Estatística II

Correlação e Regressão Linear Simples

Enrico A. Colosimo

Departamento de Estatística
Universidade Federal de Minas Gerais
http://www.est.ufmg.br/~enricoc

Associação entre Variáveis Numéricas

- Avaliar se existe **associação** entre características quantitativas é objetivo de muitos estudos em biologia e ciências da saúde.

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

- **Exemplo:** Vinte e cinco pacientes são atendidos em uma clínica oftalmológica e os seguintes valores são registrados para *pressão intra-ocular* (PIO) e *idade*:

Associação entre Variáveis Numéricas

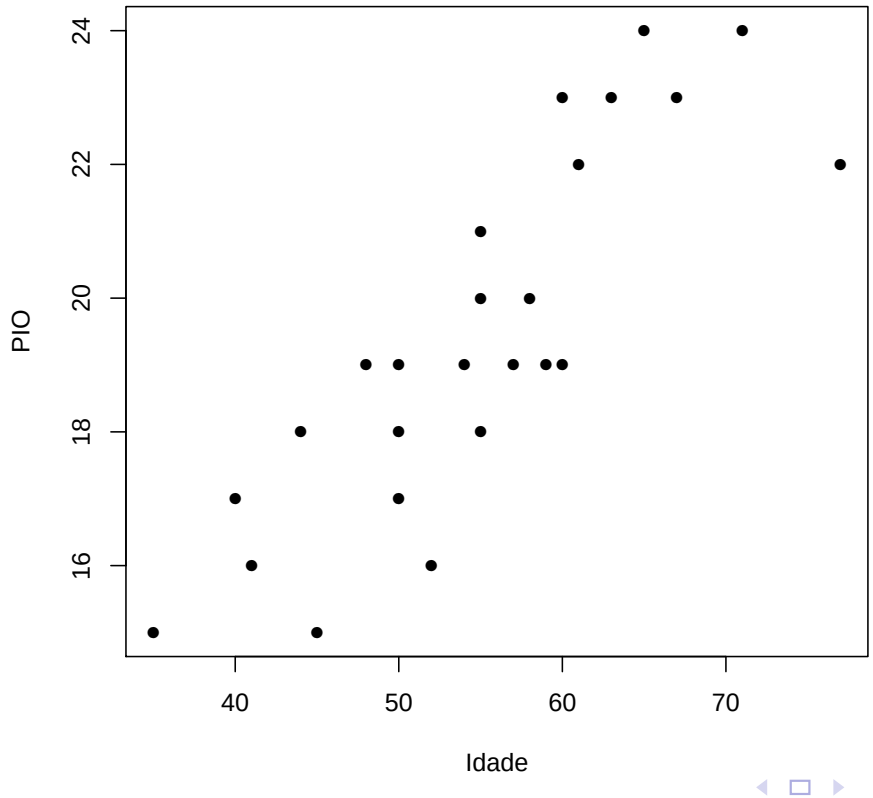
Estudo Transversal - Tabela de dados brutos

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			

Associação entre Variáveis Numéricas

Exemplo - Diagrama de dispersão

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



Associação entre Variáveis Numéricas

Exemplo

- Através do diagrama de dispersão é possível observar que, em geral, valores de idade mais altos são associados com valores de pressão intra-ocular mais altos (as variáveis parecem relacionadas).

- **Perguntas:**
  - Qual é o tipo da relação entre as variáveis *Idade* e *PIO*?
  - Qual é a força desta relação?
  - A relação é estatisticamente significativa?

Associação entre Variáveis Numéricas

Tipos de relação

- A relação entre duas variáveis quantitativas pode ser descrito como *positivo*, *negativo*, *linear* e *não-linear*.

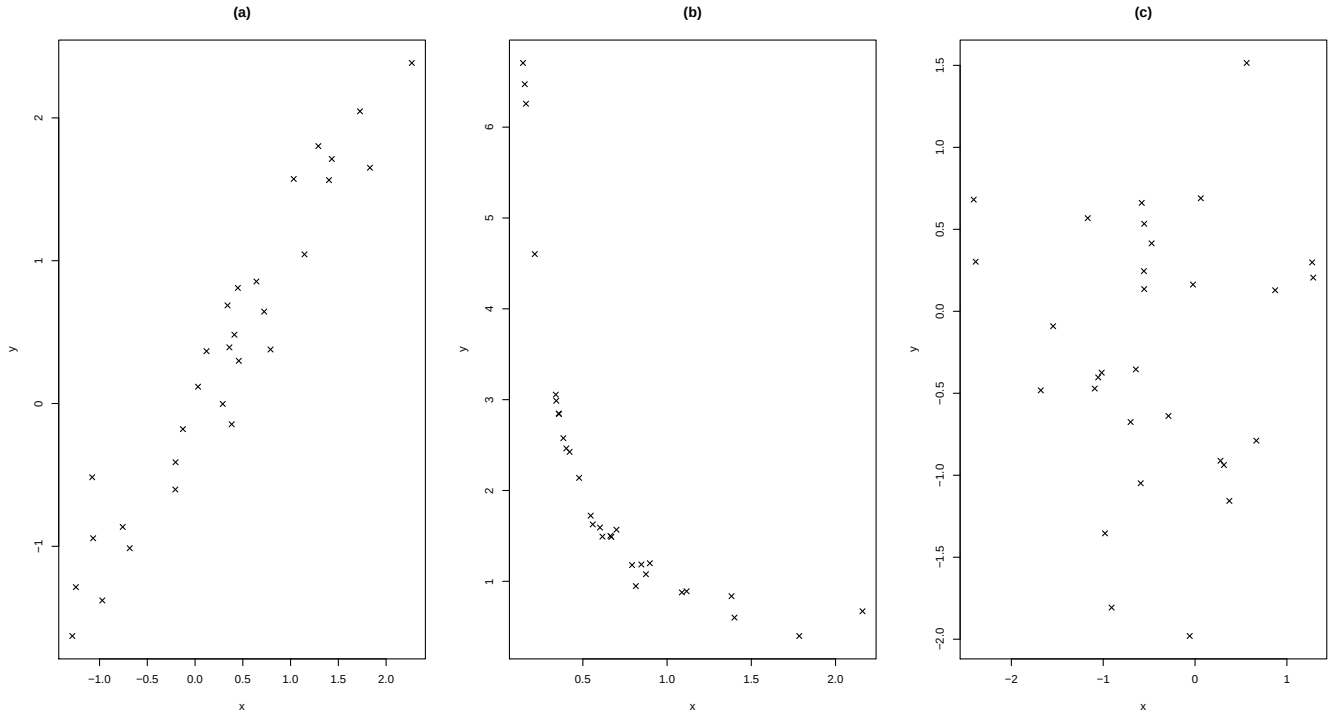


Figura: Tipos de relação entre x e y: (a) relação linear (*linha reta*)positiva; (b) relação não-linear negativa; (c) sem relação.

Associação entre Variáveis Numéricas

Tipos de relação

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

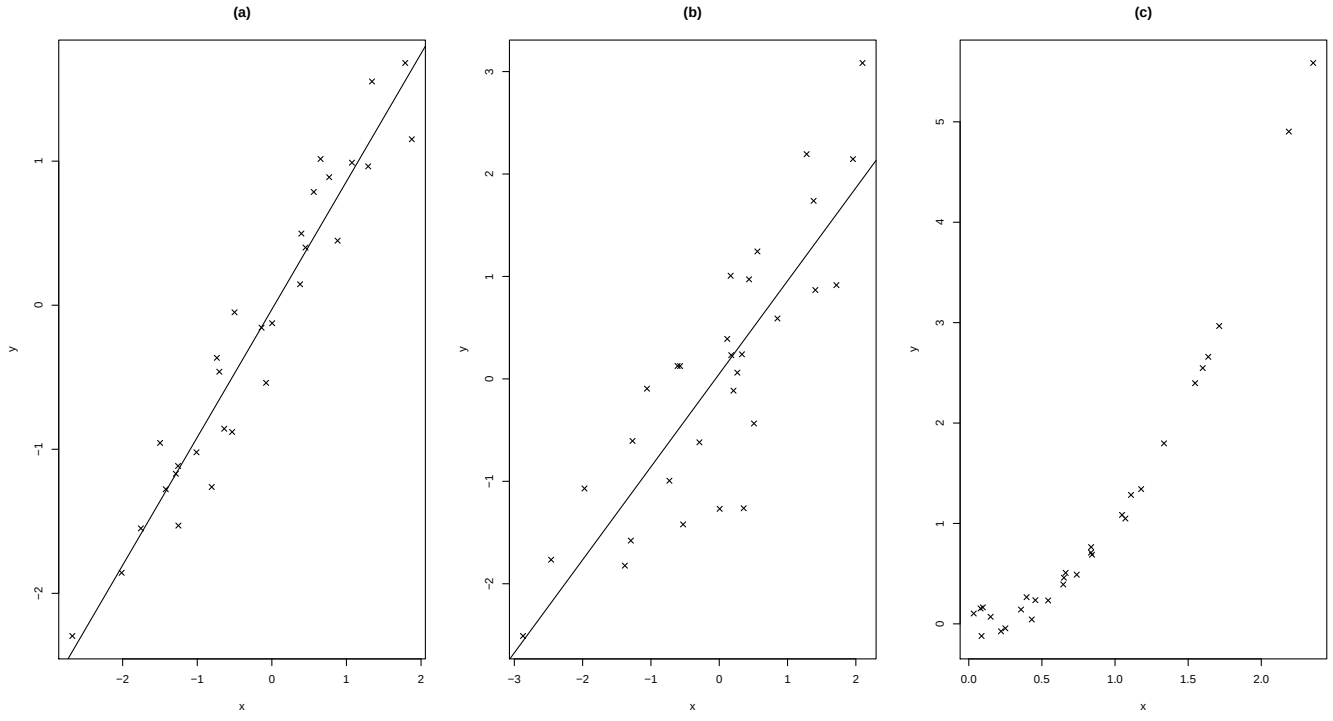


Figura: (a) relação linear positiva forte; (b) relação linear positiva fraca; (c) relação não-linear forte.

Coeficiente de correlação

- A força de uma associação pode ser medida por um **coeficiente de correlação**.
- O *coeficiente de correlação de Pearson* é uma medida da intensidade de associação linear existente entre duas variáveis quantitativas.
- Também é conhecido como *coeficiente de correlação de Pearson*, pois sua fórmula de cálculo foi proposta por Karl Pearson em 1896.
- O coeficiente de correlação de Pearson é denominado por ρ e sua estimativa por r.

Coeficiente de correlação

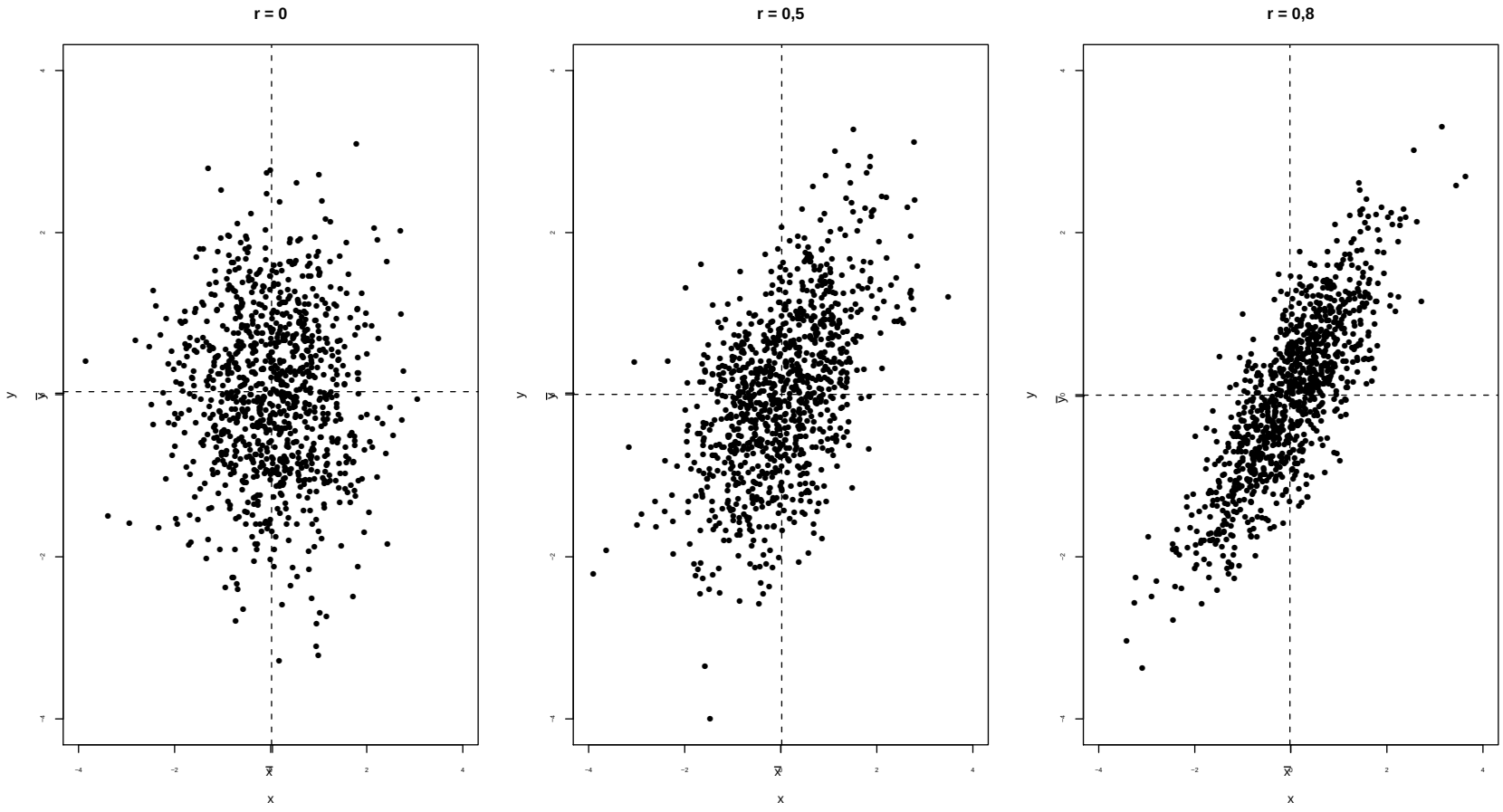
- O coeficiente de correlação varia entre -1 e 1.
- Valores negativos indicam uma correlação do tipo inversa (*negativa*):
  - Quando x aumenta, y em média diminui (ou vice-versa).
- Valores positivos ocorrem quando a correlação é direta (*positiva*):
  - x e y variam no mesmo sentido.
- **Exemplos:**
  - As taxas sanguíneas de insulina e glicose apresentam correlação negativa.
  - No exemplo da clínica oftamológica, idade e pressão intra-ocular têm correlação positiva (ainda não sabemos 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 inclinada (correlação perfeita).
- Quando não existe correlação entre x e y, os pontos se distribuem em nuvens circulares.
- Quando os pontos formam uma nuvem cujo eixo principal é uma curva, o valor de r não mede corretamente a associação entre as variáveis.

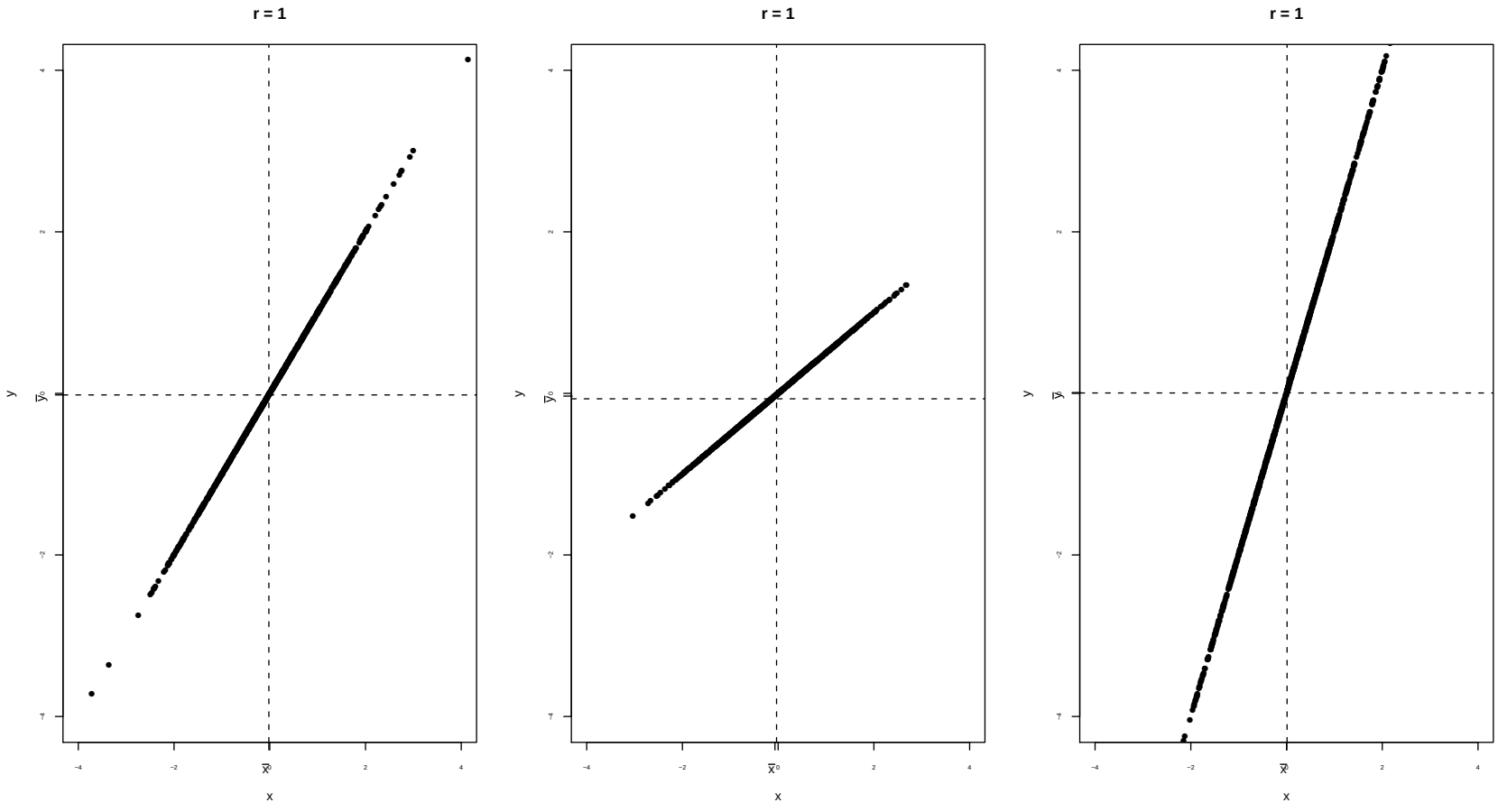
Coeficiente de correlação

Exemplos



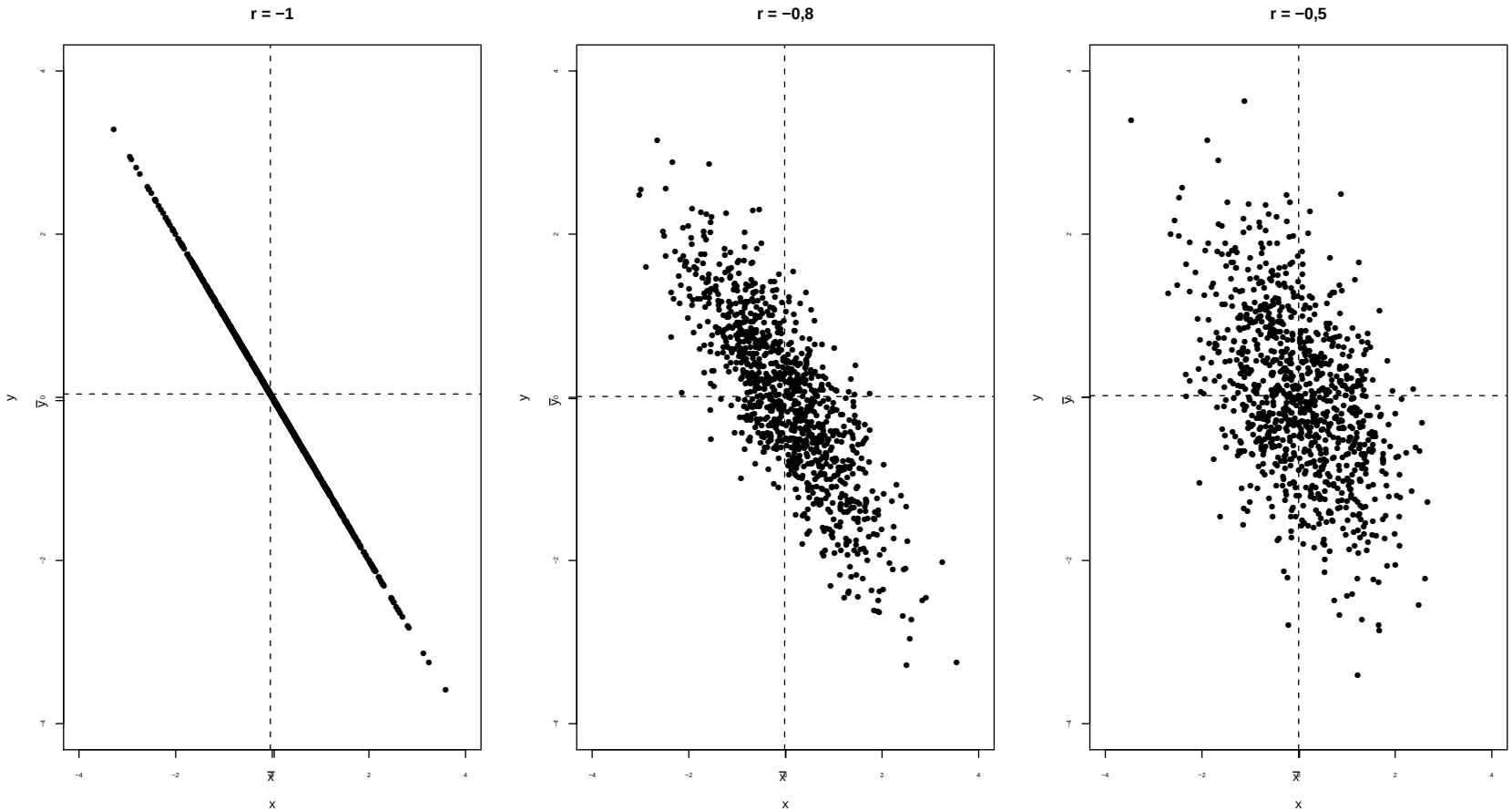
Coeficiente de correlação

Exemplos



Coeficiente de correlação

Exemplos



Coeficiente de correlação

Cálculo do coeficiente de correlação

- A fórmula para se obter o coeficiente de correlação de Pearson em uma amostra é

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

  - $cov_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ é a covariância de x e y
  - s_x e s_y são os desvios-padrões de x e y respectivamente

- Desta forma

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \times \sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Coeficiente de correlação

Cálculo do coeficiente de correlação

Fórmula alternativa para calcular r

A seguinte fórmula é mais conveniente, pois exige um número menor de operações aritméticas.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=\mathbf{1}}^{n} x_i \sum_{i=\mathbf{1}}^{n} y_i}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=\mathbf{1}}^{n} x_i)^{\mathbf{2}}}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=\mathbf{1}}^{n} y_i)^{\mathbf{2}}}{n}\right]}}.$$

- No exemplo da clínica oftamológica:

$$r = \frac{27217 - (1372 \times 486)/25}{\sqrt{(77634 - 1882384/25)(9626 - 236196/25)}} = 0,8448.$$

Coeficiente de correlação

Teste de hipóteses sobre a correlação

- Quando se calcula o coeficiente r em uma amostra, é necessário ter em mente que se está, na realidade, estimando a associação verdadeira entre x e y existente na população.
- A correlação na população é designada por ρ.
- Para avaliar a significância do coeficiente de correlação, geralmente testa-se a hipótes nula $H_0 : \rho = 0$.

Estatística de teste

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

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

Coeficiente de correlação

Teste de hipóteses sobre a correlação

- **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, $t_{0,05;23} = 2,069$

$$t_{calc} = 0,8448\sqrt{\frac{25-2}{1-0,8448^2}} = 7,5718 > 2,069$$

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

Coeficiente de correlação

Força da correlação

- Uma vez determinada a existência de correlação na população, pode-se avaliá-la qualitativamente quanto à intensidade.

$\|r\|$	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

Correlação Espúria

- É comumente encontrada uma forte relação positiva linear entre o número de médicos em uma cidade eo número de mortes por ano nessa cidade!
- À primeira vista, podemos ser tentados a concluir que ter mais médicos conduz a mais mortes.
- Assim, se reduzissemos o número de médicos, esperaríamos menos mortes.
- Podemos calcular correlações para qualquer par de variáveis, mas sempre devemos ter cuidado ao assumir que uma causa variação na outra.

Outros Coeficientes de Correlação

Correlação de Spearman

- O coeficiente de correlação de Pearson é paramétrico no sentido que assume distribuição normal para as duas variáveis.
- Ele deve ser preferido caso esta suposição for validada pois é mais preciso.
- Caso a suposição não for validada, podemos utilizar o coeficiente de correlação de Spearman.
- Este coeficiente de correlação é baseado nos postos das observações e, portanto, não paramétrico.

Regressão linear simples

- O estudo da **regressão** aplica-se àquelas situações em que há razões para supor uma relação de causa-efeito entre duas variáveis quantitativas e se deseja expressar matematicamente essa relação.
- Geralmente chama-se a *variável desfecho* (variável resposta) de Y e a *covariável* (fator, 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 do estudo da regressão são:
  1. Avaliar uma possível associação de Y em relação a x.
  2. Expressar matematicamente esta relação por meio de uma equação.
- Muitas são as relações de causa e efeito que podem ser resumidas por linhas retas, evitando-se, assim, o uso de tabelas de dados para mostrar a relação.
- 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 uma variável independente apenas (por isso, “simples").

Regressão linear simples

Equação da reta

- A equação da reta pode ser dada por:

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

em que

  - Y: variável resposta
  - 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

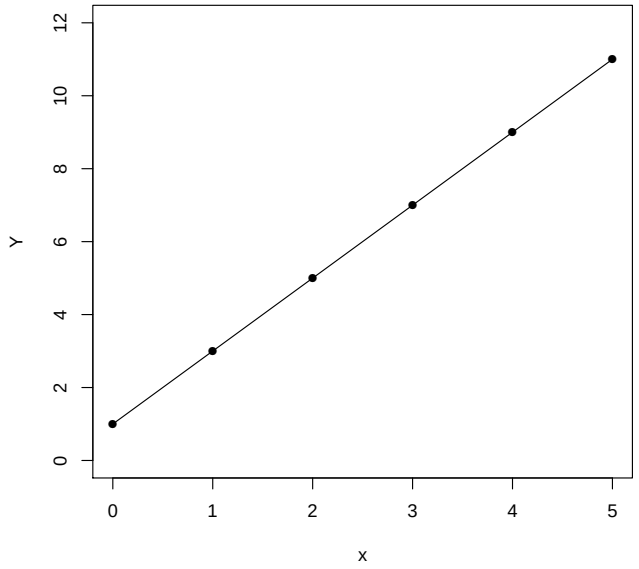
- Outras representações:
  - $E(Y|X = x) = \alpha + \beta x$
  - $\mu_{Y|X=x} = \alpha + \beta x$

Regressão linear simples

Equação da reta

- **Exemplo:** Suponha que o número de médicos (Y) em uma clínica é completamente determinado pelo número de pacientes (x), da seguinte forma:

$$Y = 1 + 2x,$$



Regressão linear simples

Obtendo a reta de regressão

- A reta de regressão (verdadeira) seria obtida se fossem conhecidos os valores de x e Y para todos os indivíduos da população.
- No entanto, o mais comum é estudar a regressão entre x e Y utilizando uma amostra da população de pontos.
- São calculados $\hat{\alpha}$ e $\hat{\beta}$, que são estimativas de α e β.
- Esses valores são obtidos pelo *Método dos Mínimos Quadrados*, assim chamado porque garante que a reta obtida é aquela para a qual se tem as menores distâncias (ao quadrado) entre os valores observados (y) e a própria reta.

Regressão linear simples

Obtendo a reta de regressão

- O coeficiente β é estimado da seguinte maneira:

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

- O coeficiente linear 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

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

Regressão linear simples

Obtendo a reta de regressão

- **Exemplo:** (Clínica Oftalmológica) Desconfia-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:

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

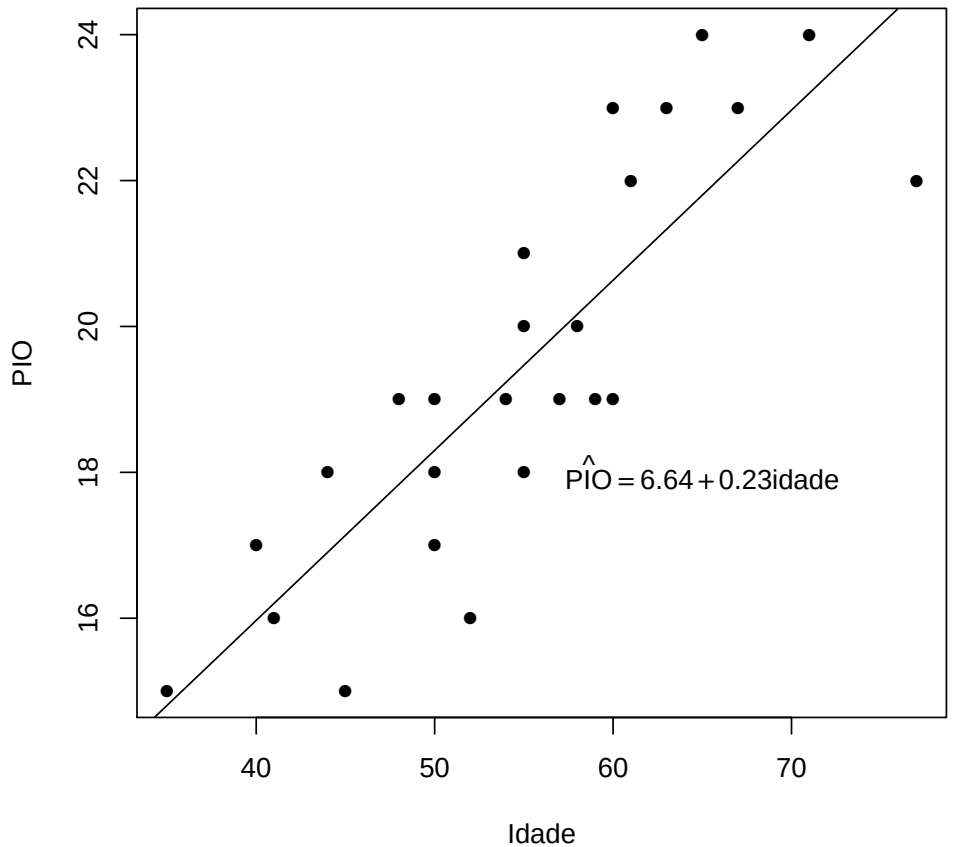
- $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{27217 - (1372)(486)/25}{77634 - 1882384/25} = 0,23318$$

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

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

Regressão linear simples

Obtendo a reta de regressão - Exemplo (Clínica Oftalmológica)



Regressão linear simples

Teste de significância da regressão

- A **dependência** de Y em relação a x é representada pelo coeficiente β.
- β é estimado com base em uma amostra de dados.
- Para afirmar que $\hat{\beta}$ representa uma depêndencia real de Y em relação a x deve-se realizar um teste de hipóteses sobre a existência de regressão na população.
- O principal teste de interesse é verificar se x influencia na resposta, o que é equivalente a testar

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

Regressão linear simples

Teste de significância da regressão

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

$$t = \frac{\hat{\beta}}{ep_{\hat{\beta}}},$$

em que $ep_{\hat{\beta}}$ é o *erro padrão* do estimador $\hat{\beta}$ e é dado por

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

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

Regressão linear simples

Utilidades da reta de regressão

A reta de regressão permite:

1. Representar a dependência de uma variável quantitativa em relação à outra por meio de uma equação simples.
2. Prever valores para a variável dependente Y de acordo com valores determinados (inclusive não-observados) da variável independente x.

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

Regressão linear simples

Requisitos ao uso da regressão linear

Certas exigências devem ser satisfeitas para se realizarem inferências válidas a reta de regressão linear (algumas se referem à estimação de α e β e outras dizem respeito ao teste de hipóteses para β).

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

Se (1), (2) e (3) forem violadas, transformações nos dados são maneiras contornar o problema.

Regressão linear simples

Análise de resíduos

Um teste para a validade das pressuposições para a regressão pode ser feito do seguinte modo:

1. Calculam-se os **resíduos** para cada valor de y
   - Os resíduos (ϵ) representam a diferença entre aquilo que foi realmente observado e o que foi predito pelo modelo de regressão:

$$\epsilon = (y - \hat{y}).$$

2. Desenha-se um gráfico no qual os resíduos são colocados no eixo vertical e os valores esperados de y ($\hat{y}$), no horizontal.

Regressão linear simples

Análise de resíduos

- Os pontos devem ficar distribuídos de forma equilibrada acima e abaixo de uma linha imaginária paralela ao eixo x na altura do resíduo zero, formando uma faixa aproximadamente retangular.
- A violação do pressuposto de homocedasticidade produz faixas em forma de cone, enquanto violações do pressuposto de linearidade produz faixas curvas

Regressão linear simples

Análise de resíduos

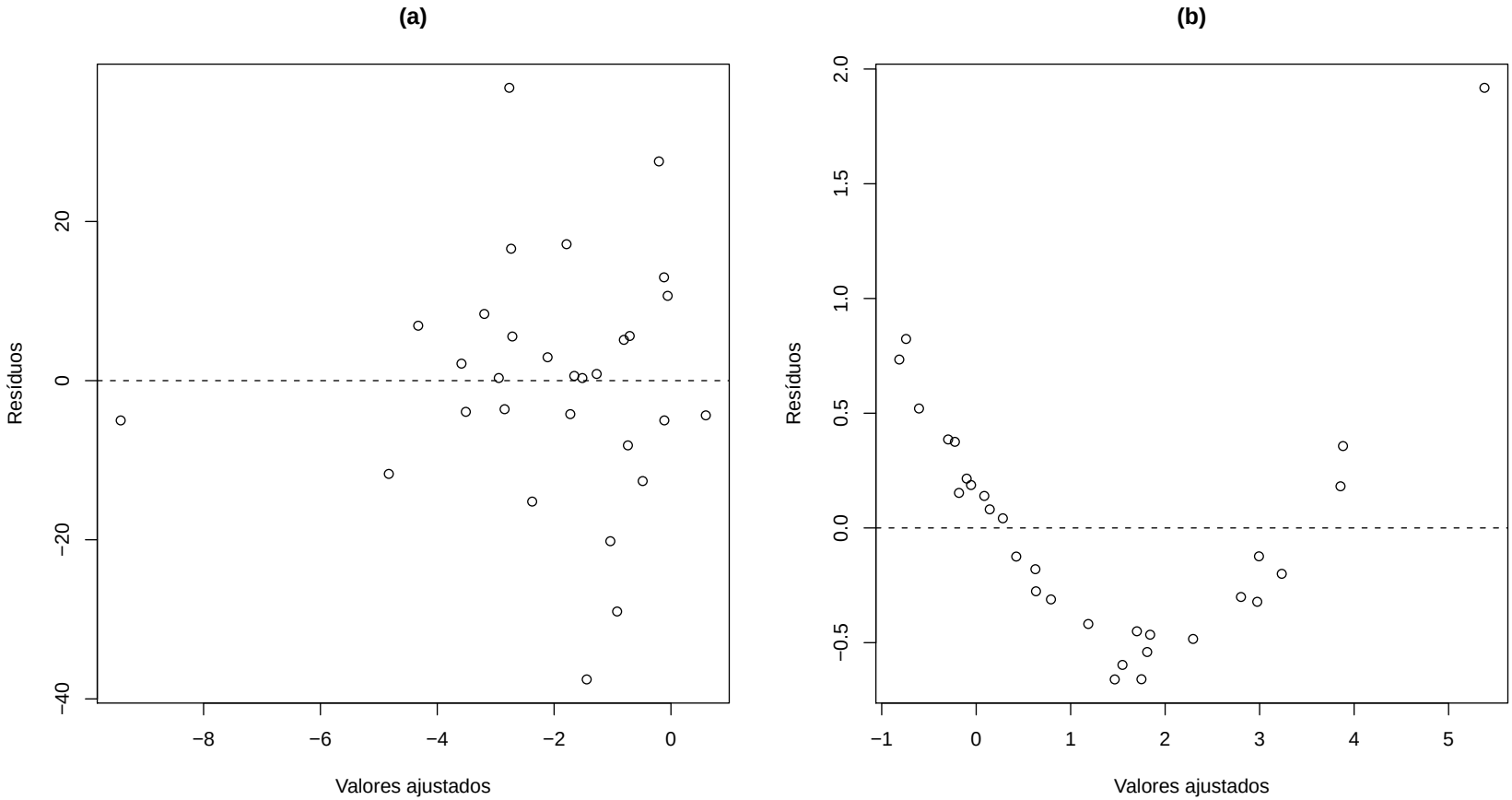


Figura: (a) violação da homocedasticidade; (b) a reta não é a equação mais adequada para descrever o fenômeno.