (BIA)

Termo	Estimativa	Erro-padrão	valor-p
idade	2	0,5	$< 0,01$
sexo (menina)	4,0	1	$< 0,01$

- **Covariável Binária** (sexo) As meninas têm em média 4% (IC 95%, 2;6) a mais de gordura do que os meninos.
- **Covariável Quantitativa** (idade) O aumento de um ano na idade, aumenta em média, 2% (IC 95%, 1;3) no percentual de gordura das crianças.

Regressão linear múltipla

Interpretação Interação

Resposta: porcentagem de gordura (BIA)

Termo	Estimativa	Erro-Padrão	valor-p
conicidade	160	30	$< 0{,}01$
cinturaEstatura	350	120	$< 0{,}01$
conicidade:cinturaEstatura	-250	80	$< 0{,}01$

- **Conicidade** (cinturaestatura varia entre 0,4 e 0,7.)
  - Cinturaestatura = 0,5
    O aumento da conicidade em 0,1 unidades, aumenta em $3,5\%$ o percentual de gordura das crianças.
  - Cinturaestatura = 0,6
    O aumento da conicidade em 0,1 unidades, aumenta em 1% o percentual de gordura das crianças.
- **CinturaEstatura** (conicidade varia entre 1,1 e 1,4.)
  - conicidade = 1,2
    O aumento da cinturaestatura em 0,1 unidades, aumenta em 5% o percentual de gordura das crianças.
  - conicidade = 1,4
    O aumento da cinturaestatura em 0,1 unidades, aumenta em 0% o

Resposta: porcentagem de gordura (BIA)

Termo	Estimativa	Erro-Padrão	valor-p
Idade	24	10	$< 0{,}01$
Idade^2	-1.5	0,4	$< 0{,}01$

- **7 para 8 anos**
  - 8 anos: 24*8 - 1,5*64 = 96
  - 7 anos: 24*7 - 1,5*49 = 94,5
    O aumento de 7 para 8 anos, aumenta em $1,5\%$ o percentual de gordura das crianças.
- **8 para 9 anos**
  O aumento de 8 para 9 anos, reduz em $1,5\%$ o percentual de gordura das crianças.
- **9 para 10 anos**
  O aumento de 9 para 10 anos, reduz em $4,5\%$ o percentual de gordura das crianças.