Análise de Resíduos

Profª Silvia Nagib Elian
Sala 215 - Bloco A

Instituto de Matemática e Estatística
Universidade de São Paulo

Agenda

1. Análise gráfica
2. Verificação das suposições
3. Outliers
4. Outros gráficos
5. Teste de Levene para igualdade de variâncias

Análise de Resíduos

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i = y_i - (\alpha + \beta x_i) \rightarrow \text{erro}$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \qquad e_i = y_i - \hat{y}_i \rightarrow \text{resíduo}$$

e_i: quantidade que a equação de regressão não consegue explicar

- efeito de variáveis externas (variáveis explicativas omitidas).
- variabiliade natural entre indivíduos.
- eventuais erros de medida na variável Y.

Suposições do M.R.L.S.: $\epsilon_i \sim N(0, \sigma^2)$ independentes.

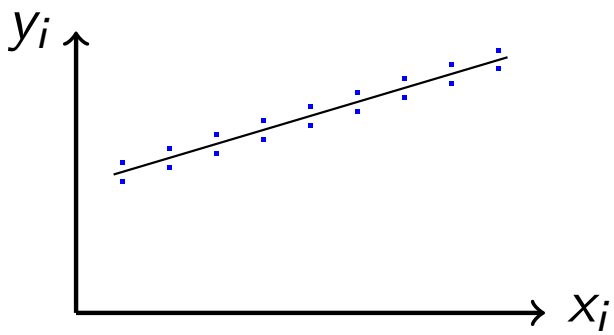
Suposições corretas: $\Rightarrow$ e_i devem apresentar evidências de modo a confirmar ou pelo menos não rejeitar as suposições.

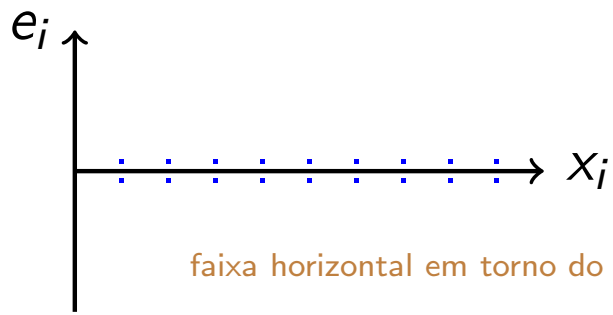
$\sum e_i = 0$ para todo modelo linear com intercepto.

(1ª equação normal: $-2\sum(y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$)

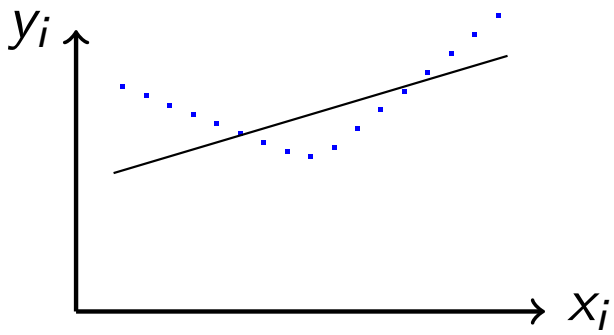
$\sum e_i \neq 0 \Rightarrow$ erro conta ou problema de aproximação.

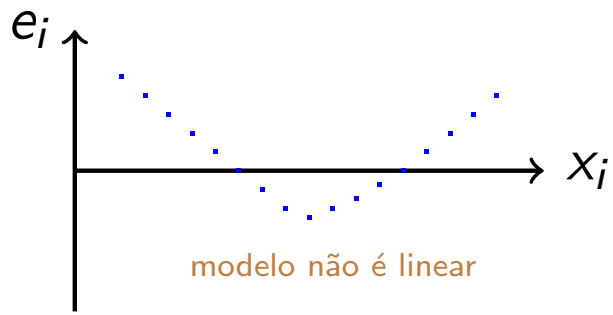
Exemplos



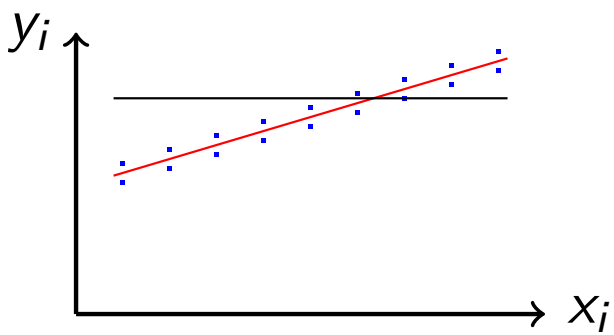


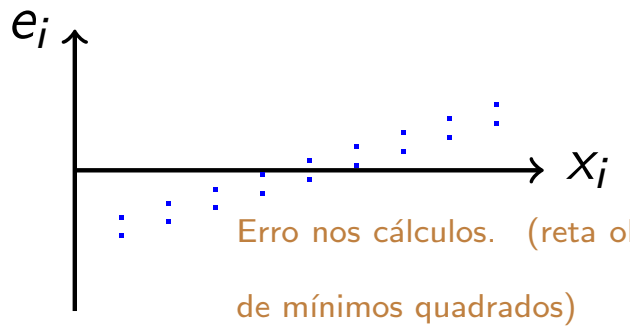
faixa horizontal em torno do zero





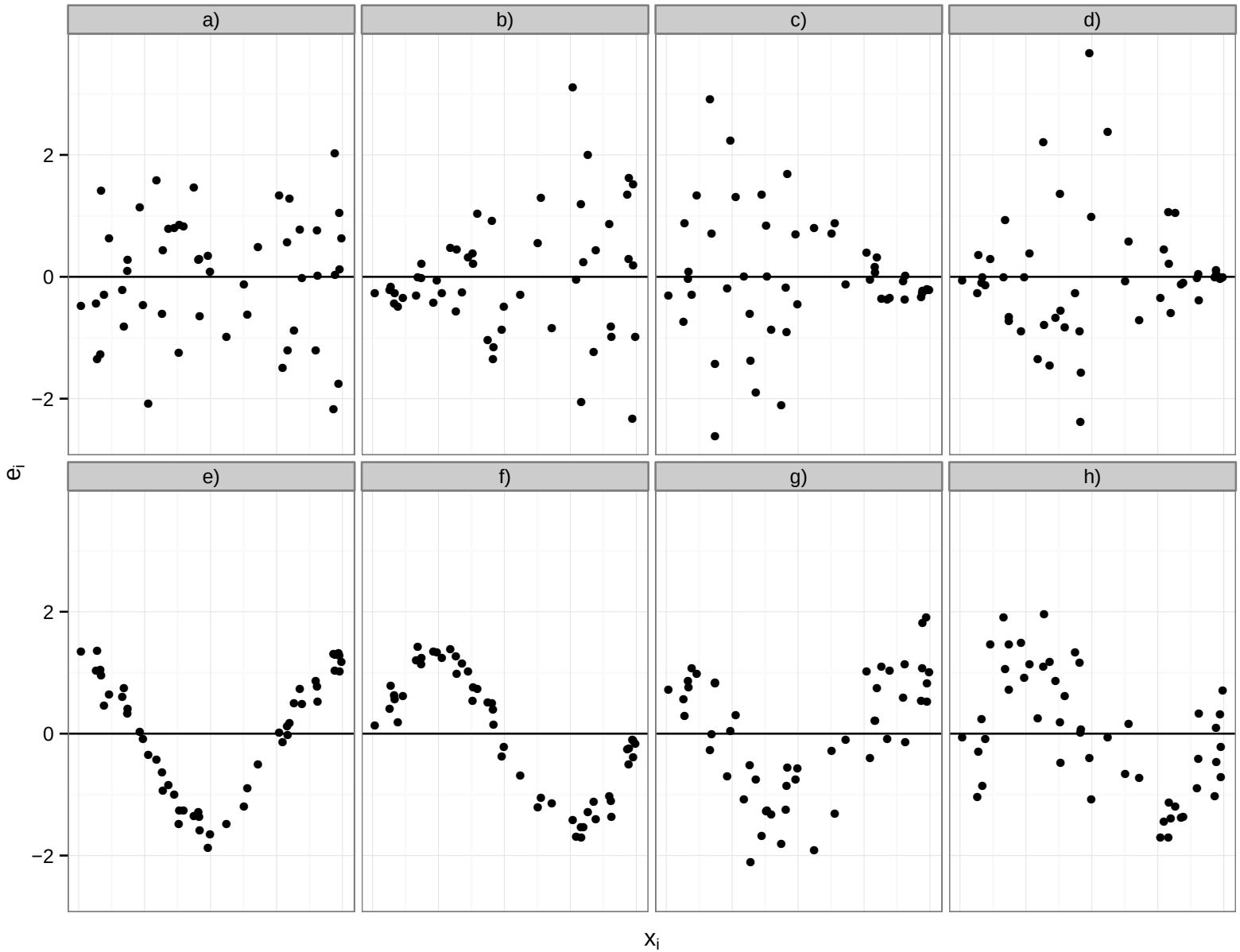
modelo não é linear



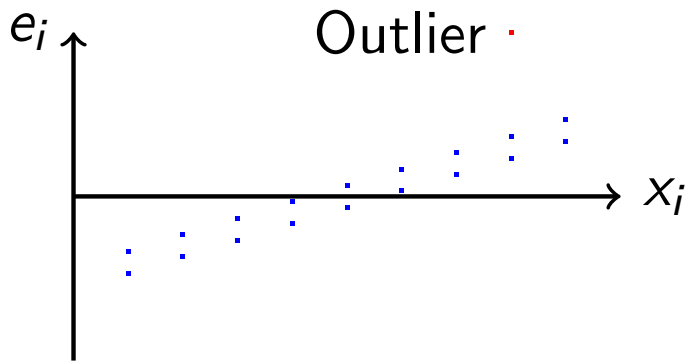


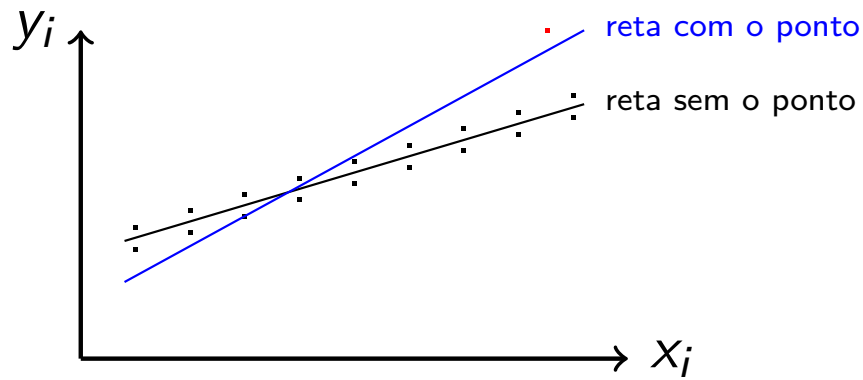
Erro nos cálculos. (reta obtida não é de mínimos quadrados)

Possíveis gráficos de resíduos



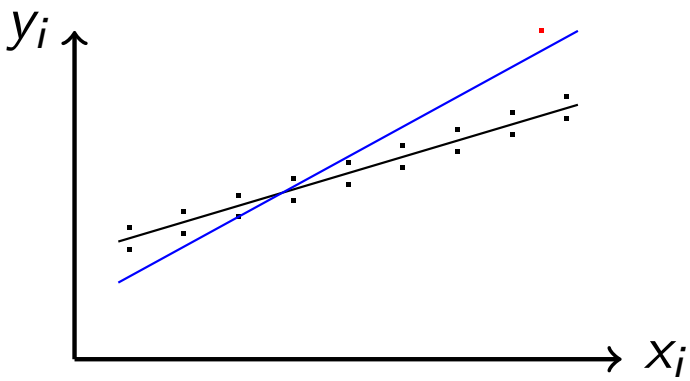
Efeito de outlier no ajuste

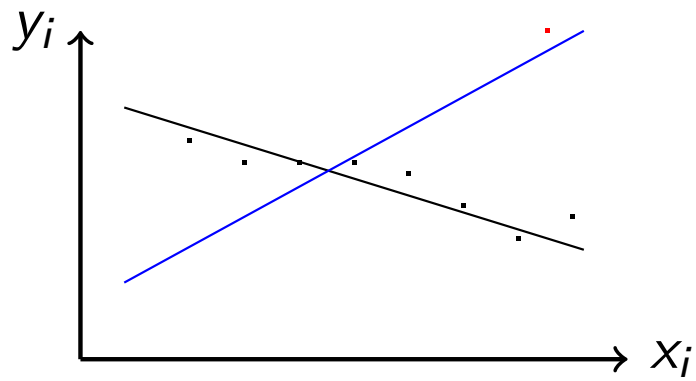




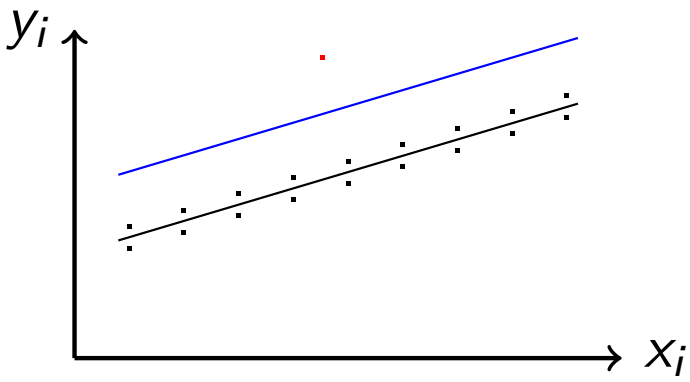
A presença de outlier pode modificar muito a equação obtida. Essa influência é tanto maior quanto mais distante estiver sua abcissa de $\bar{x}$.

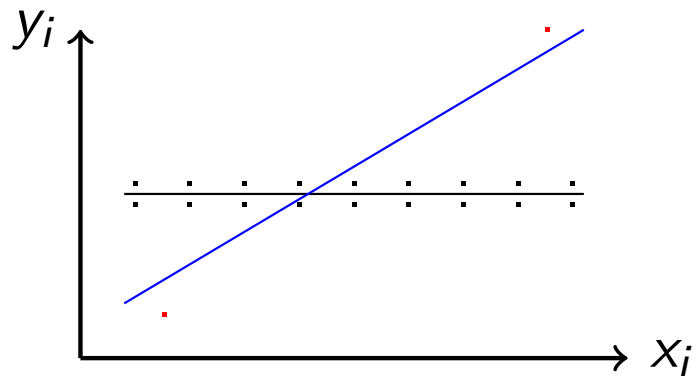
Efeito de outlier no ajuste





Abscissa do outlier distante de $\bar{x}$, alteração na inclinação e intercepto.

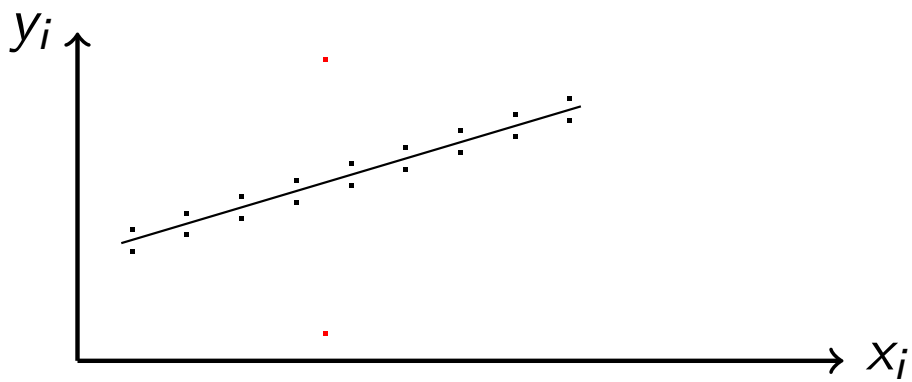


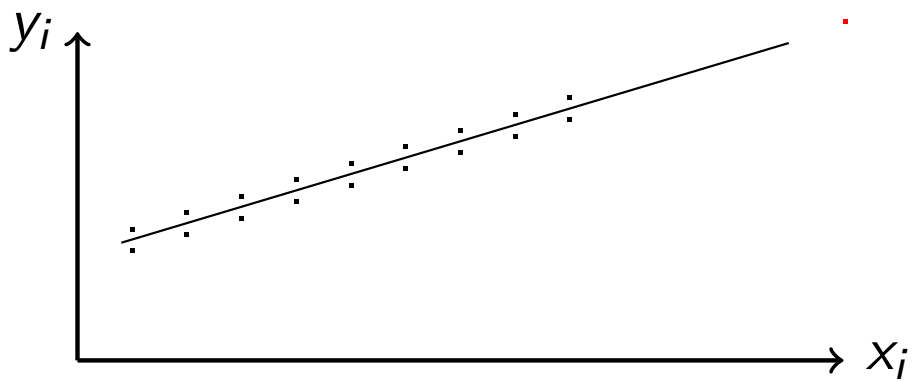


Abscissa do outlier mais próximo de $\bar{x}$, altera o intercepto, mantém a inclinação.;

Os outliers provocaram a inclinação.

Outliers que não provocam alteração





Comentários sobre os resíduos

Se o modelo é adequado:

a) Cada e_i deve ser próximo de zero.

b) Aproximadamente $\begin{cases} n/2 \text{ devem ser positivos} \\ n/2 \text{ devem ser negativos} \end{cases}$

c) e_i's não devem produzir sequências muito longas de valores positivos ou negativos

1. - - - - + + + + (embora b) seja válida) não indica linearidade.
   Pode indicar também resíduos positivamente correlacionados
2. + - + - + - + - + - é indicativo de não aleatoriedade. Pode indicar também correlação negativa entre os resíduos.

Verificação das suposições

1.Homocedasticidade.

- Se $\text{Var}(\epsilon_i|x_i) = \sigma^2, \forall i$, os resíduos devem se distribuir numa faixa horizontal em torno do zero. (Verif. gráfica)
- Testes de hipóteses.
  - ◇ Teste de Bartlett
  - ◇ Teste da Razão de Verossimilhança.
  - ◇ Outros: Weisberg pág. 135.

Soluções para a heterocedasticidade:

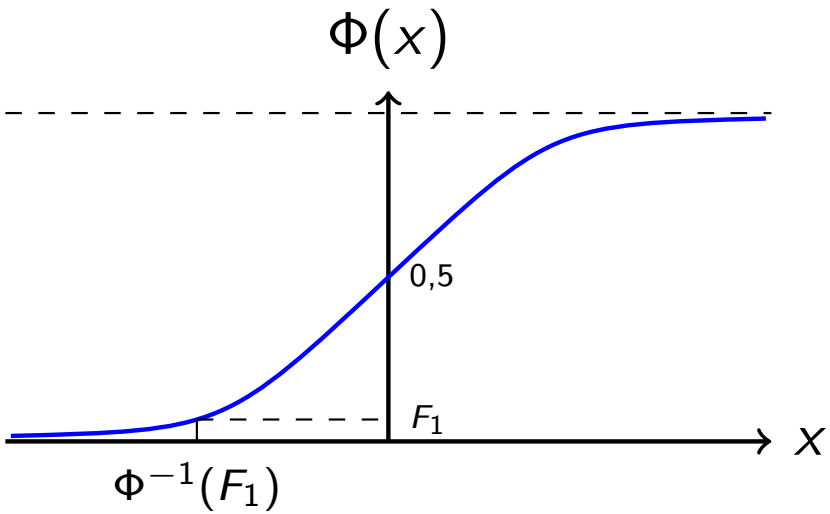
- Transformação de variáveis.
- Procedimento de mínimos quadrados ponderados.

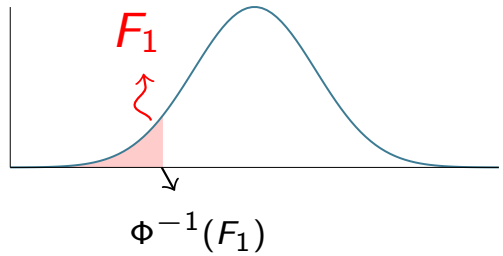
2. Normalidade $\epsilon_i \sim N(0, \sigma^2)$

- Histograma dos resíduos (se n grande)
- $z_i = \dfrac{e_i}{\sqrt{\text{MSE}}} \approx N(0,1)$ resíduo padronizado. $\pm$ 95% dos z_i's devem estar no intervalo]-1,96;1,96[.

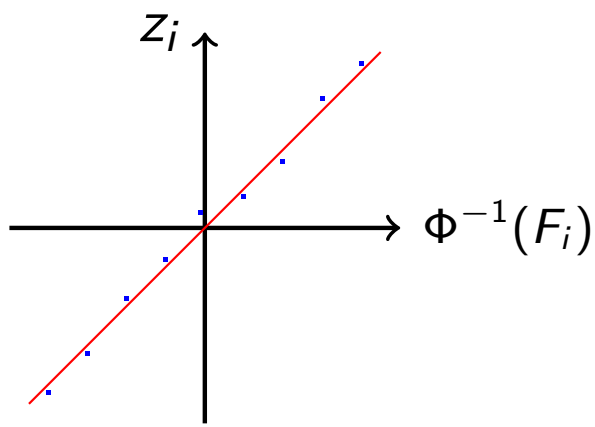
Normal probability plot

$$z_i = \frac{e_i}{\sqrt{\text{MSE}}}, \quad F_i = \frac{\#(Z \leqslant z_i)}{n}, \quad \Phi(x) = P(Z \leqslant x),\ Z \sim N(0,1).$$





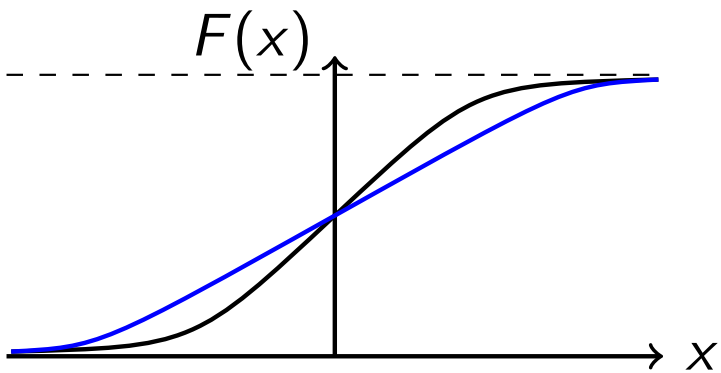
z_i ord.	F_i	$\Phi^{-1}(F_i)$
z_1	1/n	$\Phi^{-1}(F_1)$
z_2	2/n	$\Phi^{-1}(F_2)$
$\vdots$	$\vdots$	$\vdots$
z_n	1	



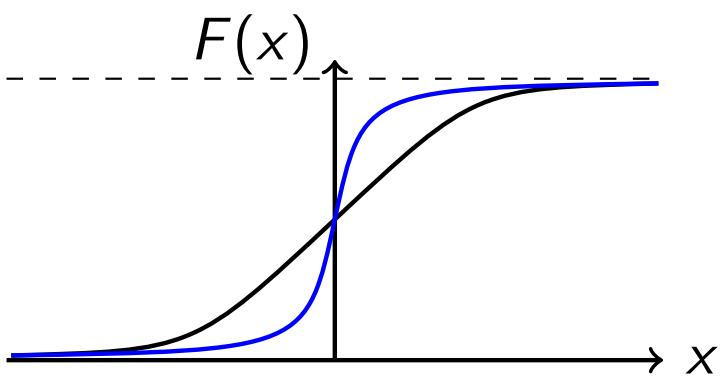
Obs: Para verificar se $W \sim N(\mu, \sigma^2)$

w_i ord.	F_i	$\Phi^{-1}(F_i)$
w_1	$1/n$	$\Phi^{-1}(F_1)$
w_2	$2/n$	$\Phi^{-1}(F_2)$
$\vdots$	$\vdots$	$\vdots$
z_n	1	

Os pontos do gráfico $W_i \times \Phi^{-1}(F_i)$ devem estar próximos da reta $y = \mu + \sigma x$.

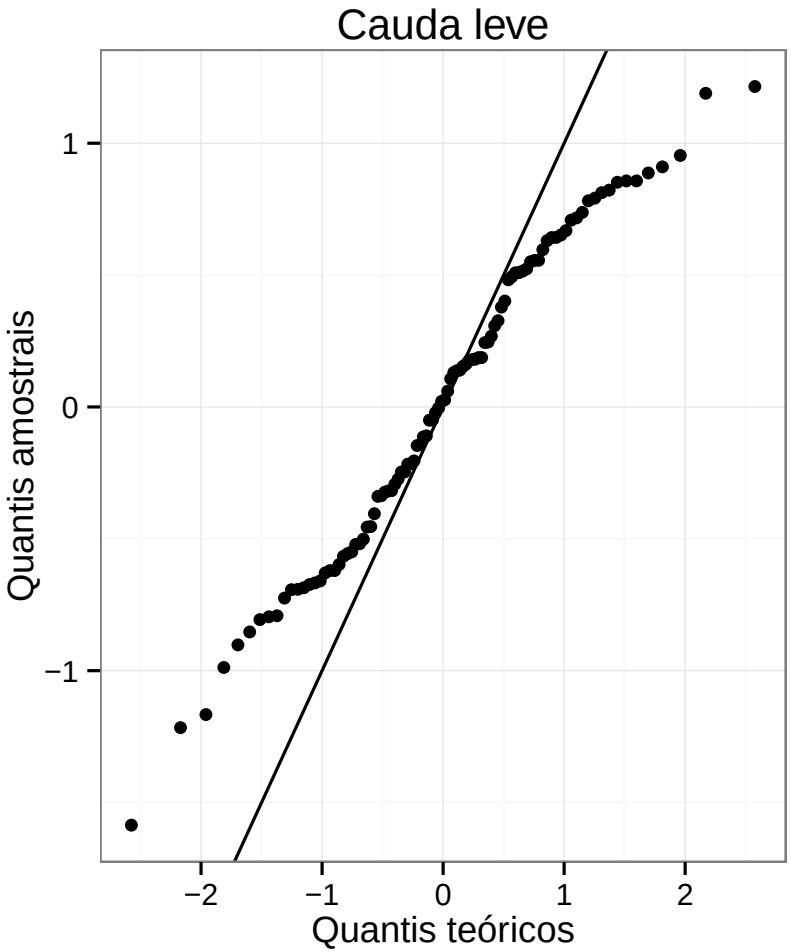


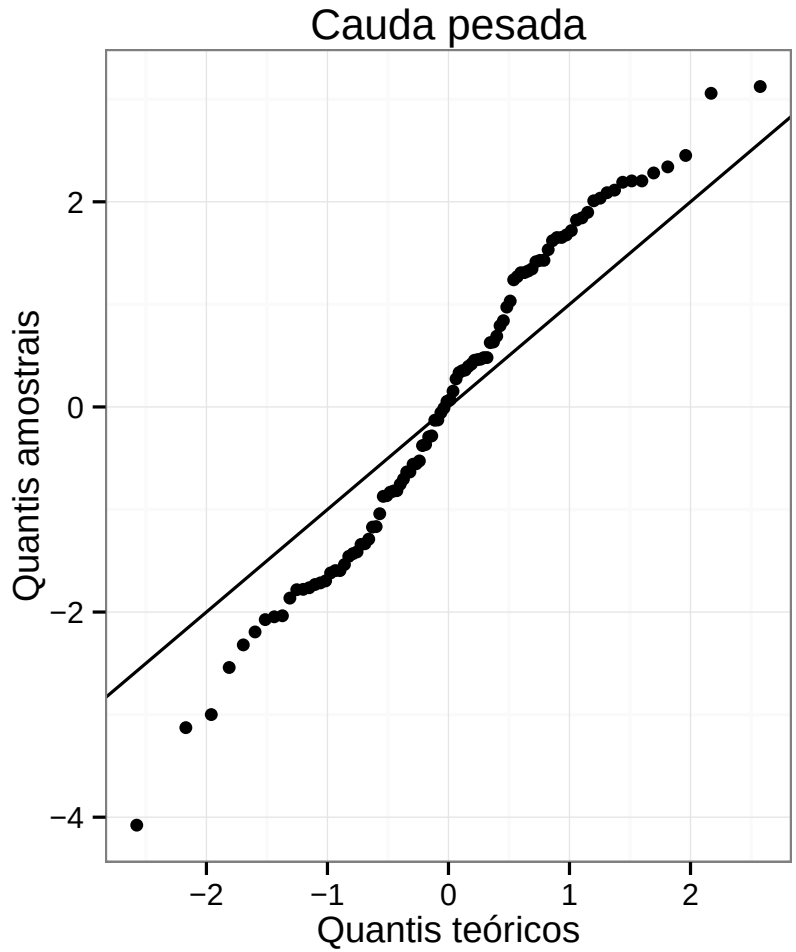
Cauda mais pesada que a normal.
(heavy tailed distribution)



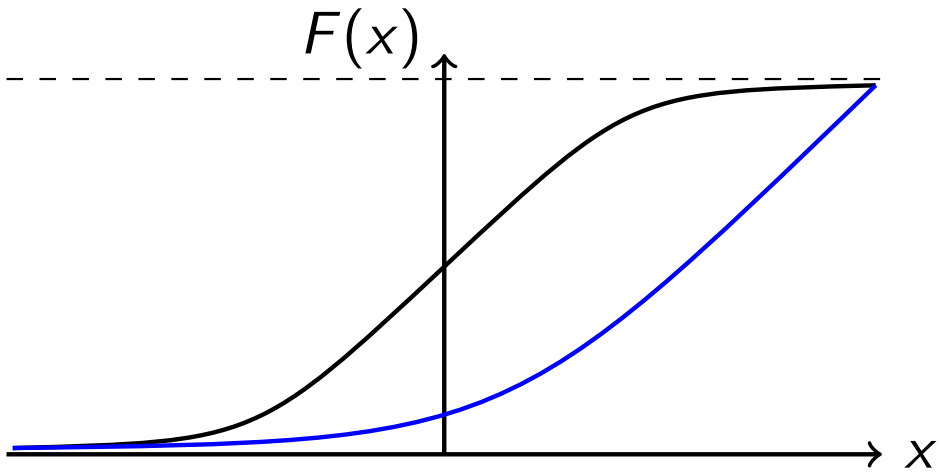
Cauda mais leve que a normal.
(light tailed distribution)

Gráfico QQ-plot para análise da normalidade dos erros

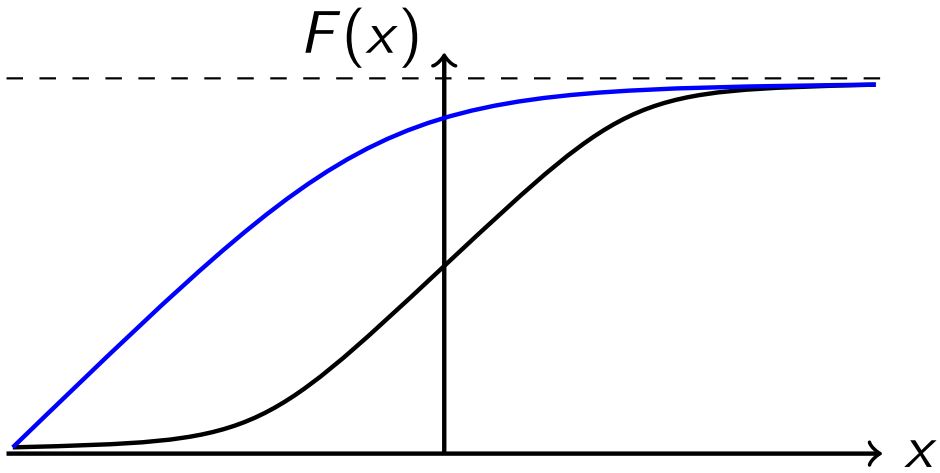




Exemplos de assimetria



Assimetria positiva



Assimetria negativa

Teste de Shapiro-Wilk

Calcula o quadrado do coeficiente de correlação entre

Φ^{-1} e z_i

Φ^{-1} e w_i

Normalidade é rejeita para pequenos valores do coeficiente de correlação.

3. Independência dos erros

Teste de Durbin-Watson

Solução:

Procedimento de mínimos quadrados generalizados.

Outliers

Resíduo extremo - muito maior em valor absoluto que os restantes.

a) erro medida - podem ser eliminados.

b) caso contrário - devem ser investigados.

Podem produzir informação.

Testes para detectar outliers:

- Seção 3.3 - Montgomery and Peck
- Seção 5.2 - Weisberg

Solução:

- Análises separadas, com e sem outlier.
- Uso de procedimentos "robustos" (menos sensíveis a "outliers", ex: regressão L1)

Outros gráficos

a) $e_i \times \hat{y}_i$

No modelo de regressão linear simples: mesma informação que $e_i \times x_i$. Importante no modelo de regressão linear múltipla.

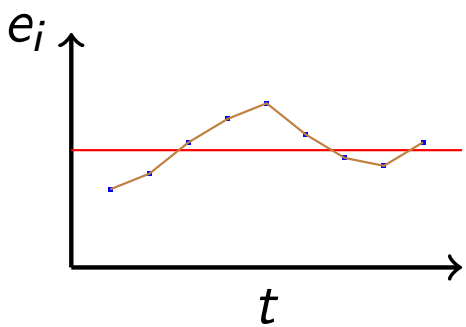
b) $e_i \times y_i$

Desaconselhável. Verifica-se que e_i e y_i são correlacionadas.

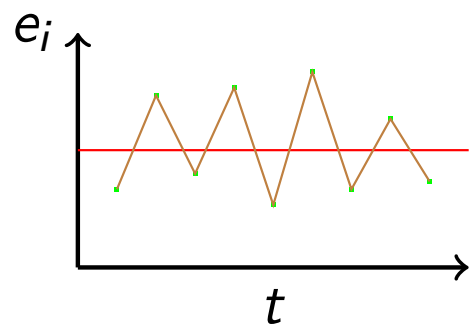
c) $e_i \times$ tempo

Se os dados forem tomados numa ordem "temporal" conhecida.

Comum: erros (ϵ_i) num período de tempo serem correlacionados com os do período seguinte.



autocorrelação positiva



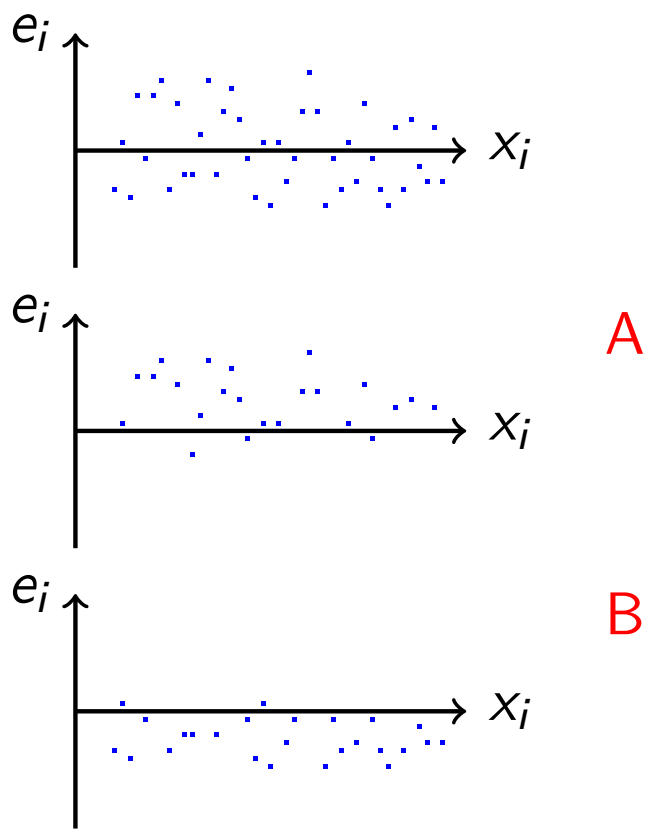
autocorrelação negativa

d) $e_i \times$ valores de uma variável independente omitida.

Qualquer "padrão" exibido por este gráfico indica que o modelo pode ser melhorado com a inclusão desta variável independente.

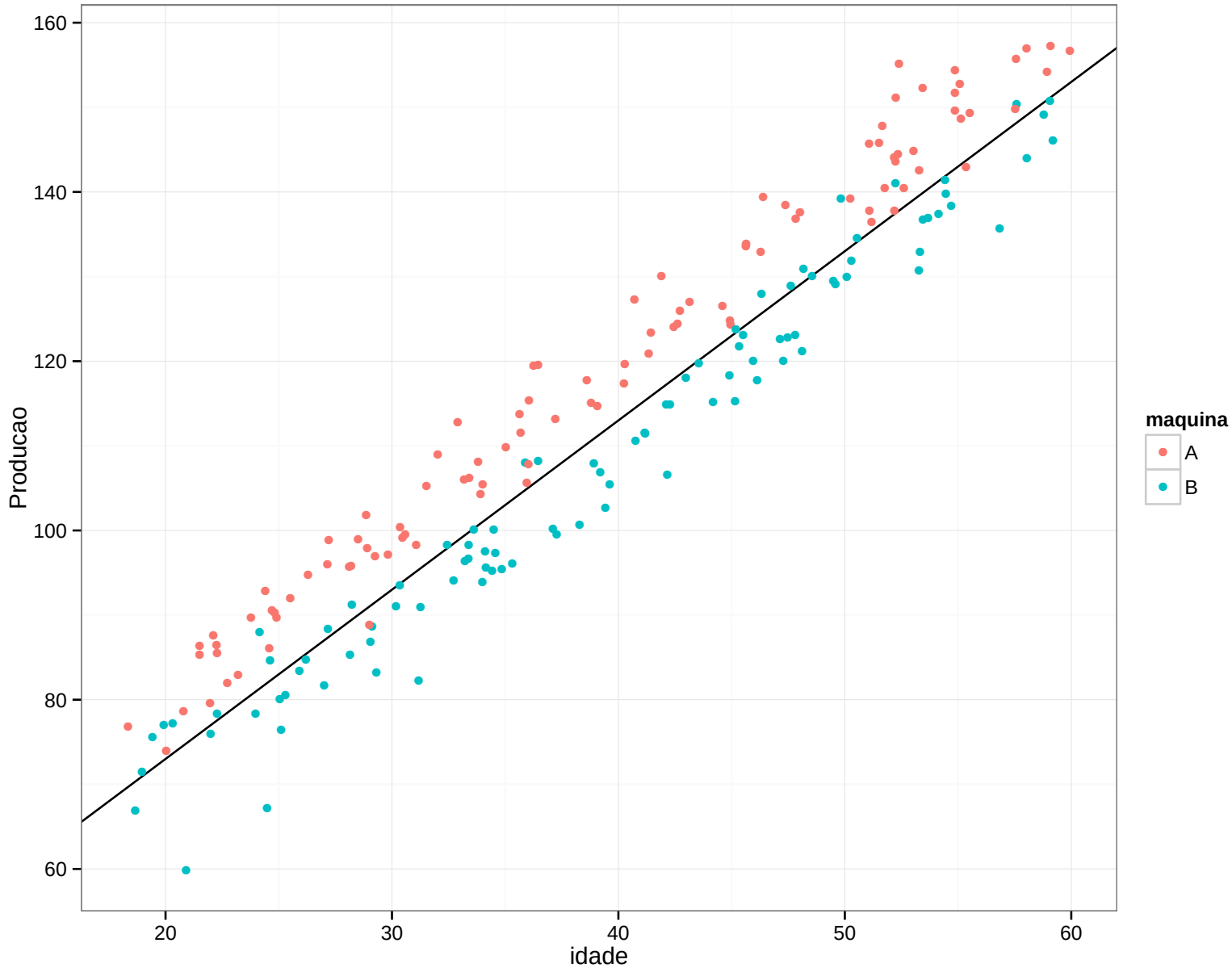
Ex: Y - Produção, X - idade do operário, W - máquina A e B

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \rightarrow e$



Introdução da variável independente máquina pode melhorar o modelo.

Exemplo variável omitida



Anscombe (1973) - Graphs in statistical analysis

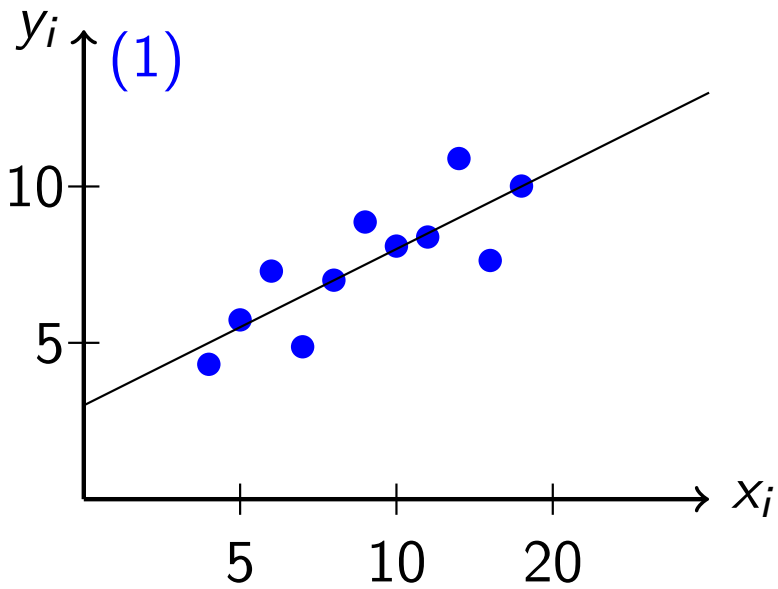
The American Statistician, vol 27, nº 1, pág. 17-21

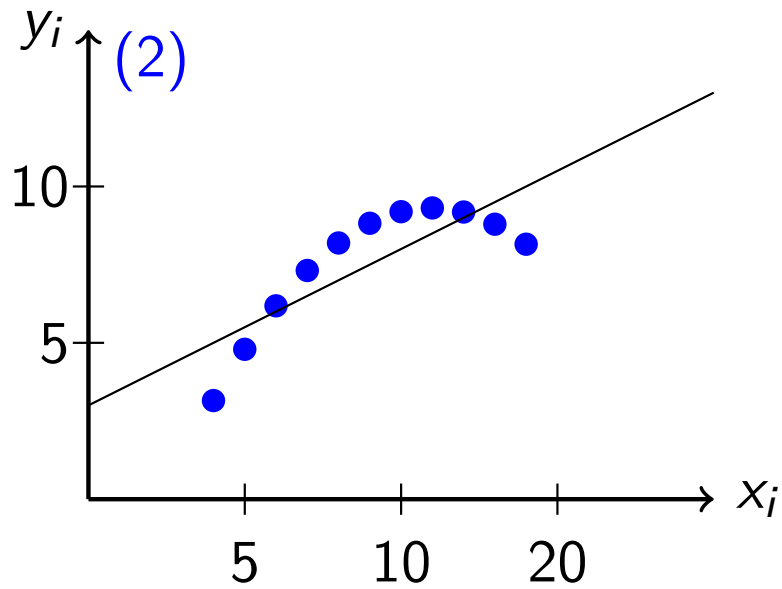
Banco de dados	1-3	1	2	3	4	4
Variável	x	y	y	y	x	y
Obs. nº 1:	10	8.04	9.14	7.46	8	6.58
2:	8	6.95	8.14	6.77	8	5.76
3:	13	7.58	8.74	12.74	8	7.71
4:	9	8.81	8.77	7.11	8	8.84
5:	11	8.33	9.26	7.81	8	8.47
6:	14	9.96	8.10	8.84	8	7.04
7:	6	7.24	6.13	6;08	8	5.25
8:	4	4.26	3.10	5.39	19	12.50
9:	12	10.84	9.13	8.15	8	5.56
10:	7	4.82	7.26	6.42	8	7.91
11:	5	5.68	4.74	5.73	8	6.89

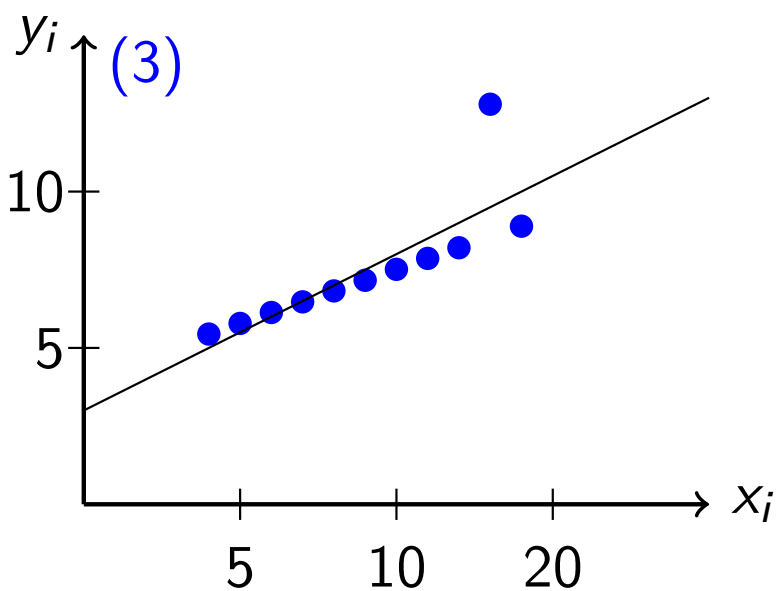
Número de observações (n) $= 11$
Média dos x's $(\bar{x}) = 9.0$
Média dos y's $(\bar{y}) = 7.5$
Coeficiente de regressão $(\hat{\beta}_1) = 0.5$
Eq. da reta de regressão: $\hat{y} = 3 + 0.5x$

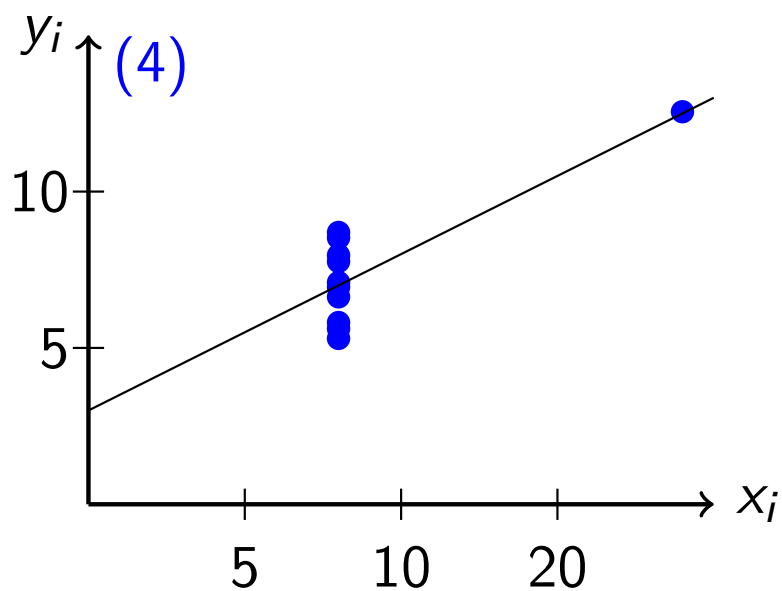
Soma de quadrados $x - \bar{x} = 110.0$
Reg. soma de quadrados $= 27.5$ (1 GL)
Res. soma de quadrados de $y = 13.75$ (9 GL)
Erro padrão estimado de $\hat{\beta}_1 = 0.118$
R^2 múltiplo $= 0.667$

Gráficos dos dados









Teste de Levene para igualdade de variâncias

(Vantagem: robusto com relação a sérios desvios da normalidade)

Considere k grupos de réplicas ($k \neq s$ valores da variável X) e a variável

$$z_{ij} = |y_{ij} - \bar{y}_i| \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k.$$

O teste é baseado na composição das médias das variáveis Z_{ij}, ou seja, se $E(z_{ij}) = \mu$, $i = 1, 2, \ldots, k$, então

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

com estatística de teste

$$F = \frac{\dfrac{1}{(k-1)} \displaystyle\sum_{i=1}^{k} n_i (\bar{z}_i - \bar{z})^2}{\dfrac{1}{(n-k)} \displaystyle\sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2}, \quad \bar{z} = \sum_i \sum_j z_{ij}/n, \ \bar{z}_i = \sum_j z_{ij}/n_i$$

Rejeita-se a igualdade das médias dos z_i, concluindo-se pela igualdade de variâncias, se $F_{obs} \geqslant F_c[k-1, n-k, \alpha]$.

Alternativa do teste de Levene

Uma alternativa desse teste é o teste de Levene usando medianas. A estatística de teste é a mesma, mas

$$z_{ij} = |y_{ij} - \tilde{y}_i| \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k.$$

onde $\tilde{y}_i$ é a mediana de $\{y_{i1}, y_{i2}, \ldots, y_{in_i}\}$

No caso, aplica-se o teste à variável resíduo, de modo que

$$z_{ij} = |e_{ij} - \tilde{e}_i| \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k.$$

Tais testes são necessários porque a desigualdade de variâncias e não normalidade geralmente "caminham juntas".