Regressão Linear Simples

M. Zevallos

IMECC-UNICAMP

Exemplo 1

É realizado o seguinte experimento em varetas: para valores fixados de temperatura são mensurados os valores do modulo de Young (Y) [- 3000] (Ku, 1969). Os resultados são:

Tabela 1: Experimento em varetas

Temp	**Young** (-3000)	**Temp**	**Young** (-3000)
500	328	750	174, 175, 154
550	296	800	152, 146, 124
600	266	850	117, 94
603	260, 244	900	97, 61
650	240, 232, 213	950	38
700	204, 203, 184	1000	30, 5

Exemplo 1

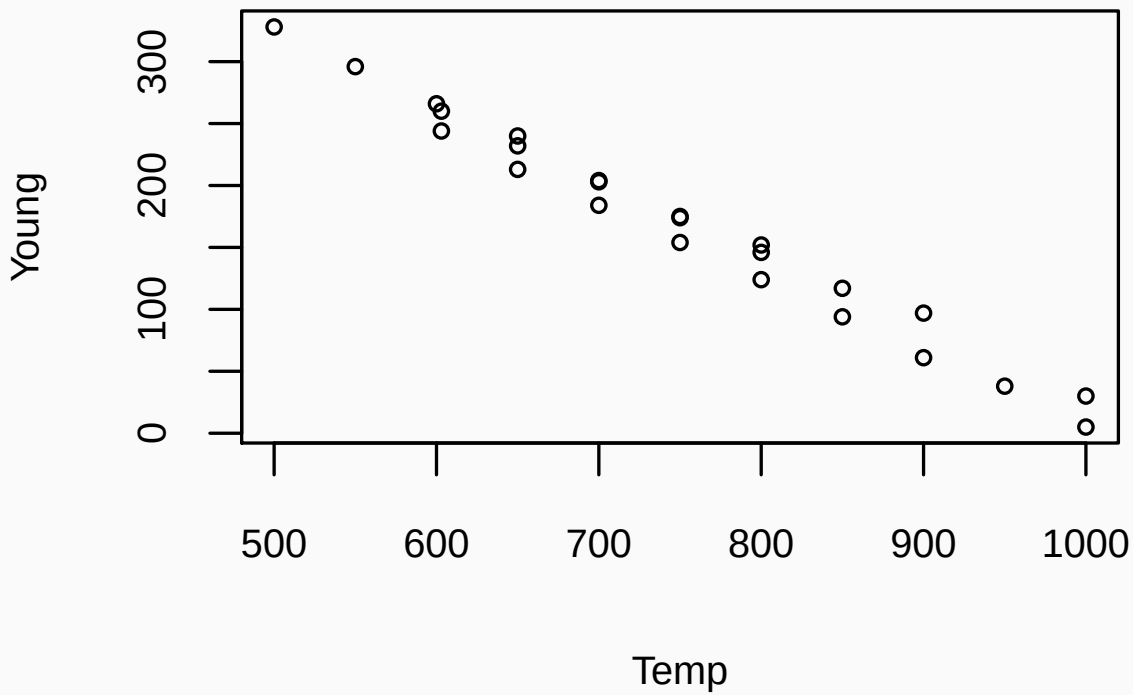


Figura 1: Gráfico de dispersão

Regressão

- **Temperatura** é a variável explanatória, denotada por X
- **Modulo de Young** é a variável resposta, denotada por Y

Regressão

A Regressão de Y dado X é definida como

$$\mu(x) = \mathbb{E}(Y \mid X = x) \tag{1}$$

Se

$$\mu(x) = \alpha + \beta x \tag{2}$$

temos uma **Regressão Linear Simples**.

Regressão Linear Simples

Em

$$\mu(x) = \alpha + \beta x \tag{3}$$

α e β são os **coeficientes** da regressão

- α é denominado **intercepto** e corresponde ao valor da reta quando $x = 0$
- β é denominada **inclinação**. Como

$$\Delta\mu(x) = \beta\Delta x \tag{4}$$

$$\beta = \frac{\Delta\mu(x)}{\Delta x} \tag{5}$$

então β é a mudança na esperança condicional de Y quando X muda em uma unidade ($\Delta x = 1$).

Modelo de Regressão Linear Simples

O modelo de Regressão Linear Simples é definido como:

$$Y = \alpha + \beta x + \varepsilon \tag{6}$$

onde ε é uma variável aleatória com $\mathbb{E}(\varepsilon) = 0$. Usualmente, para fazer inferência supomos que

$$\varepsilon \sim N(0, \sigma^2) \tag{7}$$

Então, para cada valor de x

$$Y|x \sim N(\alpha + \beta x, \sigma^2) \tag{8}$$

Estimação da Reta de Regressão

Seja a amostra $(x_1, y_1), \ldots, (x_n, y_n)$. Supondo que

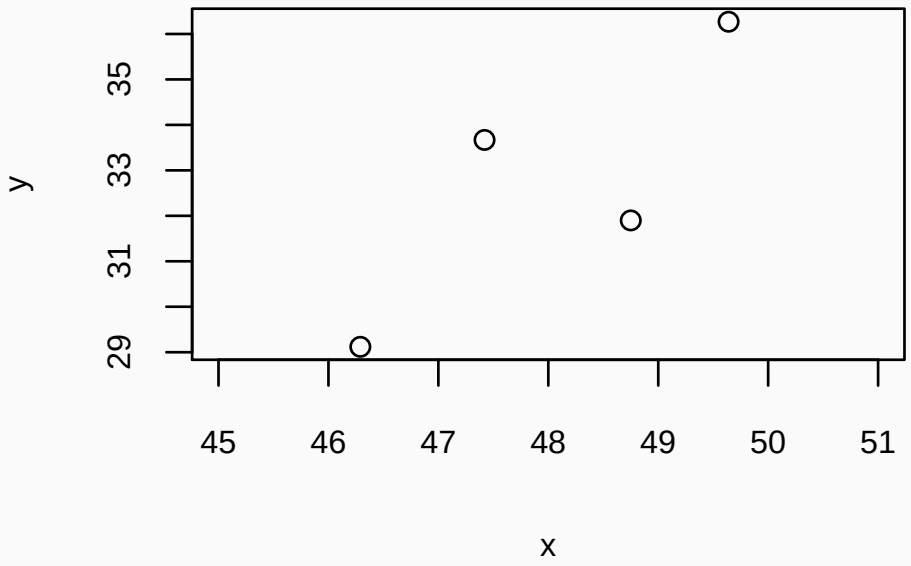
$$\mu(x) = \alpha + \beta x \tag{9}$$

Como estimar α e β a partir dos dados?

Método de Minimos Quadrados

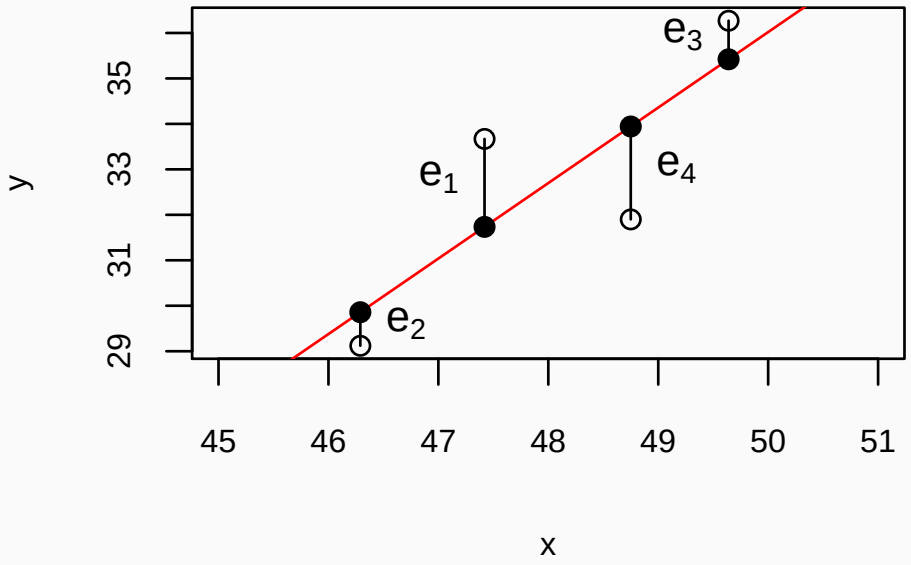
Reta de Mínimos Quadrados

- Os dados são $(x_1, y_1), \ldots, (x_n, y_n)$
- A reta estimada é $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$



Reta de Mínimos Quadrados

- Os dados são $(x_1, y_1), \ldots, (x_n, y_n)$
- A reta estimada é $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$
- Os resíduos são: $e_i = y_i - \left(\hat{\alpha} + \hat{\beta}x_i\right)$
- Critério: minimizar $Q = \sum_{i=1}^{n} e_i^2$



Exemplo 1

$\hat{\alpha} = 628.629 \quad \hat{\beta} = -0.614$

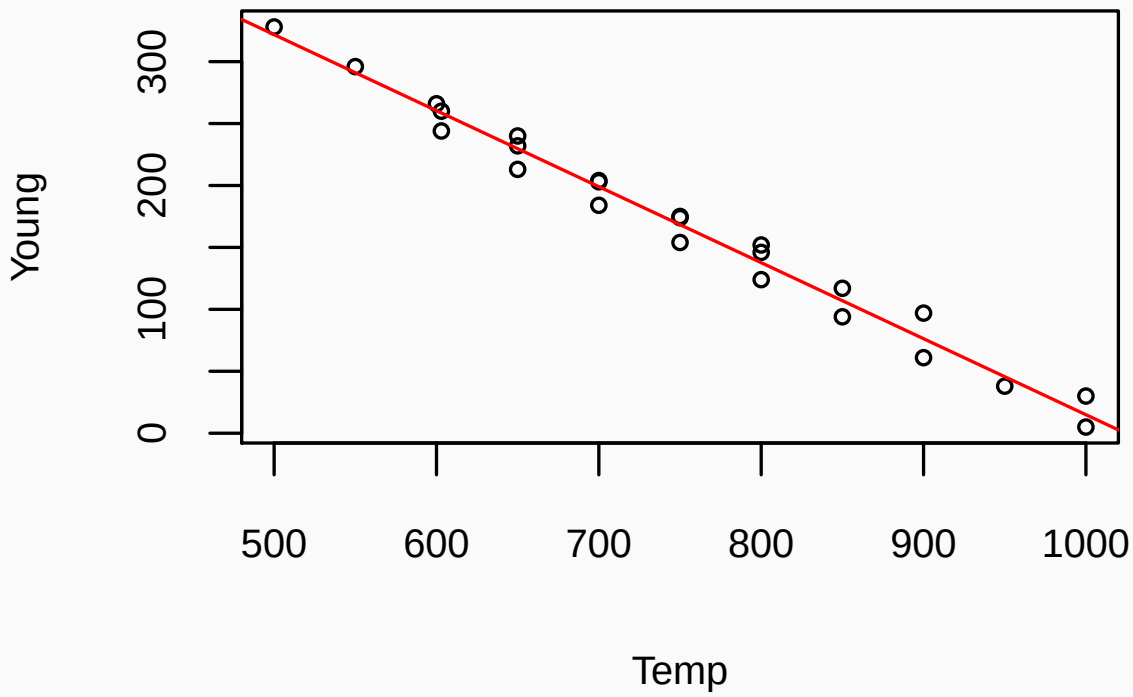


Figura 2: Gráfico de dispersão e reta de regressão estimada

Exemplo 1

- Reta estimada

$$\hat{\mu}(x) = 628.629 - 0.614x \tag{10}$$

- Se x aumenta em uma unidade então a resposta média de y diminui em 0.614
- Qual o valor da reta estimada em $x = 810$?

$$\hat{\mu}(810) = 628.629 - 0.614 \times 810 = 131.289 \tag{11}$$

- Qual o valor do resíduo para o ponto $(x, y) = (500, 328)$?

$$e = 328 - \hat{\mu}(500) = 328 - (628.629 - 0.614 \times 500) = 6.371 \tag{12}$$

Reta de Minimos Quadrados

- Para minimizar $Q = \sum_{i=1}^{n} e_i^2$ onde $e_i = y_i - \hat{\alpha} + \hat{\beta} x_i$
- Derivar com respeito a $\hat{\alpha}$ e igualando a zero (omitindo constantes)

$$0 = \frac{dQ}{d\hat{\alpha}} = \sum_{i=1}^{n} e_i \tag{13}$$

- Derivar com respeito a $\hat{\beta}$ e igualando a zero (omitindo constantes)

$$0 = \frac{dQ}{d\hat{\beta}} = \sum_{i=1}^{n} e_i x_i \tag{14}$$

- Resolver o sistema de duas equações e duas incógnitas (13)-(14)

Reta de Minimos Quadrados

Os coeficientes estimados da reta são

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} \tag{15}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{16}$$

onde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{17}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{18}$$

$$s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{19}$$

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{20}$$

Reta de Minimos Quadrados

- Como o coeficiente de correlação é

$$r = \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}}, \tag{21}$$

então

$$\hat{\beta} = r \frac{\sqrt{s_{yy}}}{\sqrt{s_{xx}}} \tag{22}$$

- A reta passa pelo ponto $(\bar{x}, \bar{y})$. Com efeito,

$$\begin{aligned} \hat{\mu}(\bar{x}) &= \hat{\alpha} + \hat{\beta}\bar{x} && (23)\\ &= (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\bar{x} && (24)\\ &= \bar{y} && (25) \end{aligned}$$

Inferência

Seja o modelo de regressão linear simples

$$Y_i = \alpha + \beta x_i + \varepsilon_i \tag{26}$$

onde as perturbaçoes (erros) ε_i, são independentes com $\varepsilon_i \sim N(0, \sigma^2)$

A seguir discutiremos três problemas:

1. Inferência para a inclinação β.

Vamos supor que o interesse é testar

$$H_0 : \beta = 0 \qquad vs \qquad H_1 : \beta \neq 0. \tag{27}$$

Para isso, podemos usar a estatística de teste

$$t = \frac{\hat{\beta}}{ep(\hat{\beta})} \sim t_{(n-2)} \tag{28}$$

onde

$$ep(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{s_{xx}}} \tag{29}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{30}$$

Inferência

Também podemos calcular um intervalo de confiança $(1-\gamma)\%$:

$$\hat{\beta} \pm ep(\hat{\beta}) \times t_{n-2,\gamma/2} \tag{31}$$

Exemplo 1

Como $\hat{\beta} = -0.614$ e $ep(\hat{\beta}) = 0.018$ então $t = 34.85$

Adicionalmente, um intervalo de confiança 95% para β é

$$-0.614 \pm 0.018 \times 2.07 = (-0.651, -0.577) \tag{32}$$

2. Intervalo de confiança $(1-\gamma)$% para a reta $\mu(x)$ em um valor $x = x_0$

$$\hat{\mu}(x_0) \pm ep(\hat{\mu}(x_0)) \times t_{n-2,\gamma/2} \tag{33}$$

onde

$$\hat{\mu}(x_0) = \hat{\alpha} + \hat{\beta} x_0 \tag{34}$$

$$ep(\hat{\mu}(x_0)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}} \tag{35}$$

Exemplo. Dados simulados com $\mu(x) = 1 + 5x$ e $\sigma = 12$. Duas amostras de $n = 10$ observações

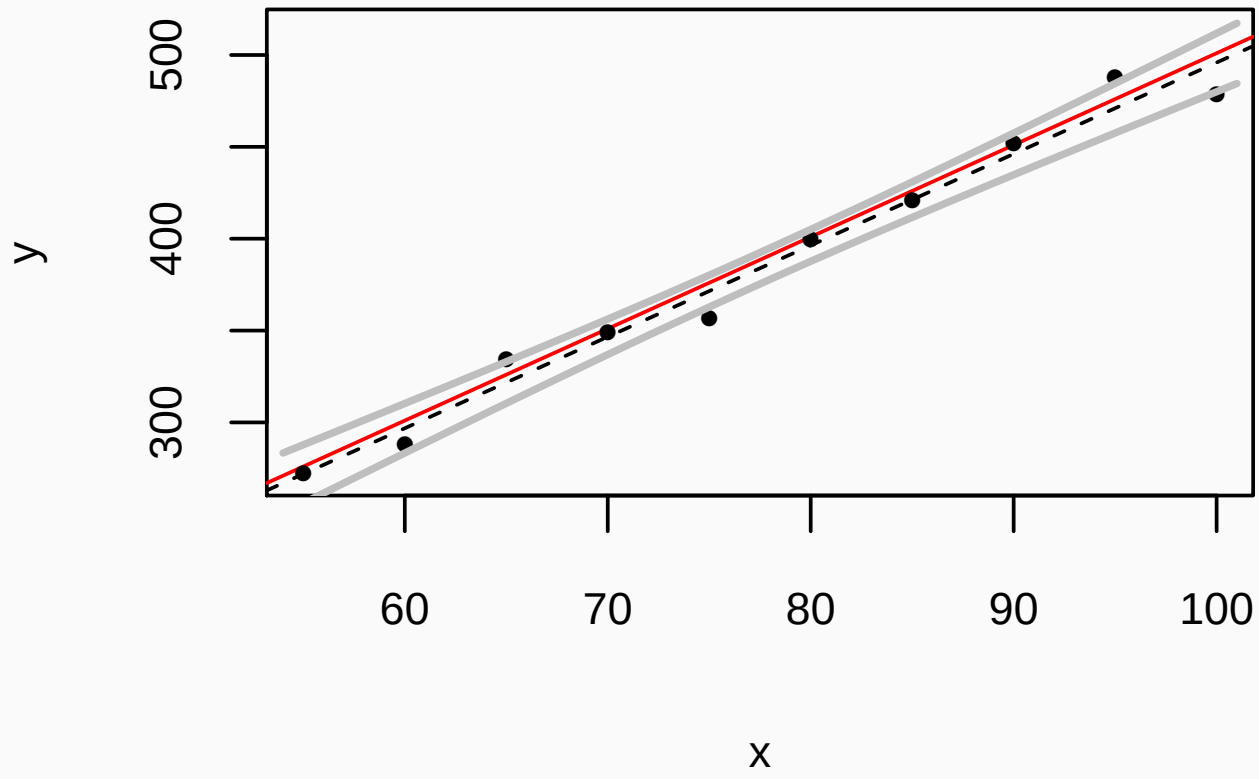


Figura 3: Dados simulados I. IC 95% para $\mu(x)$ em cinza.

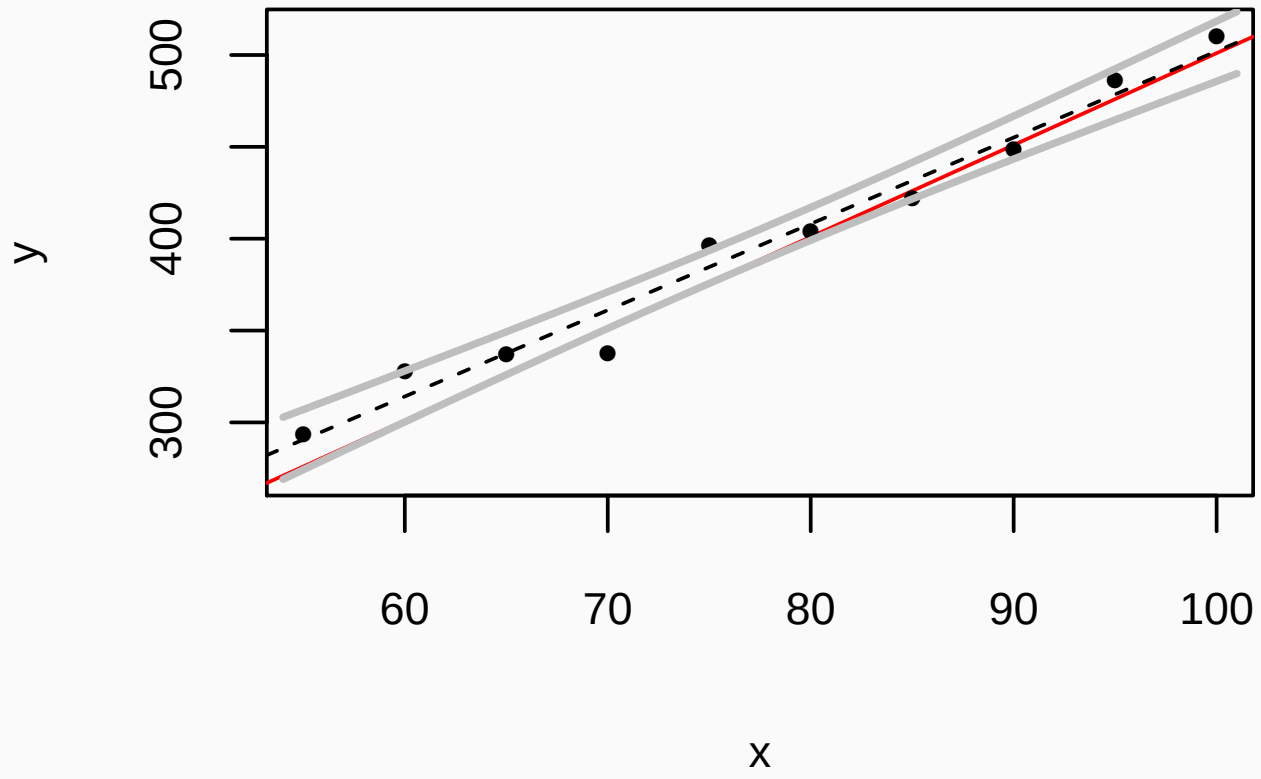


Figura 4: Dados simulados II. IC 95% para $\mu(x)$ em cinza.

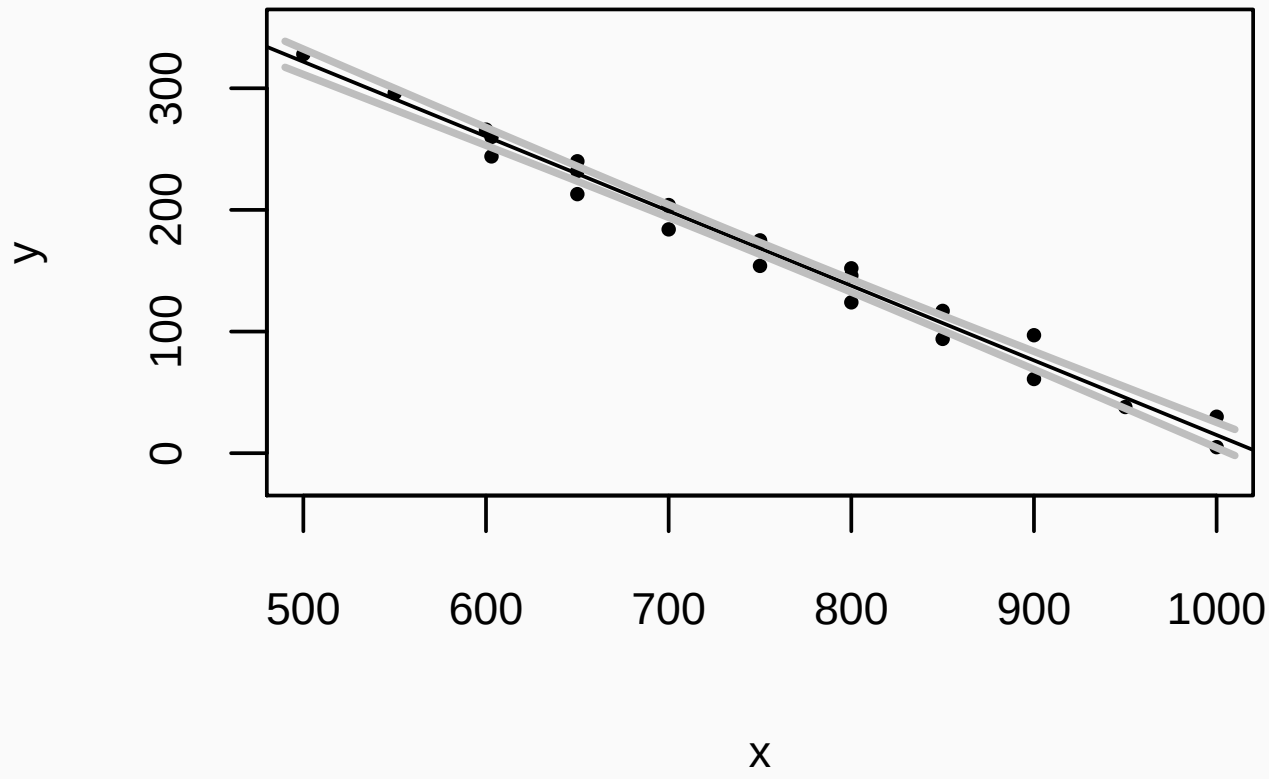


Figura 5: Intervalo de confiança 95% para a reta no Exemplo 1

3. Intervalo de previsão $(1-\gamma)$% para a resposta y em um valor x_0

$$\hat{y} \pm W \times t_{n-2,\gamma/2} \tag{36}$$

onde $\hat{y} = \hat{\alpha} + \hat{\beta} x_0$ e

$$W = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}} \tag{37}$$

Exemplo 1

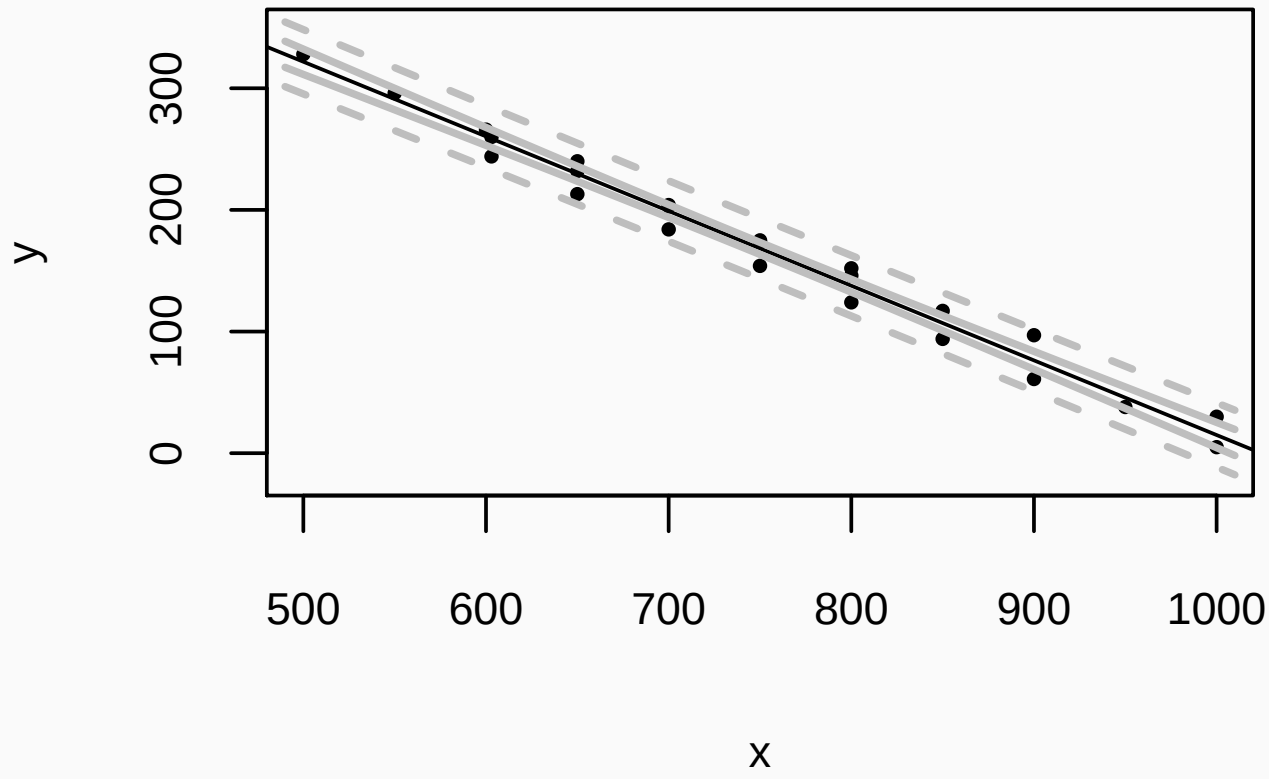


Figura 6: Intervalos de confiança para a reta (em cinza) e y (tracejado) no Exemplo 1

Coeficiente de Determinação

Para o modelo utilizado temos a seguinte **Decomposição da Soma de Quadrados**:

$$SQT = SQReg + SQE \tag{38}$$

onde

$$SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{39}$$

$$SQReg = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{40}$$

$$SQE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{41}$$

com $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$

Coeficiente de Determinação

O Coeficiente de Determinação, R^2, é a porção da variabilidade de Y explicada pela regressão:

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQE}{SQT}, \tag{42}$$

e satisfaz

$$0 \leq R^2 \leq 1. \tag{43}$$

Usualmente é expresso em porcentagem.

Exemplo 1

$R^2 = 0.98$

Análise de Resíduos

Temos suposto que as perturbações ε_i são independentes com

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

Os resíduos são:

$$e_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

e os resíduos padronizados:

$$\tilde{e}_i = \frac{e_i}{\hat{\sigma}} \tag{44}$$

Quando o modelo de regressão linear simples está corretamente especificado os resíduos padronizados são aproximadamente independentes com distribuição $N(0, 1)$. Nesse caso, o gráfico dos resíduos não indica nenhum padrão. Caso contrário devemos questionar a validez das suposições do modelo de regressão linear simples.

Exemplo: Dados de Anscombe

Tabela 2: Dados de Anscombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Exemplo: Dados de Anscombe I

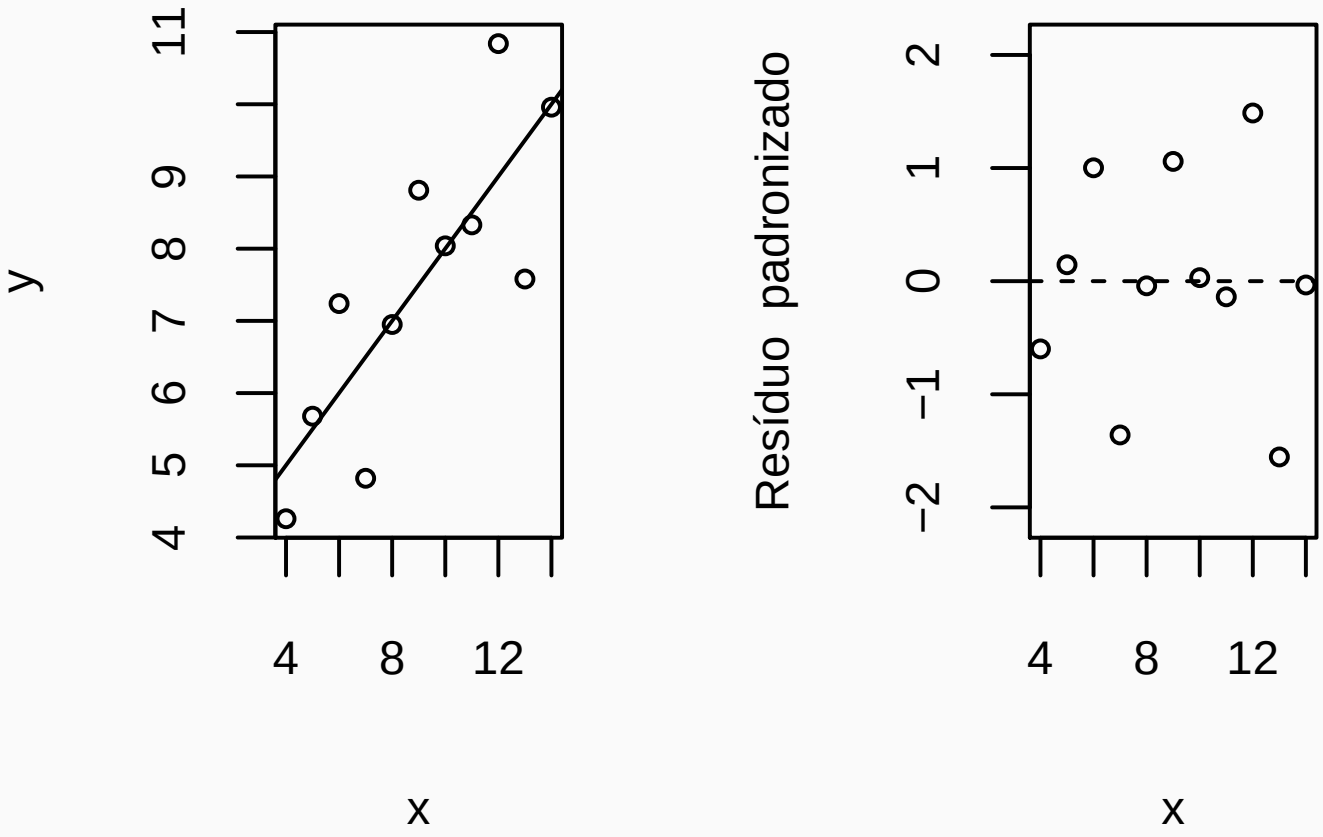


Figura 7: Dados de Anscombe I

Exemplo: Dados de Anscombe II

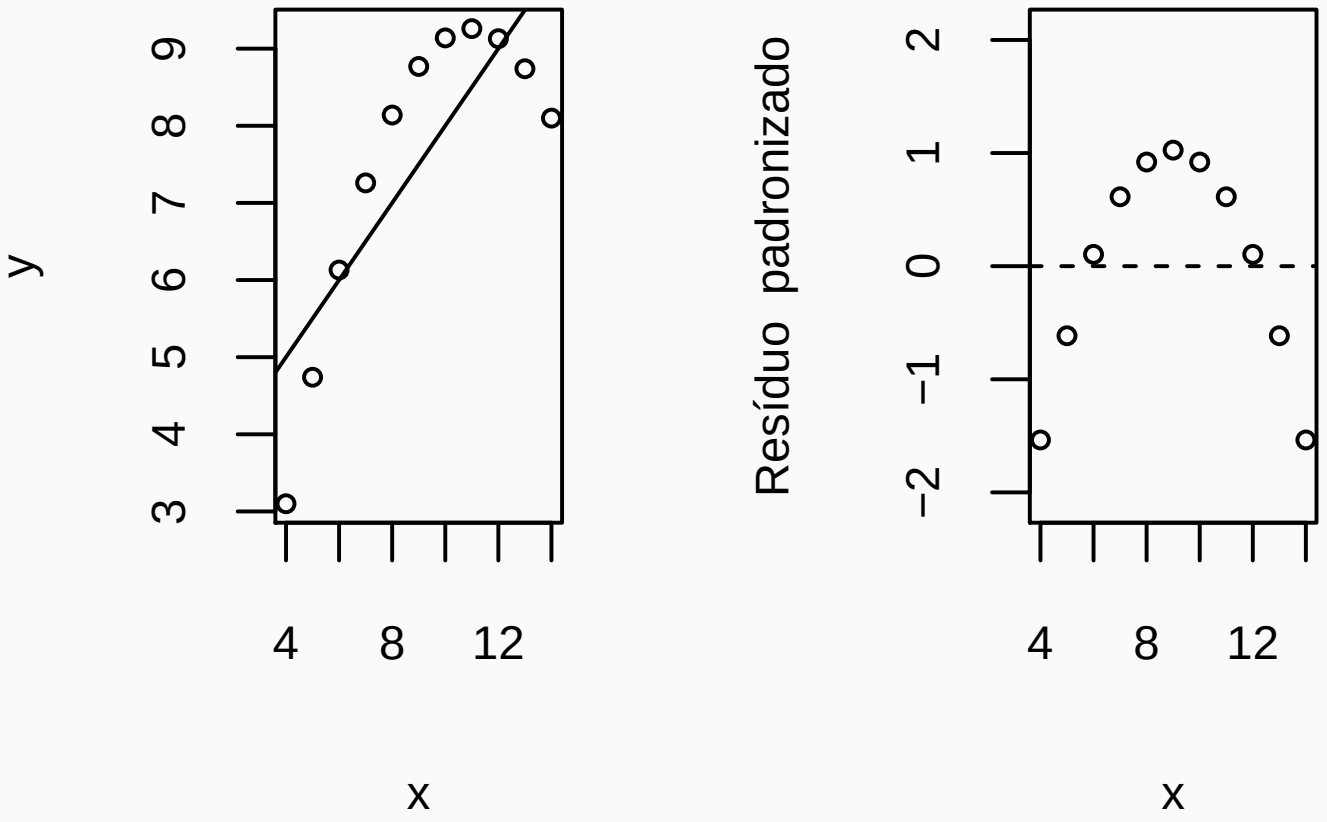


Figura 8: Dados de Anscombe II

Exemplo: Dados de Anscombe III

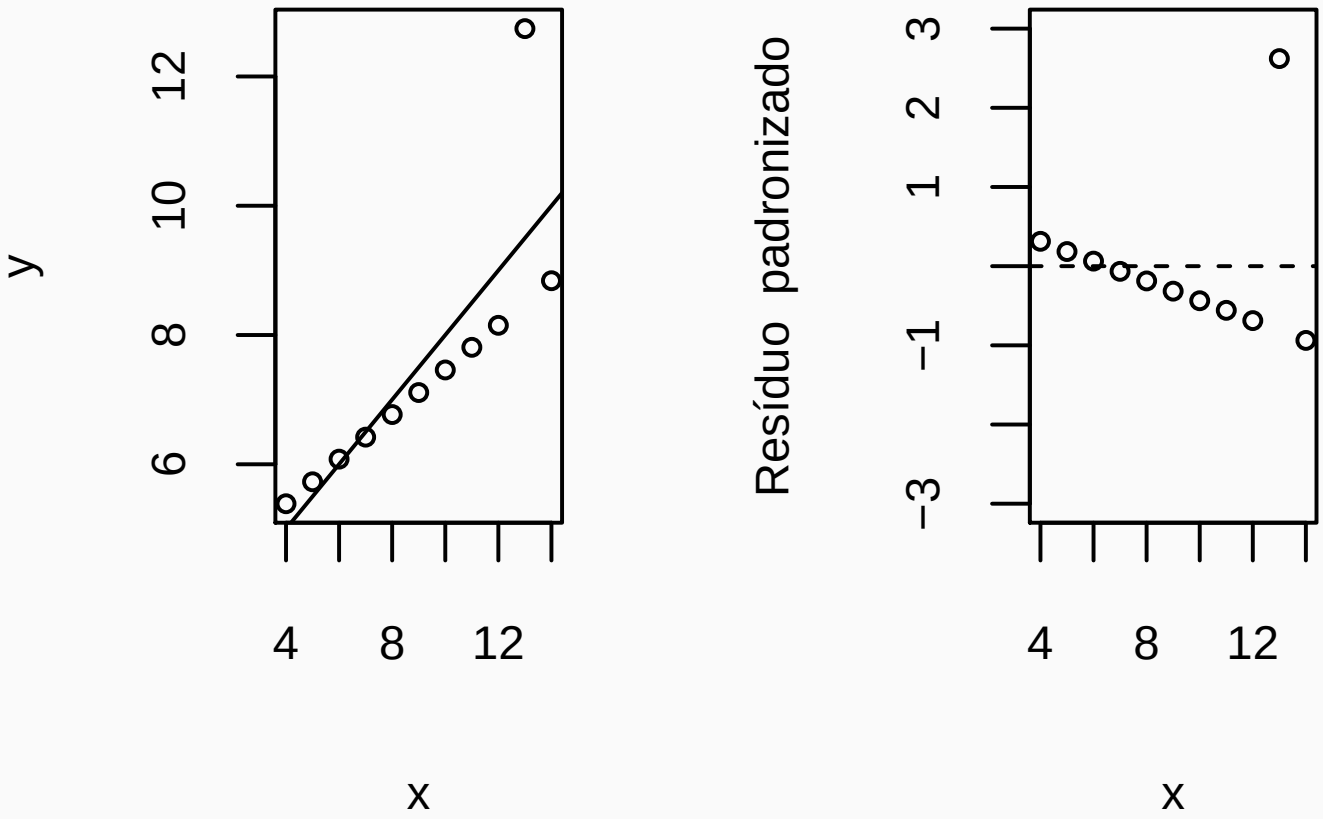


Figura 9: Dados de Anscombe III

Exemplo: Dados de Anscombe

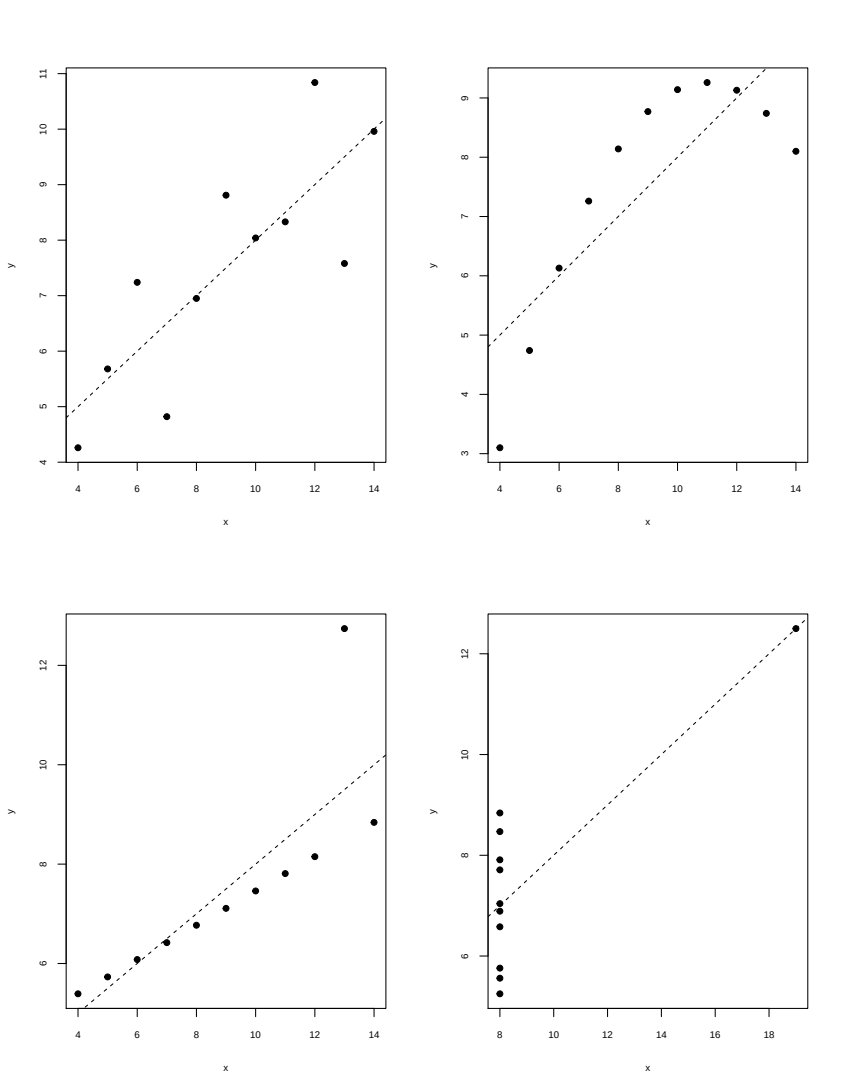


Figura 10: Dados de Anscombe. Reta estimada: $\hat{\mu}(x) = 3.00 + 0.50x$, $r = 0.82$

Exemplo

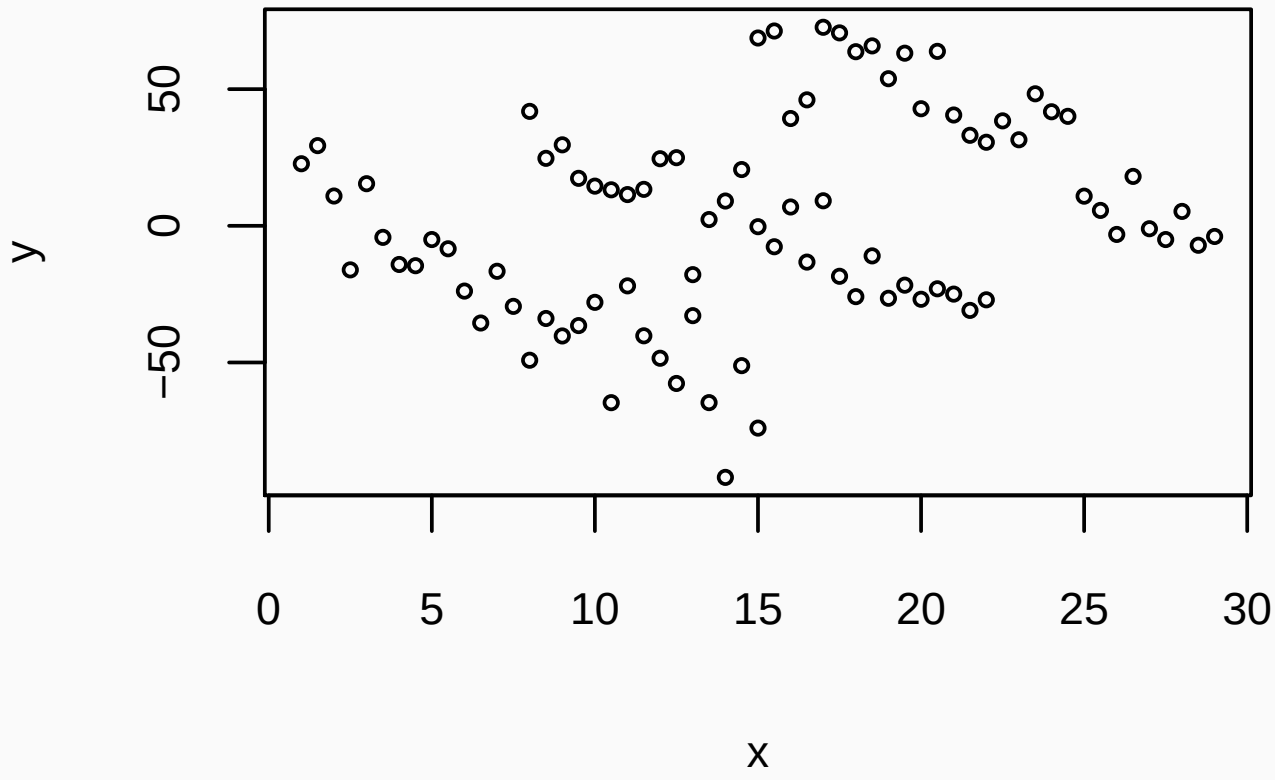


Figura 11: Dados simulados

Exemplo

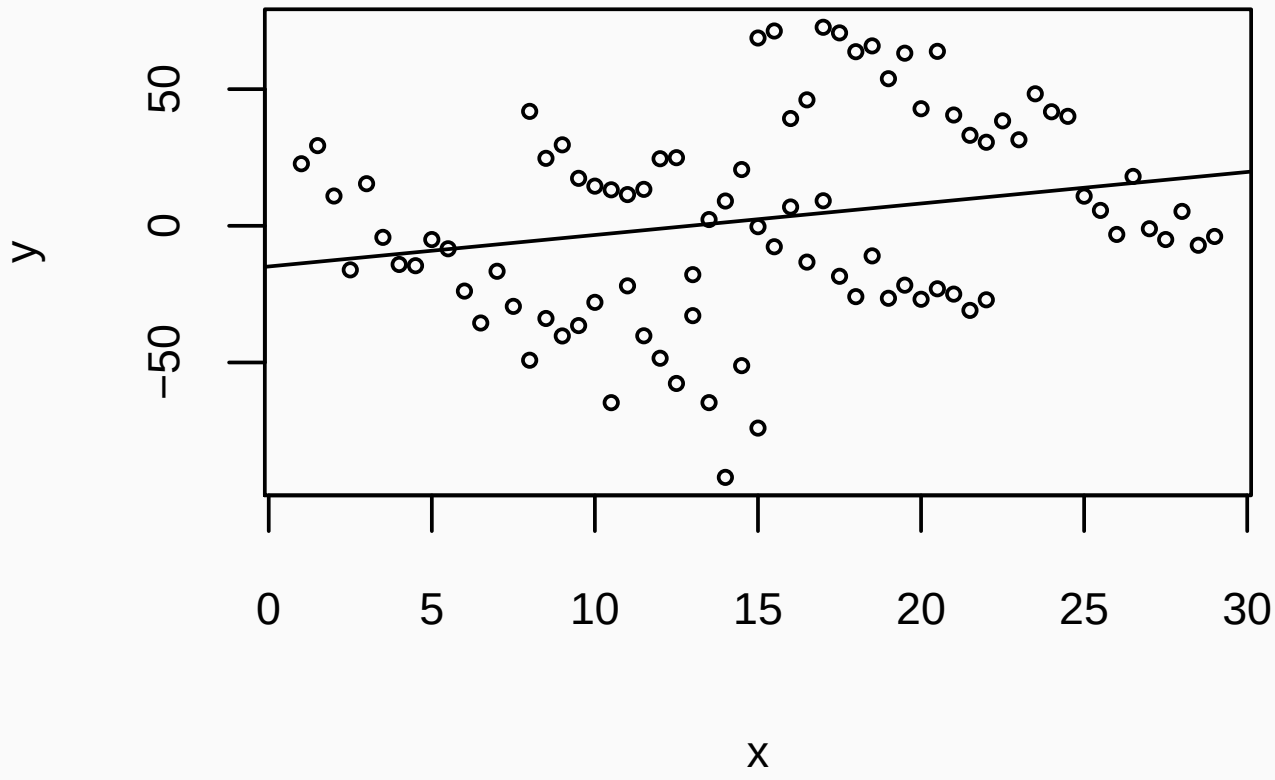


Figura 12: Dados simulados e reta de mínimos quadrados

Exemplo

Confundimento ou viés por omissão de variável

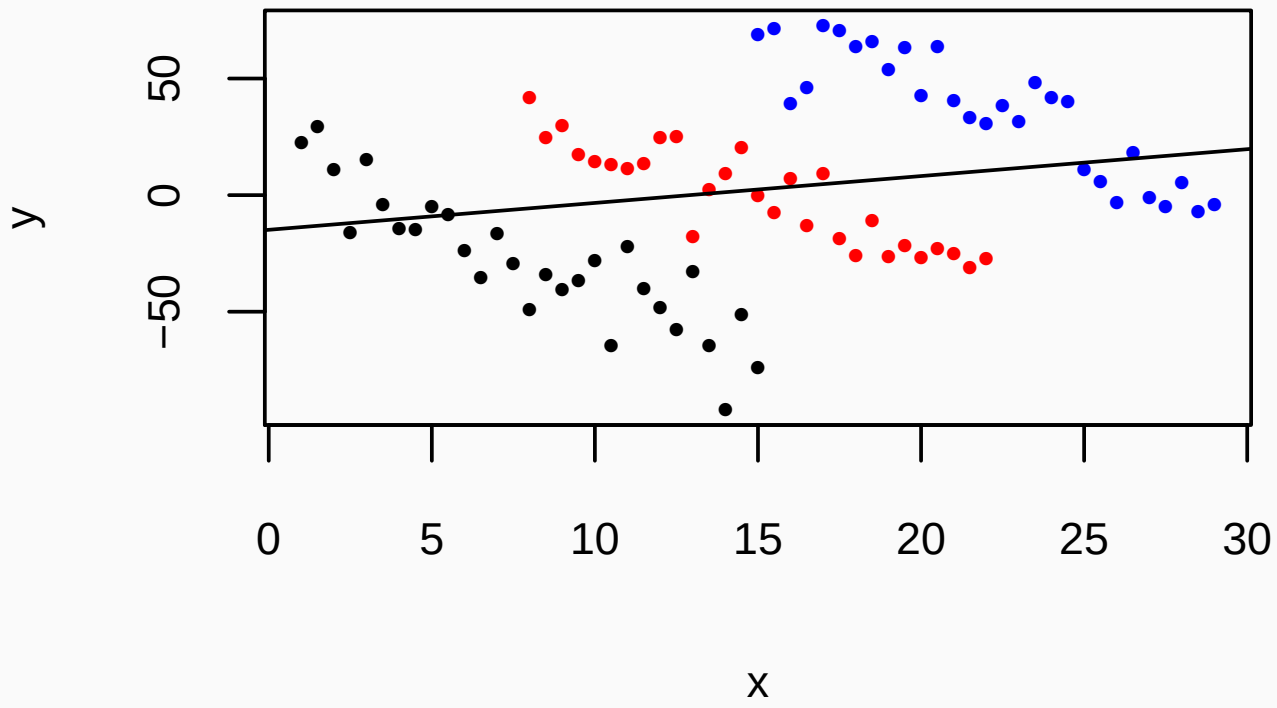


Figura 13: Dados simulados

Exemplo 2

- Estudo observacional
- Interessa explicar o **salário** em termos da **educação**
- **salário** é medido em unidades monetárias e **educação** em anos de estudo
- Fonte: Dougherty

Exemplo 2

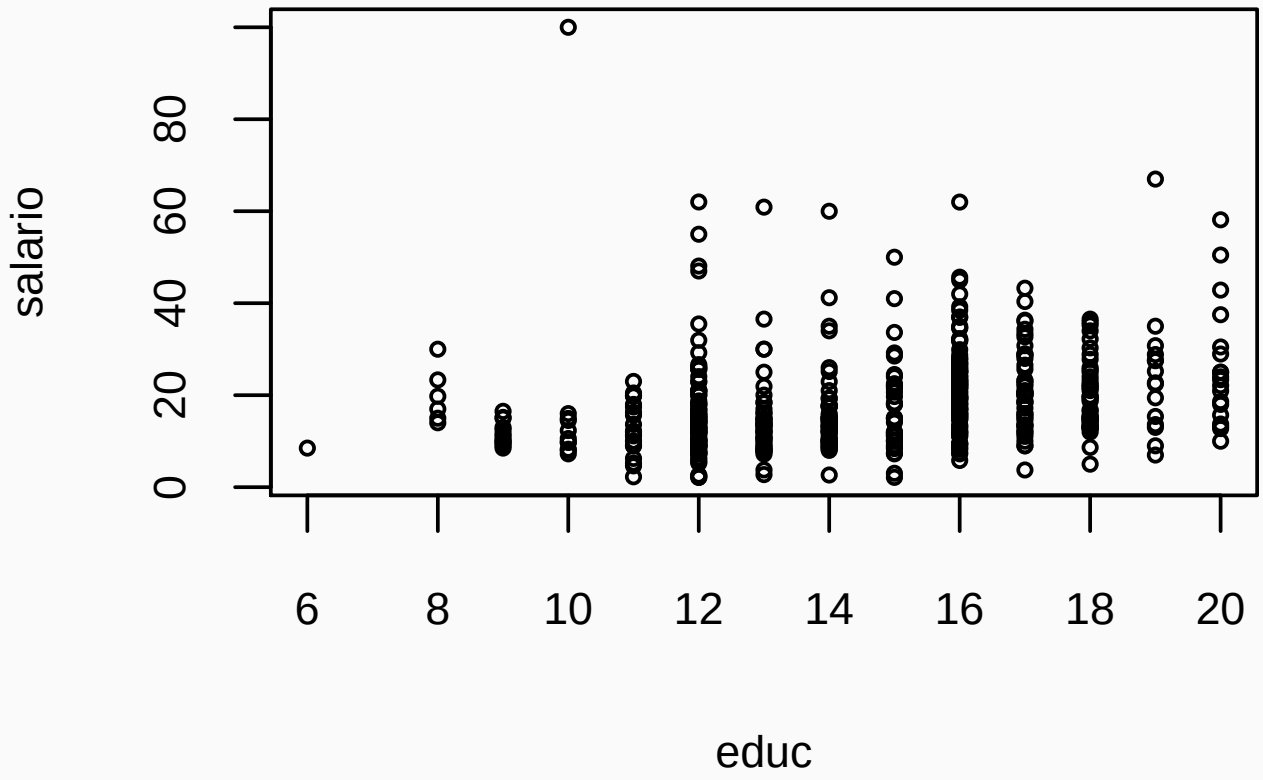


Figura 14: Dados do Exemplo 2

Exemplo 2

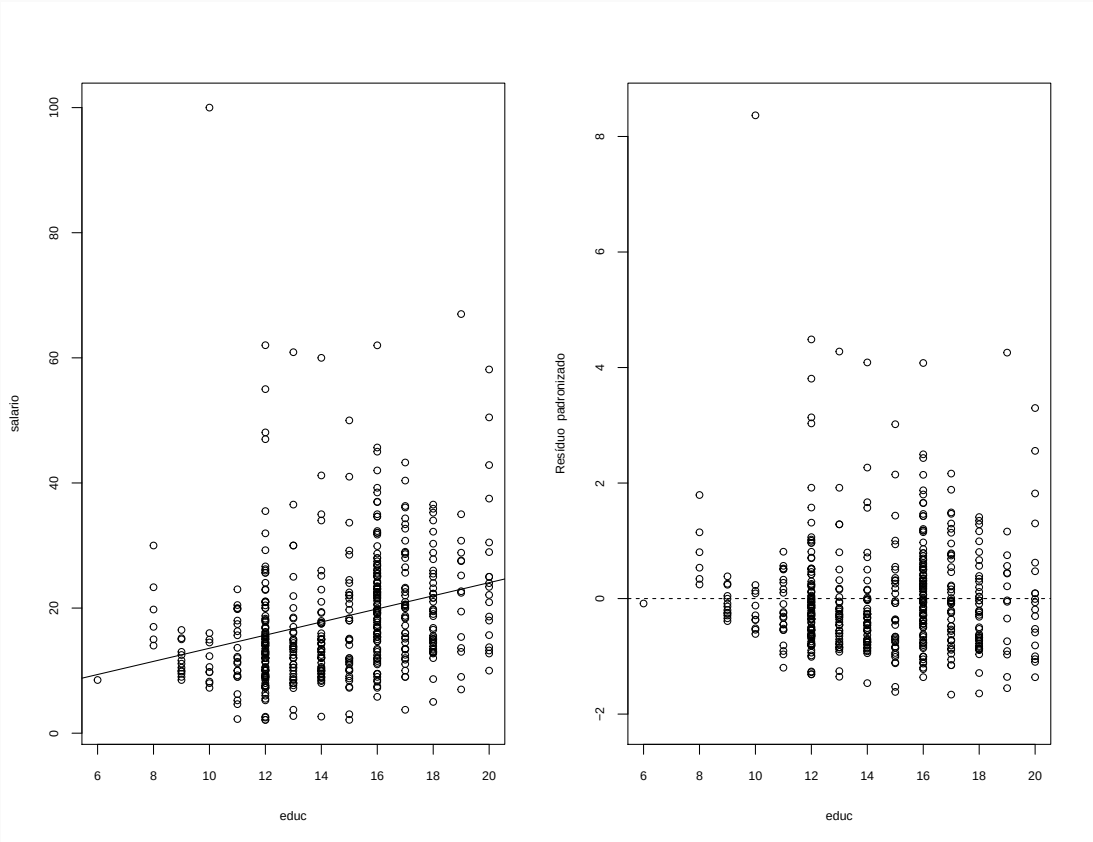


Figura 15: Exemplo 2. Dados e gráfico de resíduos padronizados (à direita)

Exemplo 2

- Presença de heteroscedasticidade
- Será aplicada uma transformação na variável resposta. Usando o logaritmo obtemos interpretabilidade.

Exemplo 2

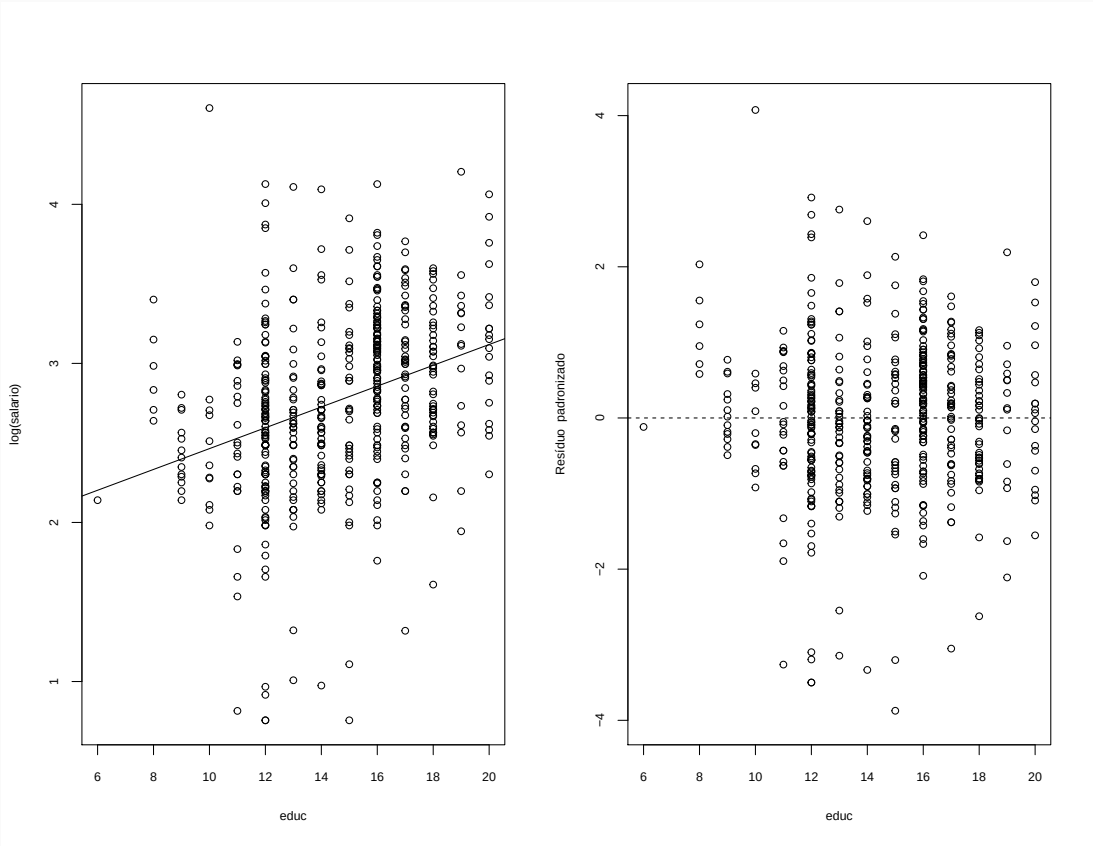


Figura 16: Dados

Exemplo 2

- $Y = \log(salario)$, $X = educ$
- Estimativas de mínimos quadrados, erros padrões e t-ratios

$$\hat{\alpha} = 1.810 \quad ep(\hat{\alpha}) = 0.125 \quad t = 14.4 \tag{45}$$

$$\hat{\beta} = 0.065 \quad ep(\hat{\beta}) = 0.008 \quad t = 5.8 \tag{46}$$

- O salario médio aumenta em $100 \times 0.065\% = 6.5\%$ por cada ano adicional de educação
- $R^2 = 0.11$
- Necessidade de incluir outras variáveis explanatórias: sexo, experiência, ...

Regressão Múltipla

Referencias

Dougherty, C. (2002). Introduction to Econometrics. Second edition.

Rice, J.A. (1995). Mathematical Statistics and Data Analysis. Second edition.

Ross, S. M. (2010). Introductory Statistics.