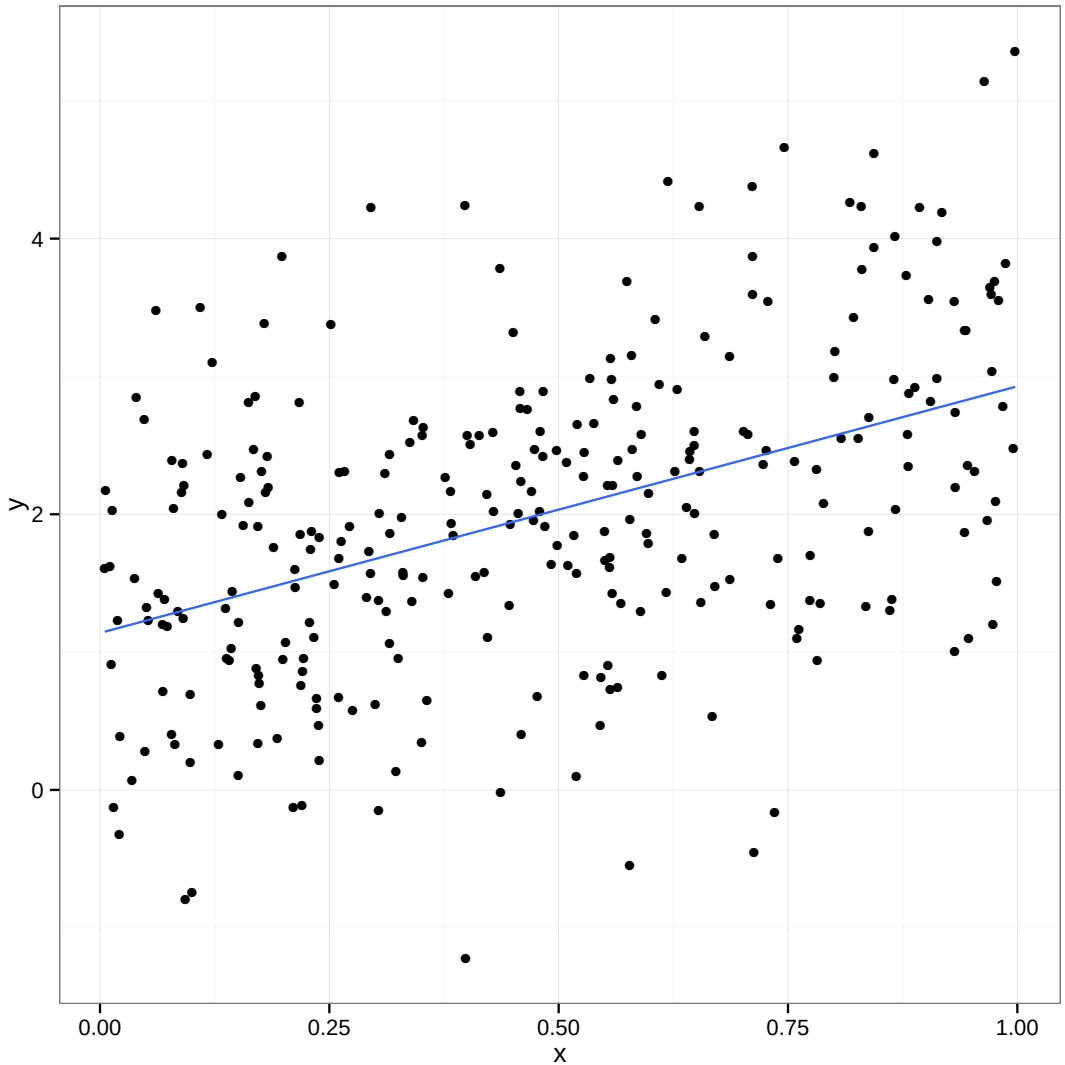
Métodos Estatísticos

Prof. Bruno Santos

Instituto de Matemática e Estatística
Universidade Federal da Bahia

Resumindo



Modelos

Modelo de regressão linear simples

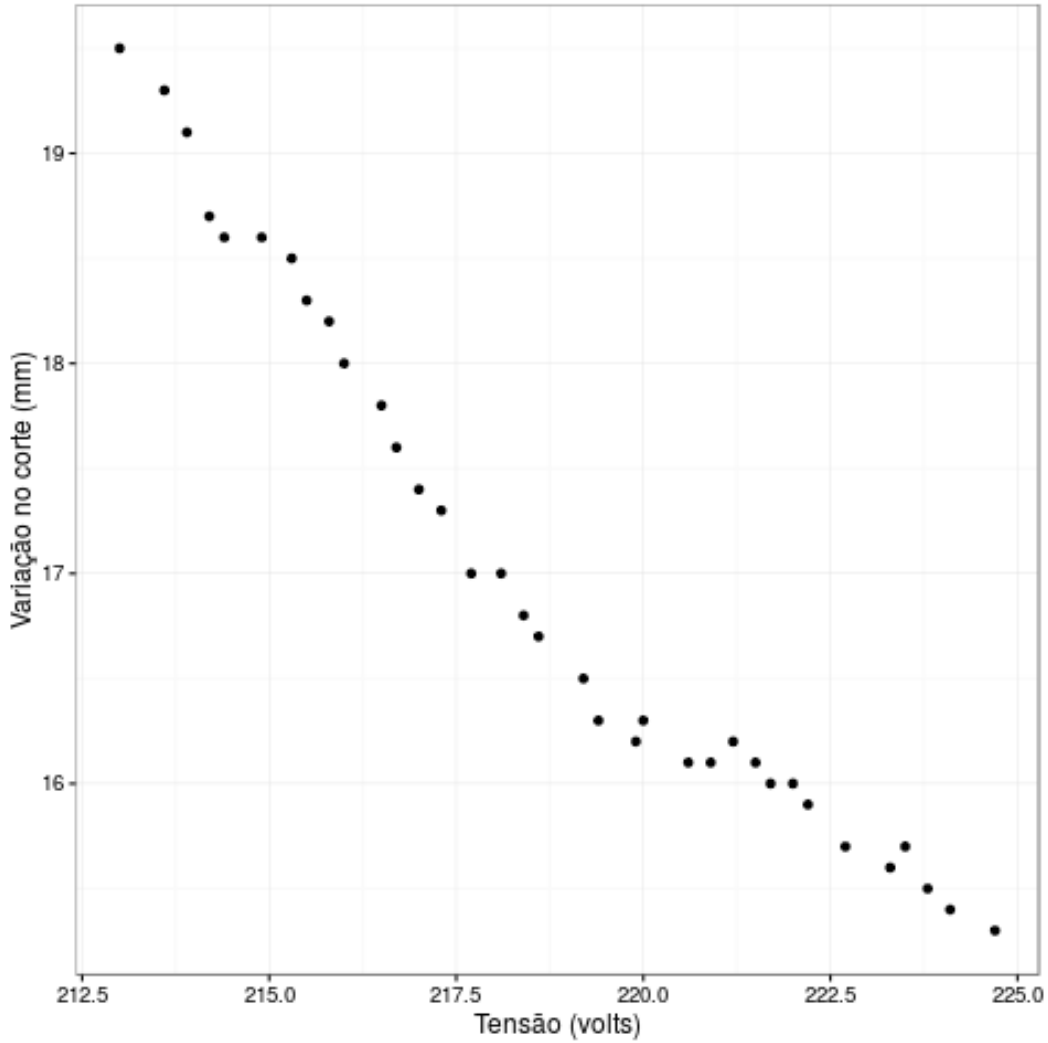
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Modelo de regressão linear múltipla

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

- Y: variável resposta ou dependente
- X: variáveis explicativas ou preditoras ou independentes.

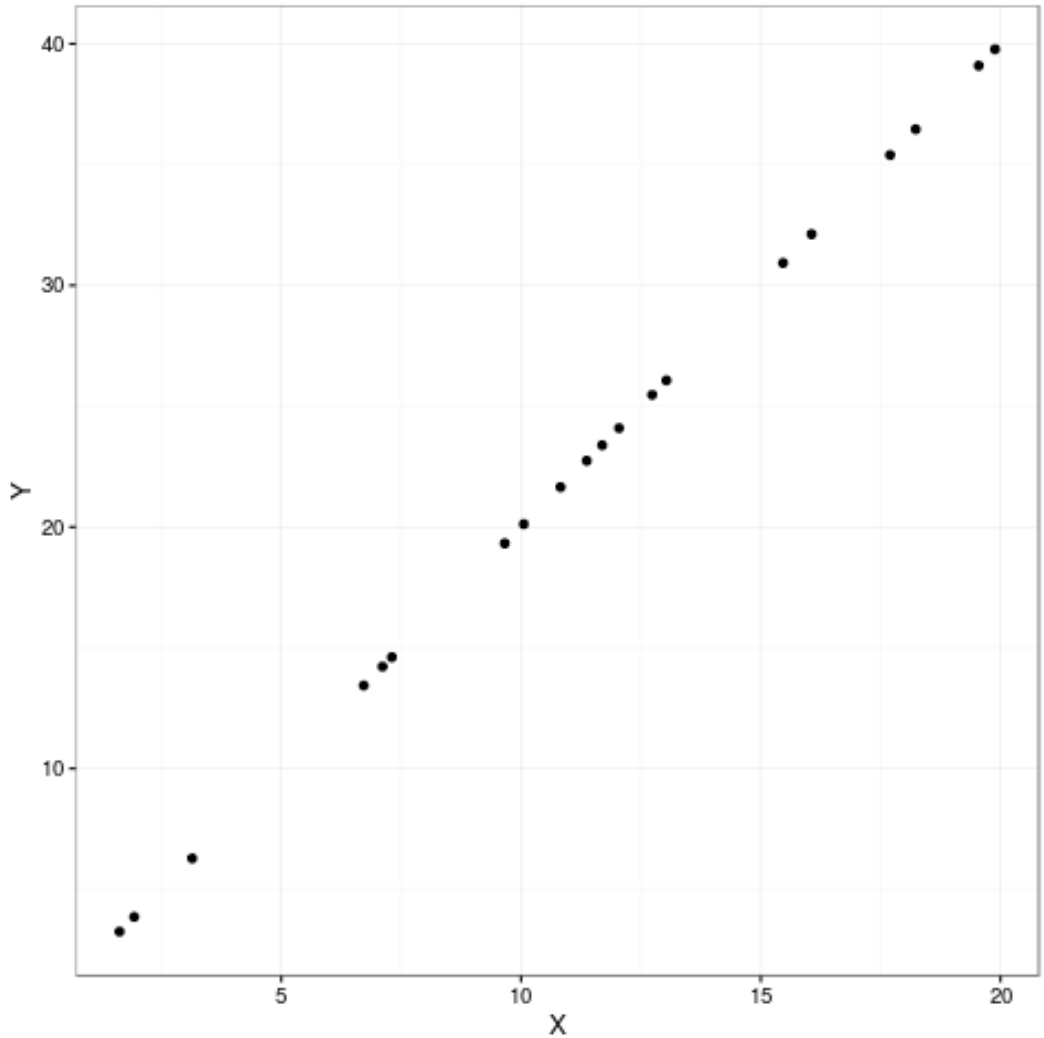
Diagrama de dispersão da Tensão da Rede Elétrica e da Variação no Corte



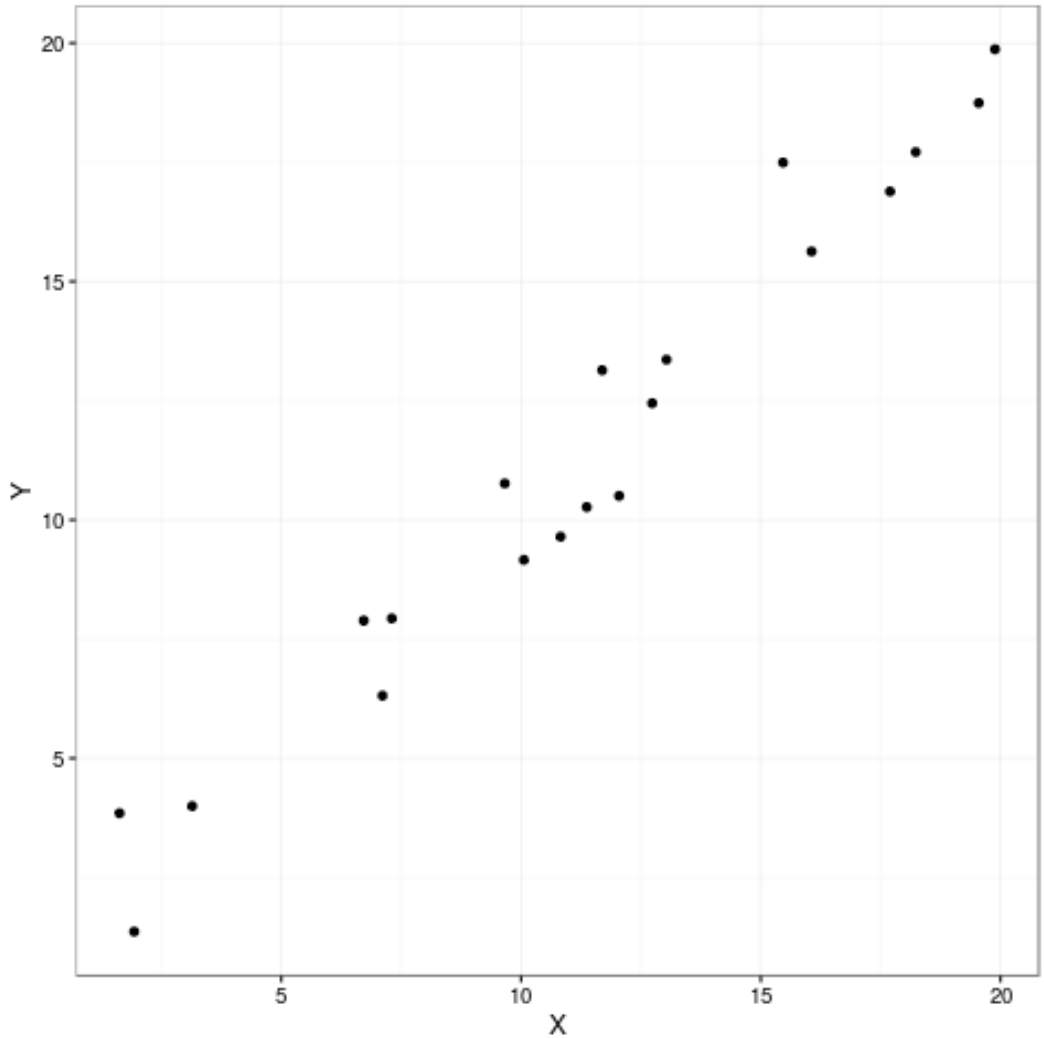
Coeficiente de correlação

- Quantifica a associação entre X e Y.
- Índice que varia entre -1 e 1.
- Valores próximos de -1 indicam uma relação linear negativa.
- Valores próximos de 1 indicam uma relação linear positiva.
- Valores próximos indicam ausência de relação entre as variáveis.
- Pode ser representado pela letra r.

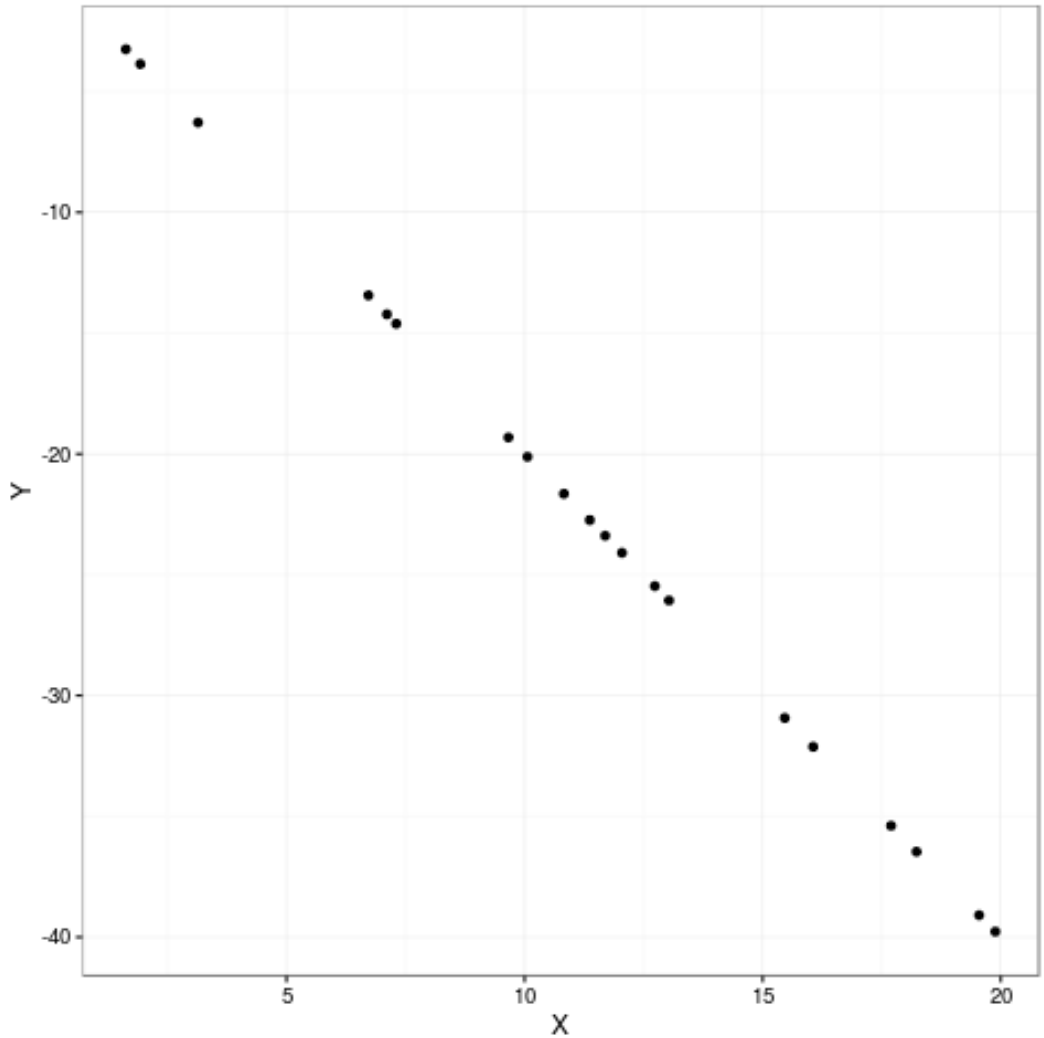
Exemplos
r=1



Exemplos
r>0 e r≈1

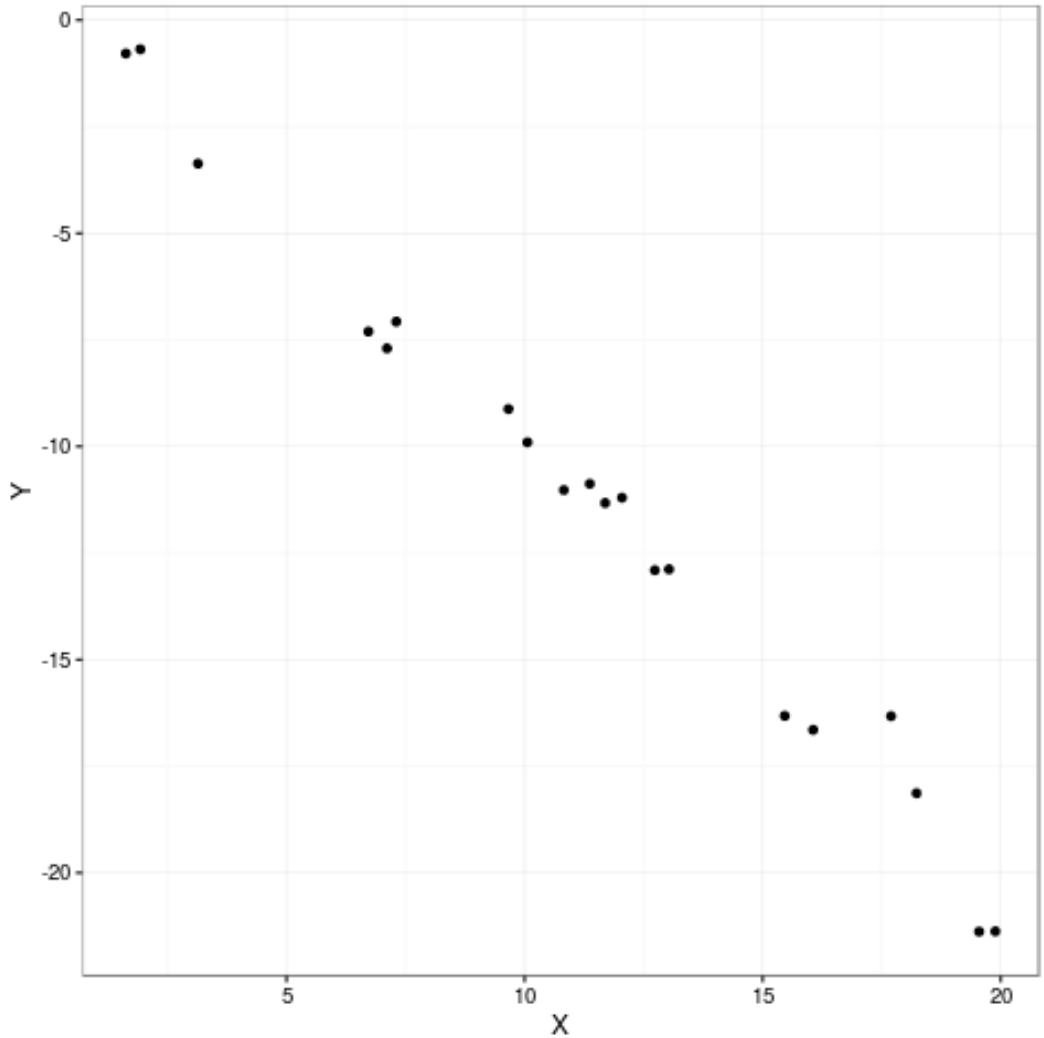


Exemplos
r=-1



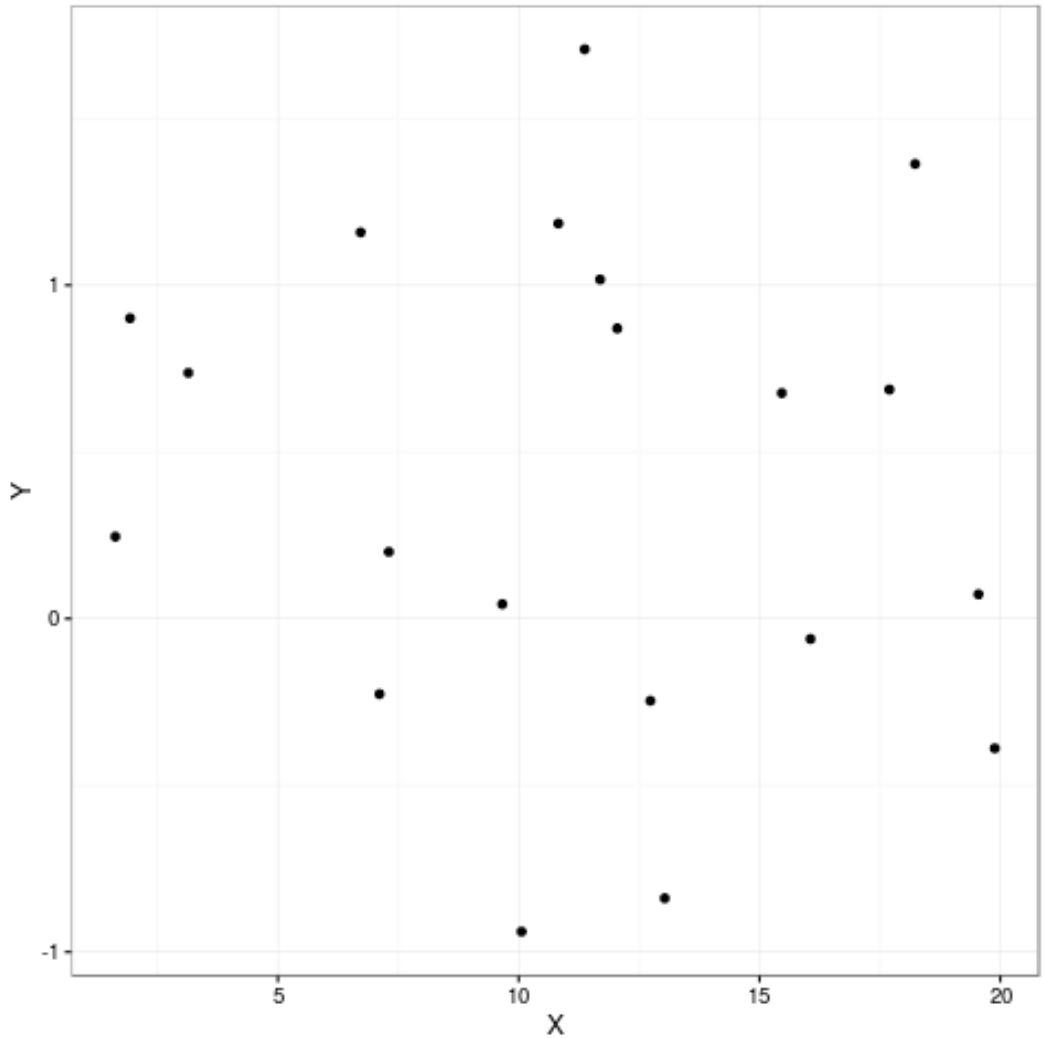
Exemplos

r<0 e r≈-1



Exemplos

r≈0



Cálculo do coeficiente de correlação de Pearson

Coeficiente de correlação de Pearson:

$$r = \frac{\text{cov}(X, Y)}{\text{DP}(X)\text{DP}(Y)}$$

Que pode ser escrito como

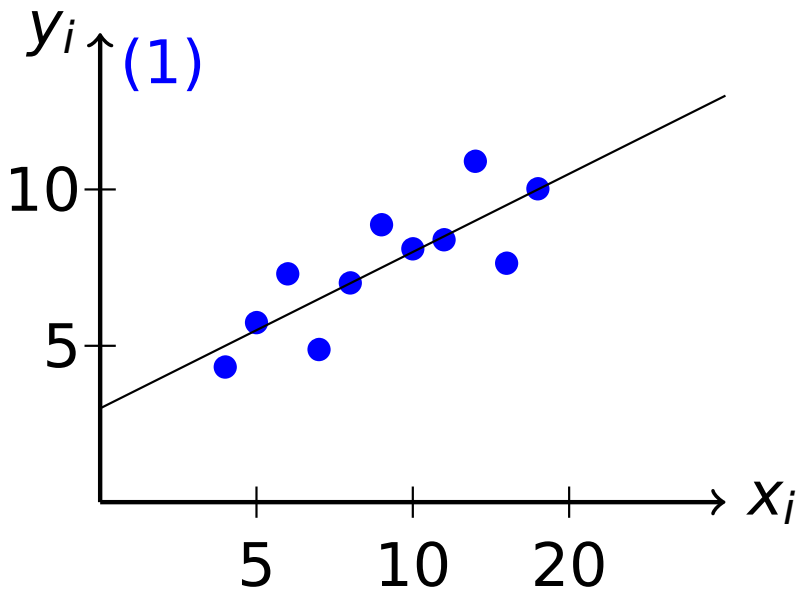
$$r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

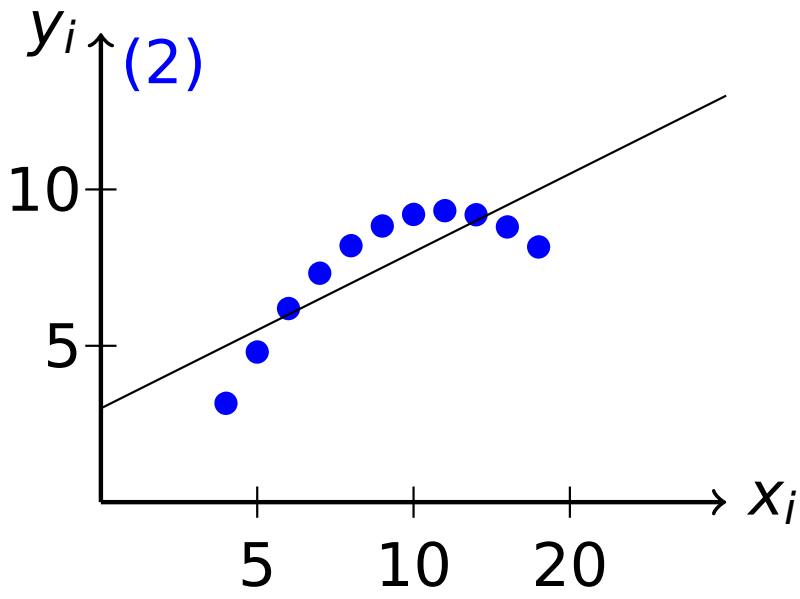
Anscombe (1973) - Graphs in statistical analysis

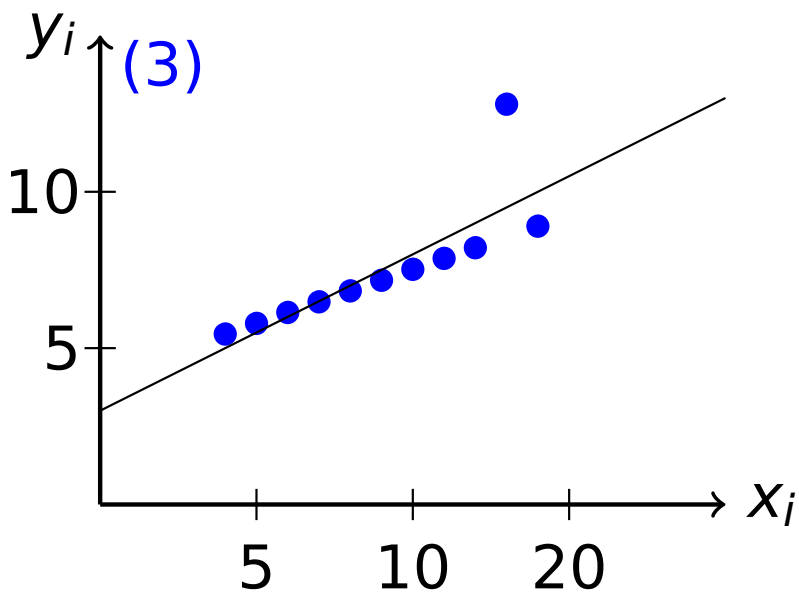
The American Statistician, vol 27, nº 1, pág. 17-21

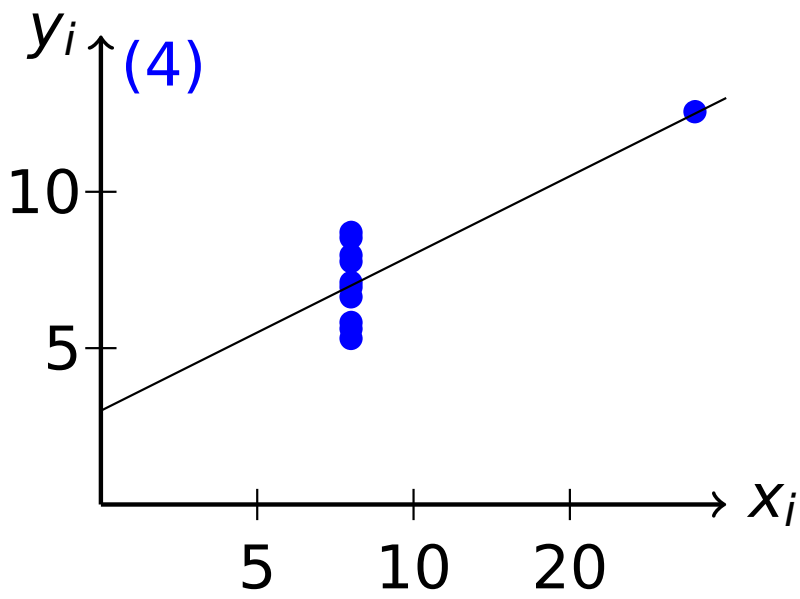
Banco de dados	1-3	1	2	3	4	4
Variável	x	y	y	y	x	y
Obs. nº 1:	10	8.04	9.14	7.46	8	6.58
2:	8	6.95	8.14	6.77	8	5.76
3:	13	7.58	8.74	12.74	8	7.71
4:	9	8.81	8.77	7.11	8	8.84
5:	11	8.33	9.26	7.81	8	8.47
6:	14	9.96	8.10	8.84	8	7.04
7:	6	7.24	6.13	6;08	8	5.25
8:	4	4.26	3.10	5.39	19	12.50
9:	12	10.84	9.13	8.15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

Gráficos dos dados



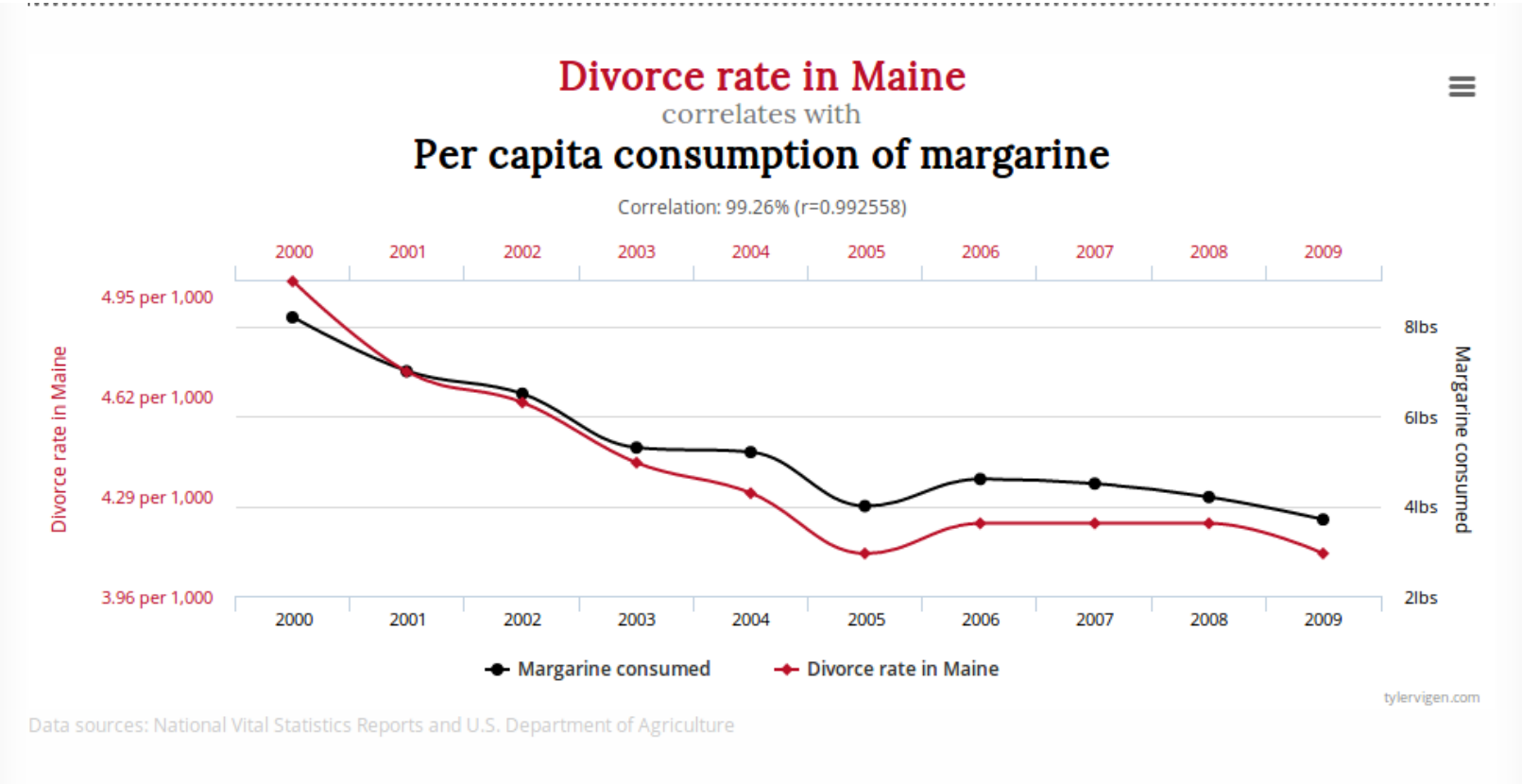






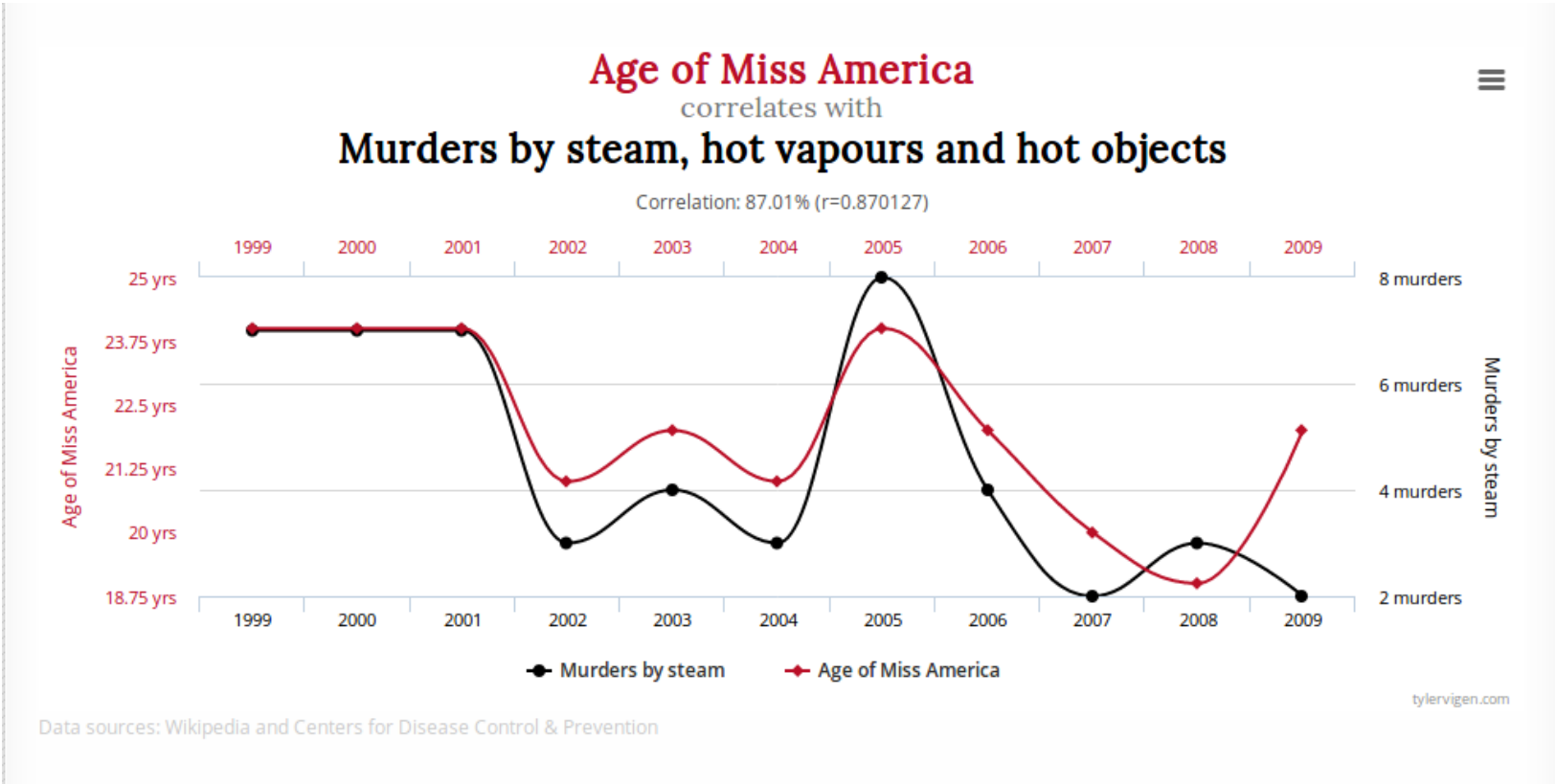
Correlação espúria

http://www.tylervigen.com/



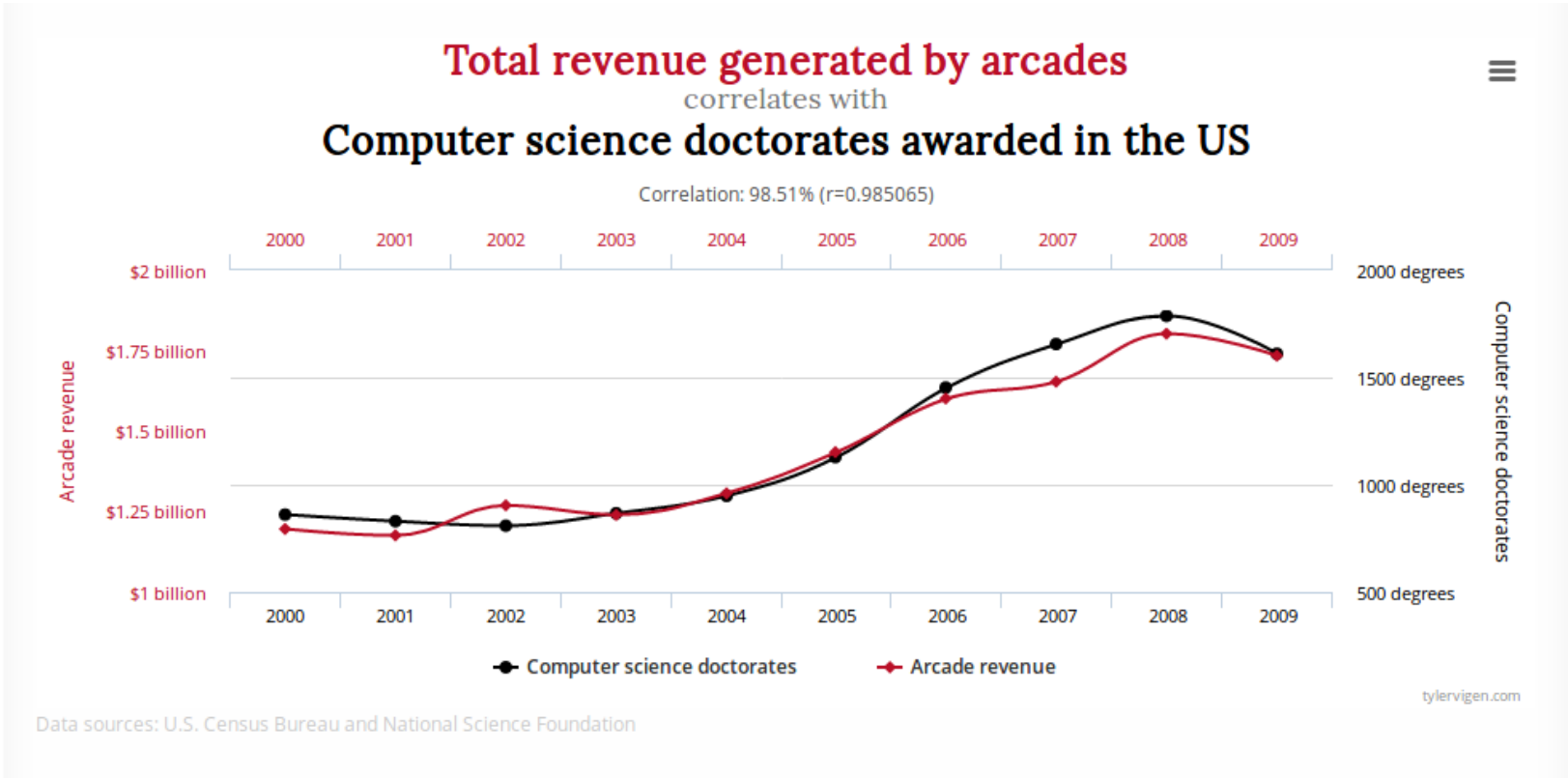
Correlação espúria

http://www.tylervigen.com/



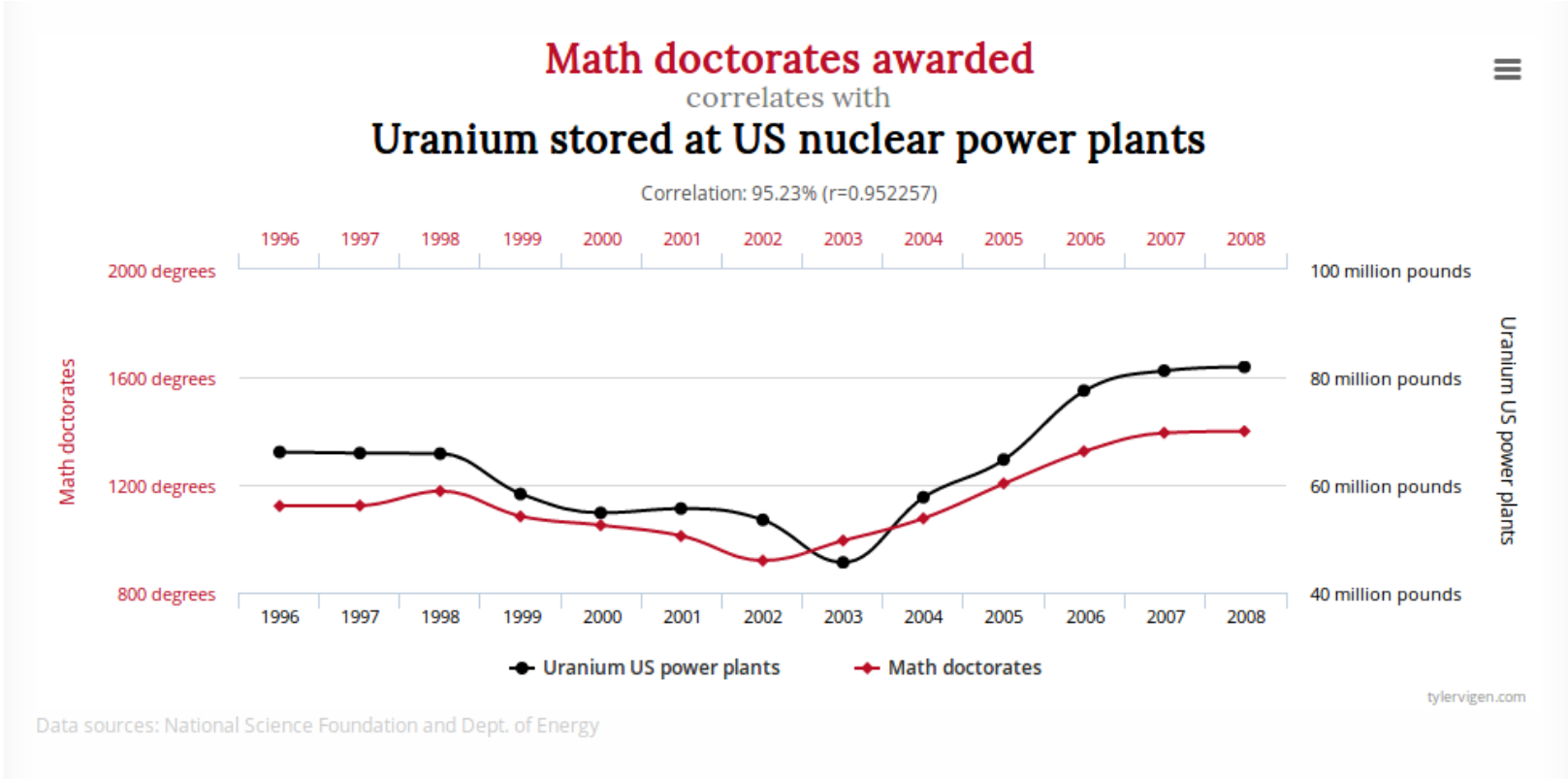
Correlação espúria

http://www.tylervigen.com/



Correlação espúria

http://www.tylervigen.com/



Análise de regressão

Modelo de regressão linear simples

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Estimador de mínimos quadrados

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

Reta estimada

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X_i$$

Expressão para os estimadores

Derivando e igualando a zero a seguinte soma

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2,$$

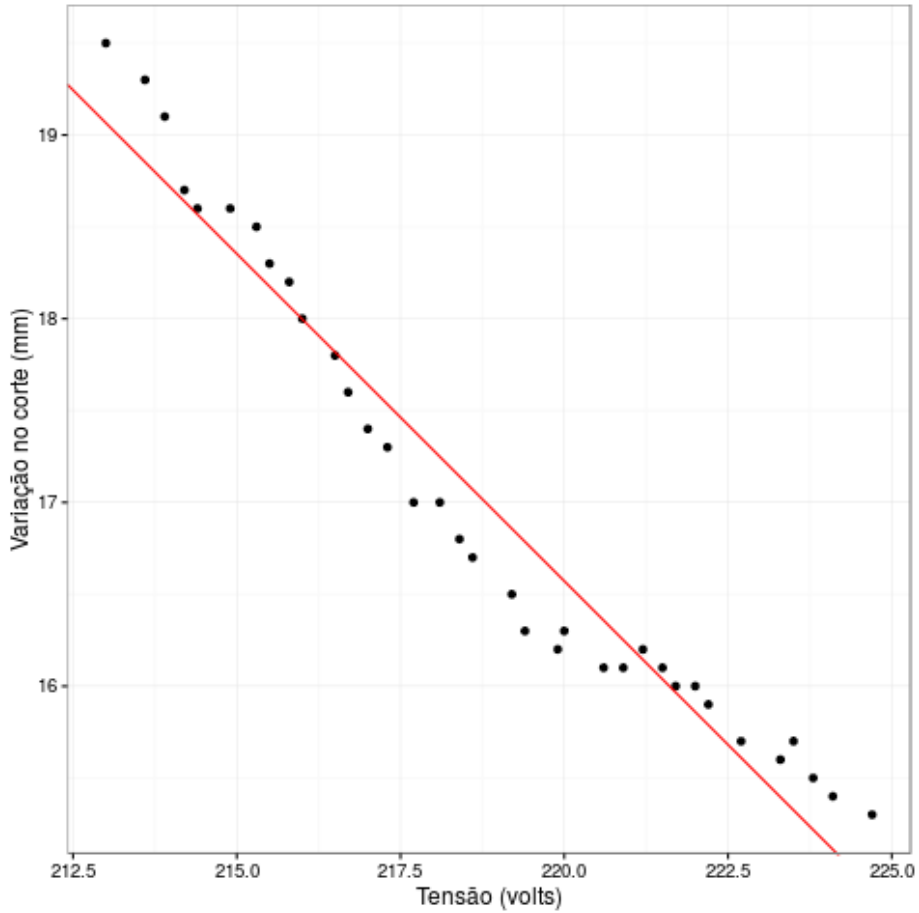
é possível obter os seguintes estimadores

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

e

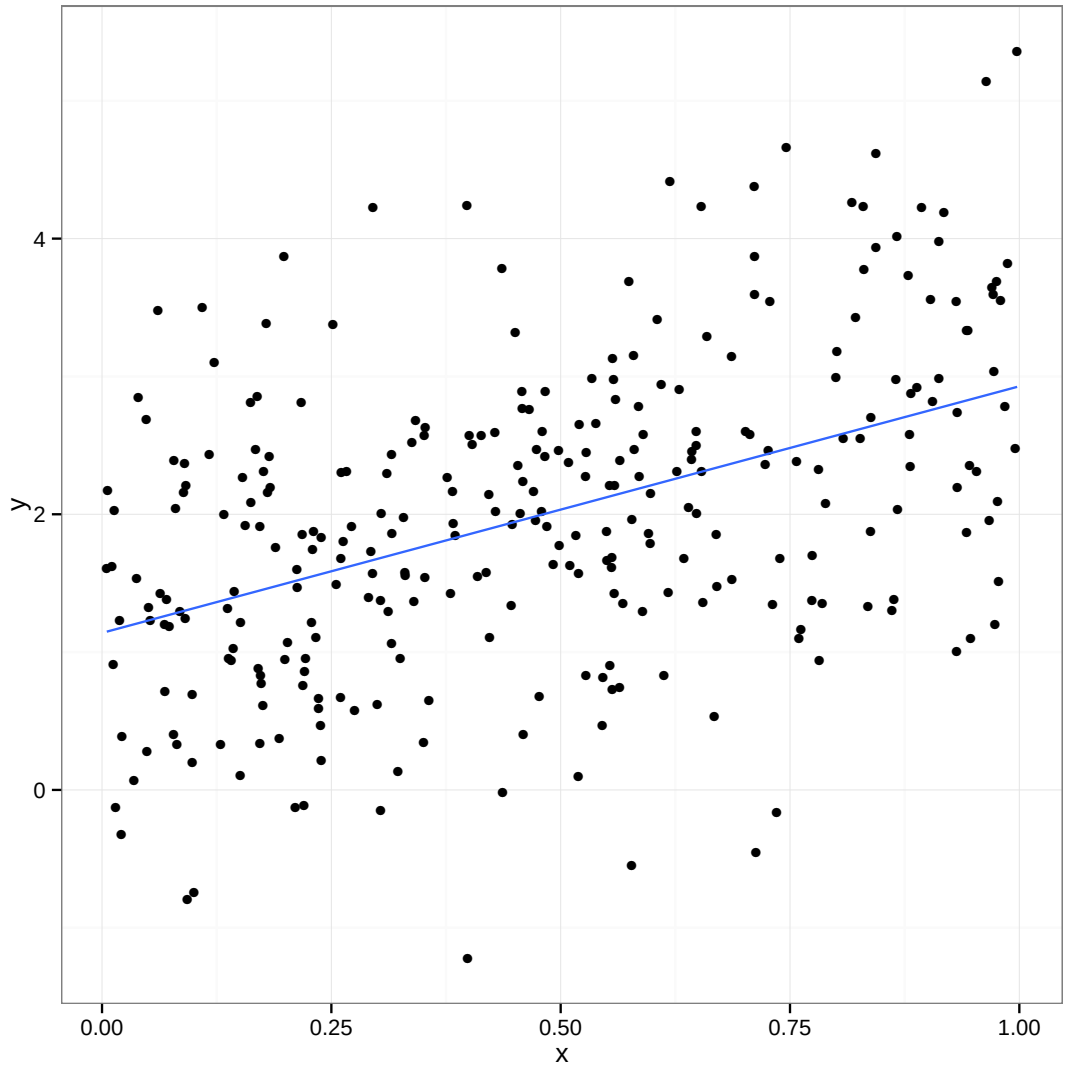
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Exemplo apostila

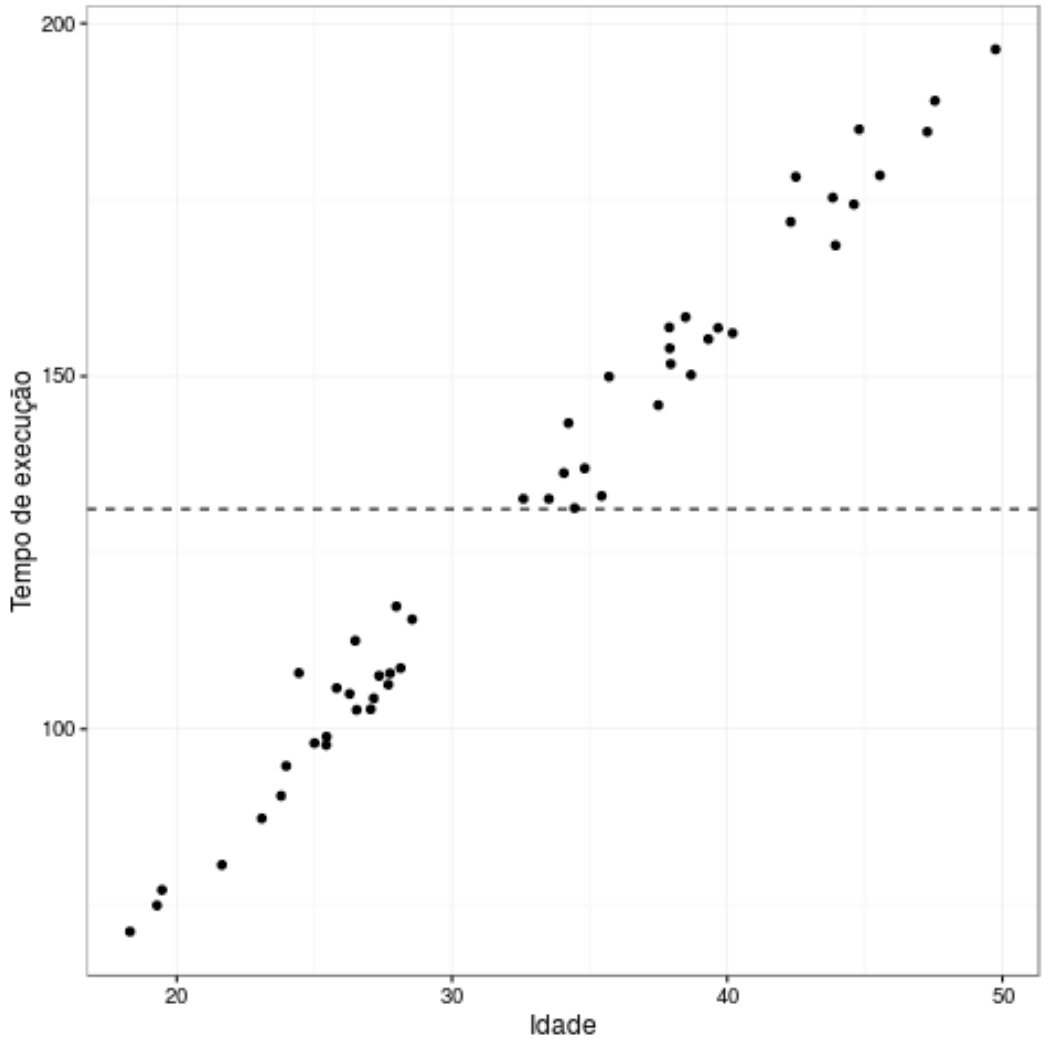


$$\hat{\alpha} = 94,9576 \quad \text{e} \quad \hat{\beta} = -0.3563$$

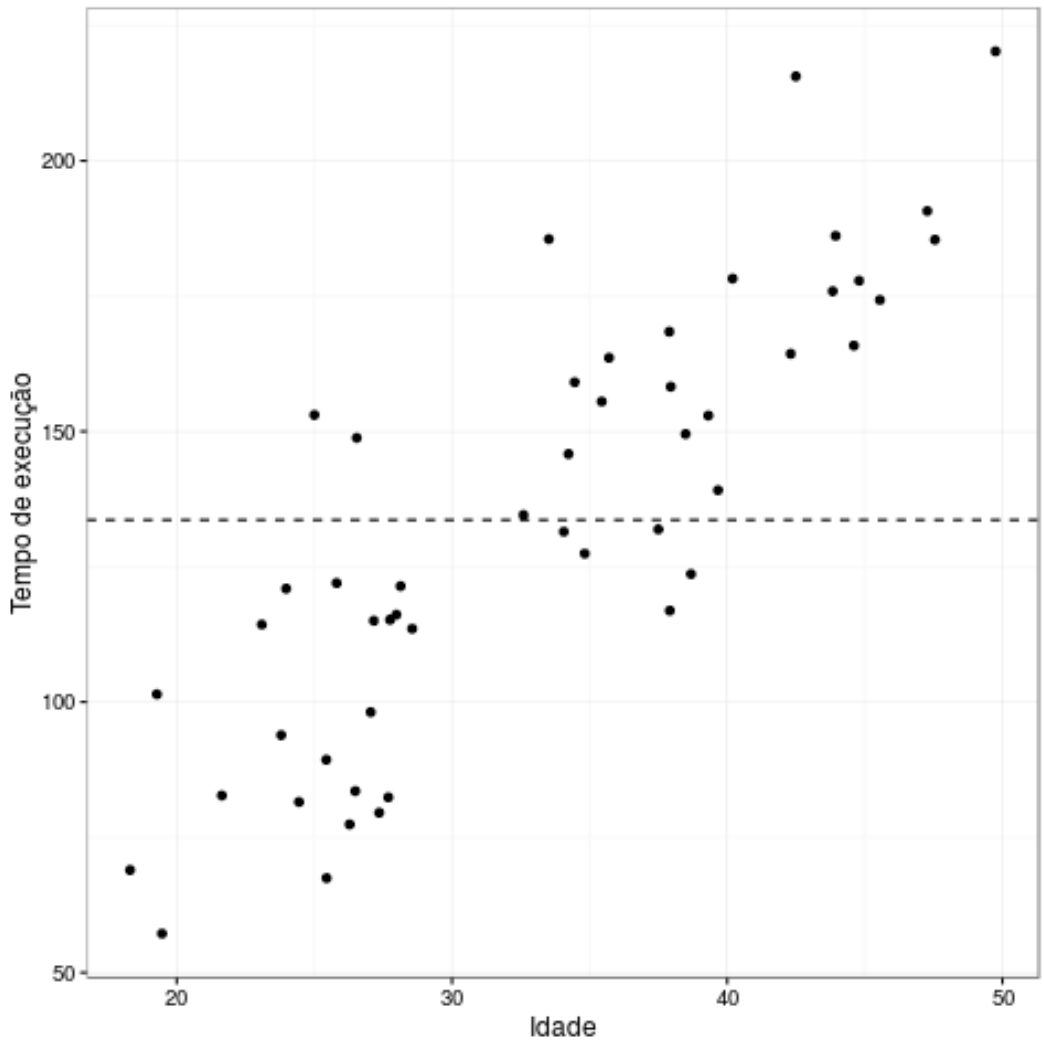
Análise de regressão



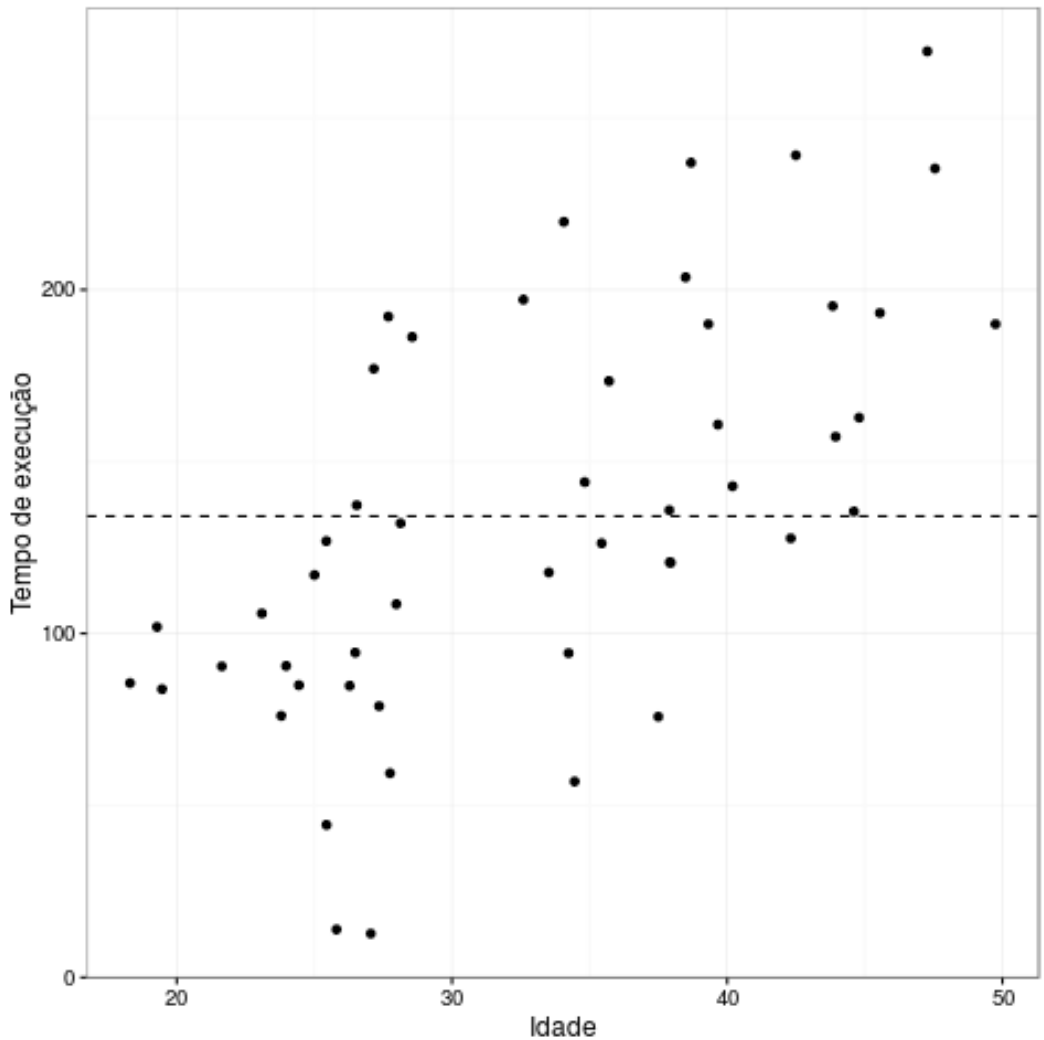
Caso 1



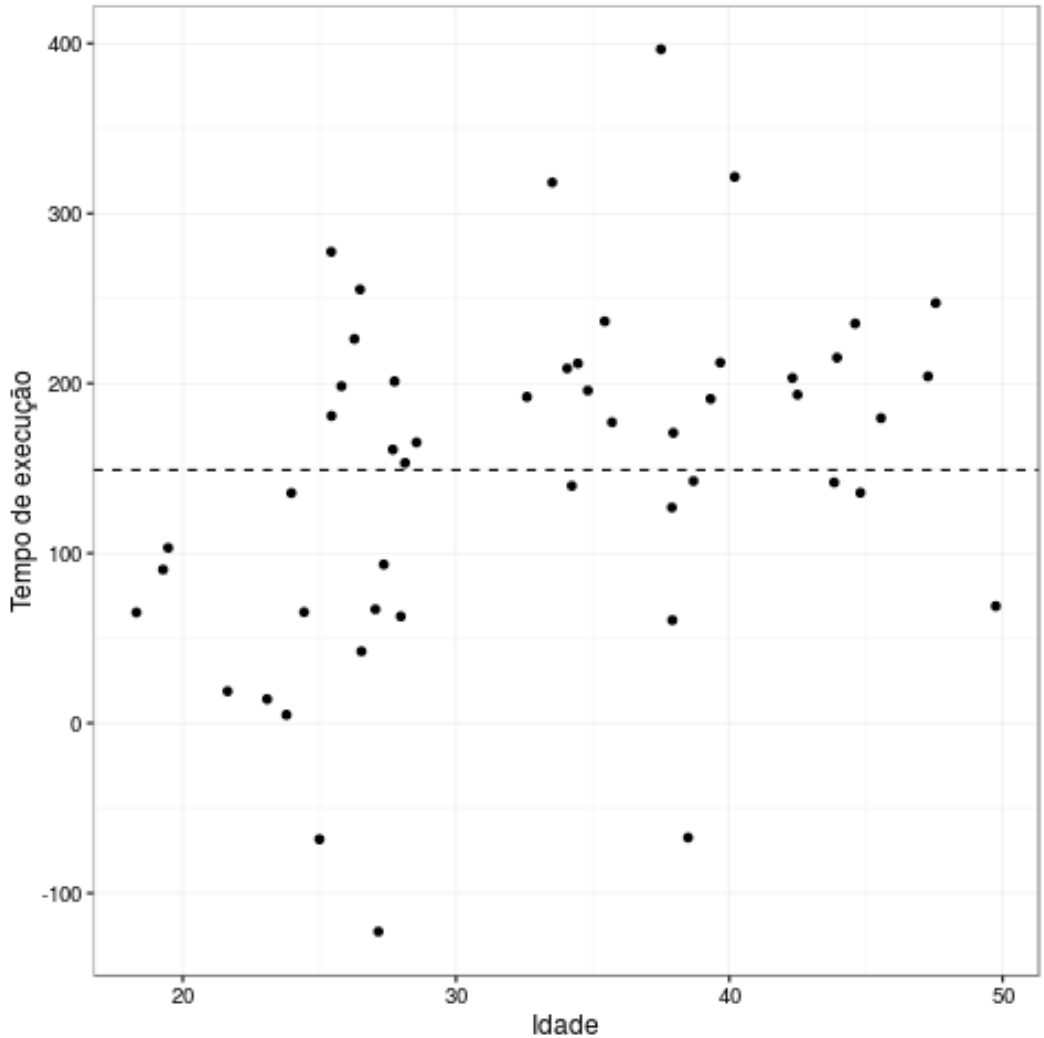
Caso 2



Caso 3



Caso 4



Soma de quadrados

Medida de variabilidade total dos dados:

$$\text{SQ}_{\text{Total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Variabilidade explicada pelo modelo de regressao:

$$\text{SQ}_{\text{Regressão}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Variabilidade não explicada pelo modelo de regressão:

$$\text{SQ}_{\text{Residual}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

É possível mostrar que

$$\text{SQ}_{\text{Total}} = \text{SQ}_{\text{Regressão}} + \text{SQ}_{\text{Residual}}$$

Análise de variância

A tabela de análise de variância (ANOVA) é utilizada para testar a seguinte hipótese:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Não rejeitar $H_0 \Rightarrow$ Não existe relação linear entre X e Y.

ANOVA:

Fonte de variação	GL	SQ	QM	F
Regressão	1	SQ_{Reg}	$QM_{Reg} = \frac{SQ_{Reg}}{1}$	$F = \frac{QM_{Reg}}{QM_{Res}}$
Residual	n-2	SQ_{Res}	$QM_{Res} = \frac{SQ_{Res}}{n-2}$	
Total	n-1	SQ_{Tot}		

Estatística de teste

Utiliza a estatística de teste

$$F = \frac{\text{QM}_{\text{Reg}}}{\text{QM}_{\text{Res}}}.$$

Supondo que $\varepsilon \sim N(0, \sigma^2)$, é possível mostrar que

$$F \sim \text{Fisher-Snedecor}_{(1,n-2)}$$

O critério do teste é o seguinte:

- Rejeita-se H_0 se $F > F_{1,n-2}(\alpha)$
- $F_{1,n-2}(\alpha)$ é o quantil $1-\alpha$ da dist. $\text{Fisher-Snedecor}_{(1,n-2)}$
- Caso contrário, não rejeita-se a hipótese.

Cálculo manual

Os seguintes valores podem ser utilizados

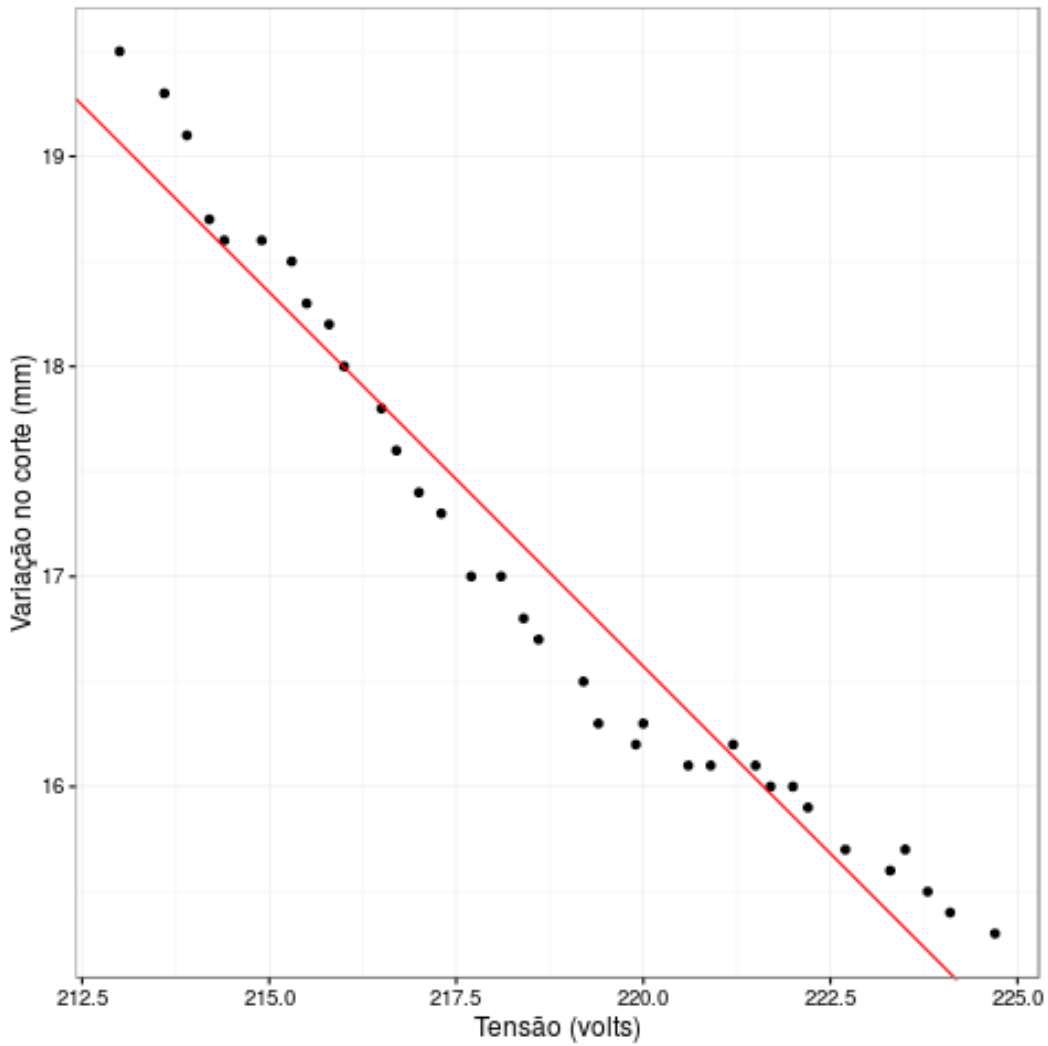
$$\text{SQ}_{\text{Total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2$$

$$\text{SQ}_{\text{Reg}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}\left[\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right]$$

E por último,

$$\text{SQ}_{\text{Res}} = \text{SQ}_{\text{Total}} - \text{SQ}_{\text{Reg}}$$

Dados do Exemplo 14.1



Valores para o exemplo

Testes de hipóteses:

- $H_0 : \beta = 0$ (Não existe relação linear entre a tensão da rede elétrica e o corte da gaveta)
- $H_1 : \beta \neq 0$ (Existe relação linear entre a tensão da rede elétrica e o corte da gaveta)

$$\begin{aligned} \text{SQ}_{\text{Total}} &= \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 \\ &= 10.178,11 - (1/35)(595,3)^2 \approx 52,907 \end{aligned}$$

$$\begin{aligned} \text{SQ}_{\text{Reg}} &= \hat{\beta}\left[\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right] \\ &= -0,3563\left[130103,39 - 35\left(\frac{7657,60}{35}\right)\left(\frac{595,30}{35}\right)\right] \\ &\approx 50,397 \end{aligned}$$

Estatística de teste

Logo,

$$\begin{aligned} \text{SQ}_{\text{Res}} &= \text{SQ}_{\text{Total}} - \text{SQ}_{\text{Reg}} \\ &= 52,907 - 50,397 = 2,513 \end{aligned}$$

Temos então

$$\text{MQ}_{\text{Reg}} = \frac{\text{SQ}_{\text{Reg}}}{1} = 50,397,\ \text{MQ}_{\text{Res}} = \frac{\text{SQ}_{\text{Reg}}}{n-2} = \frac{2,513}{33} = 0,07($$

A estatística de teste F fica dada por

$$F = \frac{50,397}{0,0762} = 661,377$$

Tabela ANOVA - Exemplo

Fonte de variação	GL	SQ	QM	$F_{\text{calculado}}$
Regressão	1	50,397	50,397	$F = 661,377$
Residual	33	2,513	0,0762	
Total	34	52,907		

Utilizando a tabela da distribuição Fisher-Snedecor, temos que a região de rejeição é definida por

$$[4,139, \infty)$$

$F_{\text{calculado}} \in$RC, portanto rejeitamos H_0

- $\Rightarrow$ Os dados indicam que existe relação linear entre a tensão da rede elétrica e a variação no corte da gaveta

Coeficiente de determinação

Definição:

- Proporção da variabilidade total explicada pelo modelo de regressão
- Varia entre 0 e 1

É calculada como

$$R^2 = \frac{\text{SQ}_{\text{Reg}}}{\text{SQ}_{\text{Total}}} = 1 - \frac{\text{SQ}_{\text{Res}}}{\text{SQ}_{\text{Tot}}}$$

No exemplo,

$$R^2 = \frac{50,397}{52,907} = 0,953$$

Intervalo de confiança para previsão

Suponha que exista o interesse em:

- fazer a previsão de Y para um determinado valor de X^*

Uma estimativa pontual pode ser obtida como

$$Y^* = \hat{\alpha} + \hat{\beta} X^*$$

Considerando que $\varepsilon \sim N(0, \sigma^2)$, então um intervalo de confiança para essa predição é dado por

$$\hat{\alpha} + \hat{\beta} X^* \pm t_{1-\alpha/2;n-2} S \sqrt{1 + \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}},$$

em que

$$S^2 = \frac{\text{SQ}_{\text{Res}}}{n-2}$$

Exemplo - Tensão elétrica

Suponha que se queira predizer a variação no corte (mm) quando a tensão é 200 volts.

$$X^* = 200$$

Estimativa pontual

$$Y^* = 95,03 - (0,36x200) = 23,03$$

O intervalo de confiança para a previsão é

$$\left[23,03 \pm 2,035 \times 0,276 \sqrt{1 + \frac{1}{35} + \frac{(200 - 218,79)^2}{397,015}} \right] = [22,3$$

Análise de Resíduos

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

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

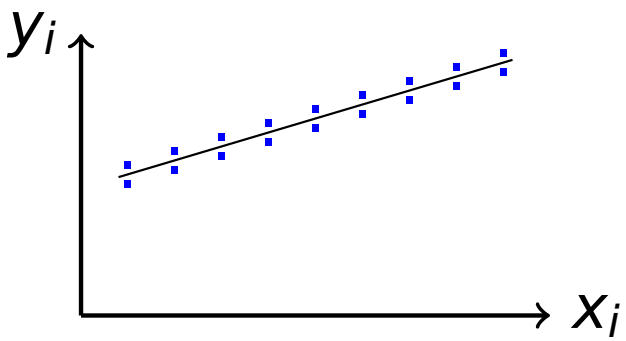
e_i: quantidade que a equação de regressão não consegue explicar

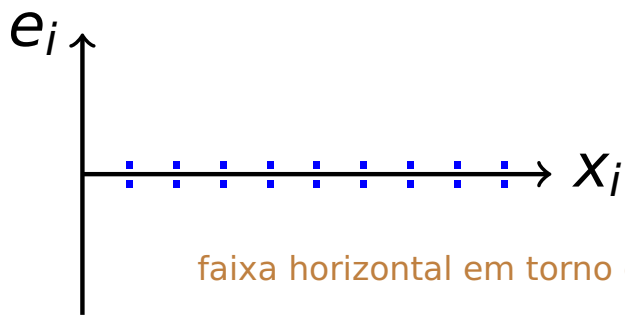
- efeito de variáveis externas (variáveis explicativas omitidas).
- variabiliade natural entre indivíduos.
- eventuais erros de medida na variável Y.

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

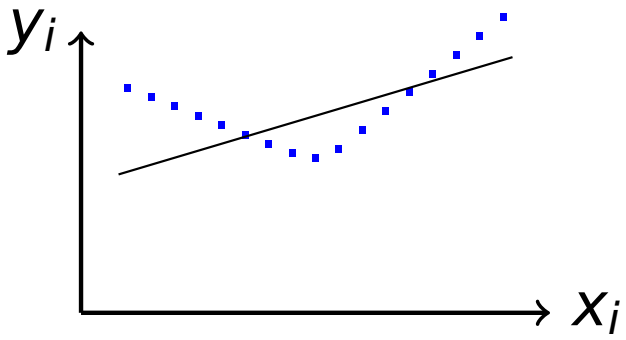
Suposições corretas: $\Rightarrow e_i$ devem apresentar evidências de modo a confirmar ou pelo menos não rejeitar as suposições.

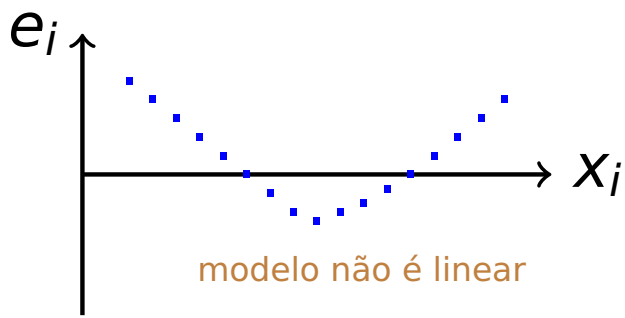
Exemplos



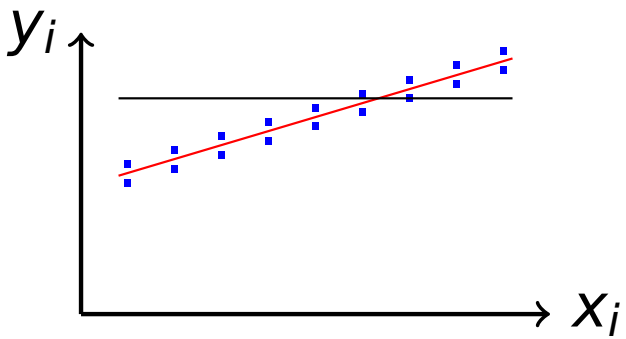


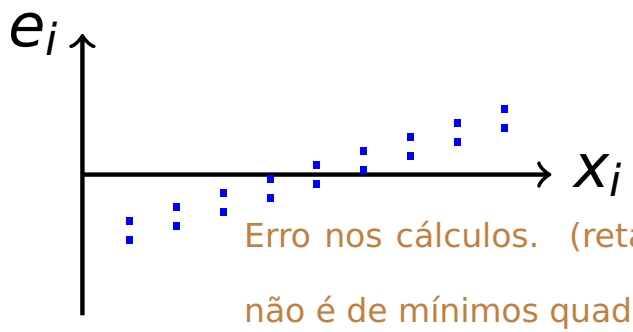
faixa horizontal em torno do zero





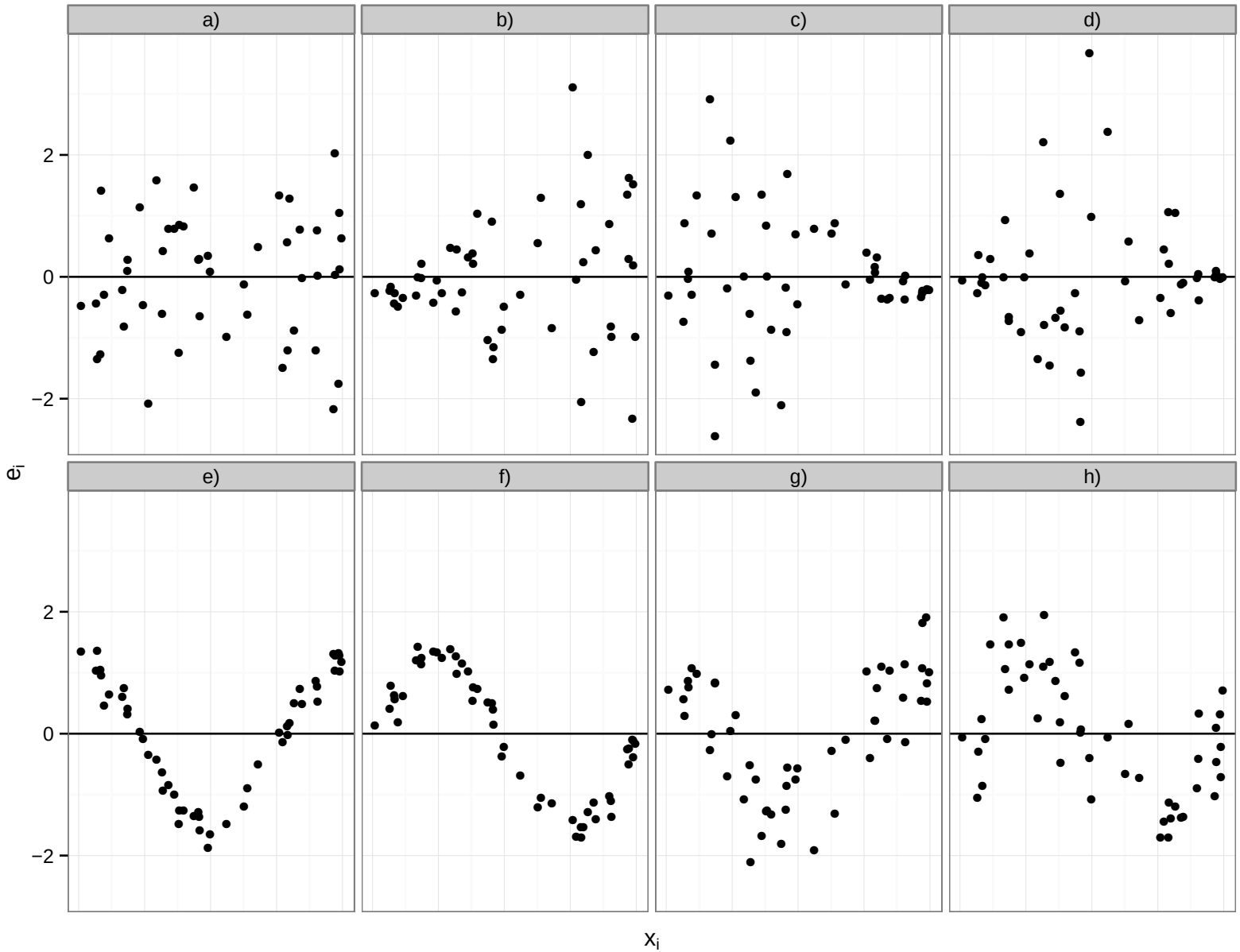
modelo não é linear





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

Gráfico de resíduos vs variável preditora (ou valores ajustados)



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)

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

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

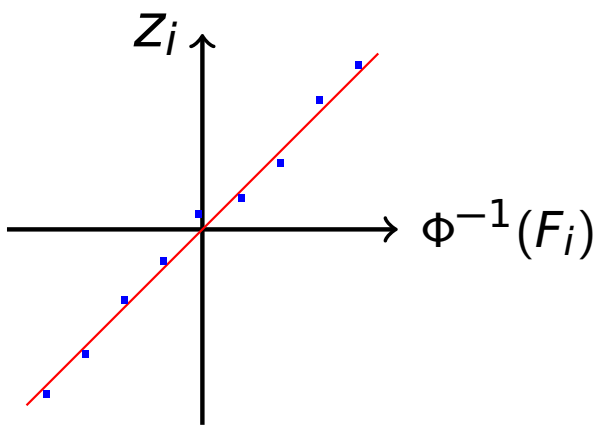
Gráfico de probabilidade normal

$$z_i = \frac{e_i}{\sqrt{\text{QM}_{\text{Res}}}},$$

$$F_i = \frac{\#(Z \leqslant z_i)}{n},$$

$$\Phi(x) = P(Z \leqslant x),\ Z \sim N(0,1).$$

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	



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.

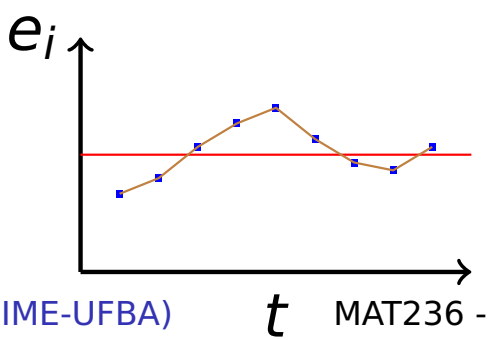
b) $e_i \times y_i$

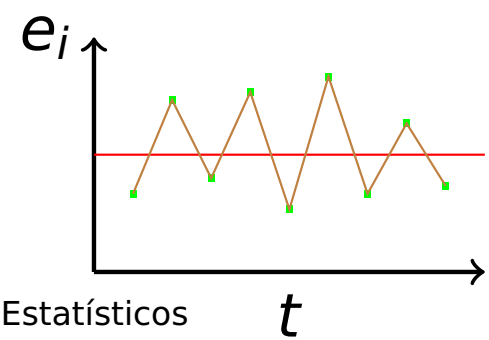
Desaconselhável. Verifica-se que e_i e y_i são correlacionadas.

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.

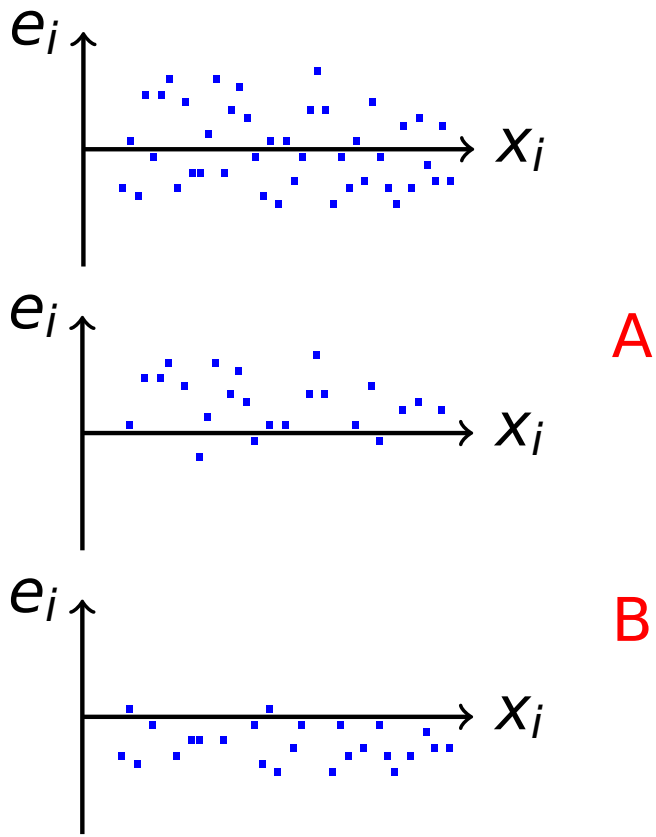




d) $e_i \times$ valores de uma variável independente omitida.
Qualquer "padrão" exibido por este gráfico indica que o modelo pode ser melhorado com a inclusão desta variável independente.

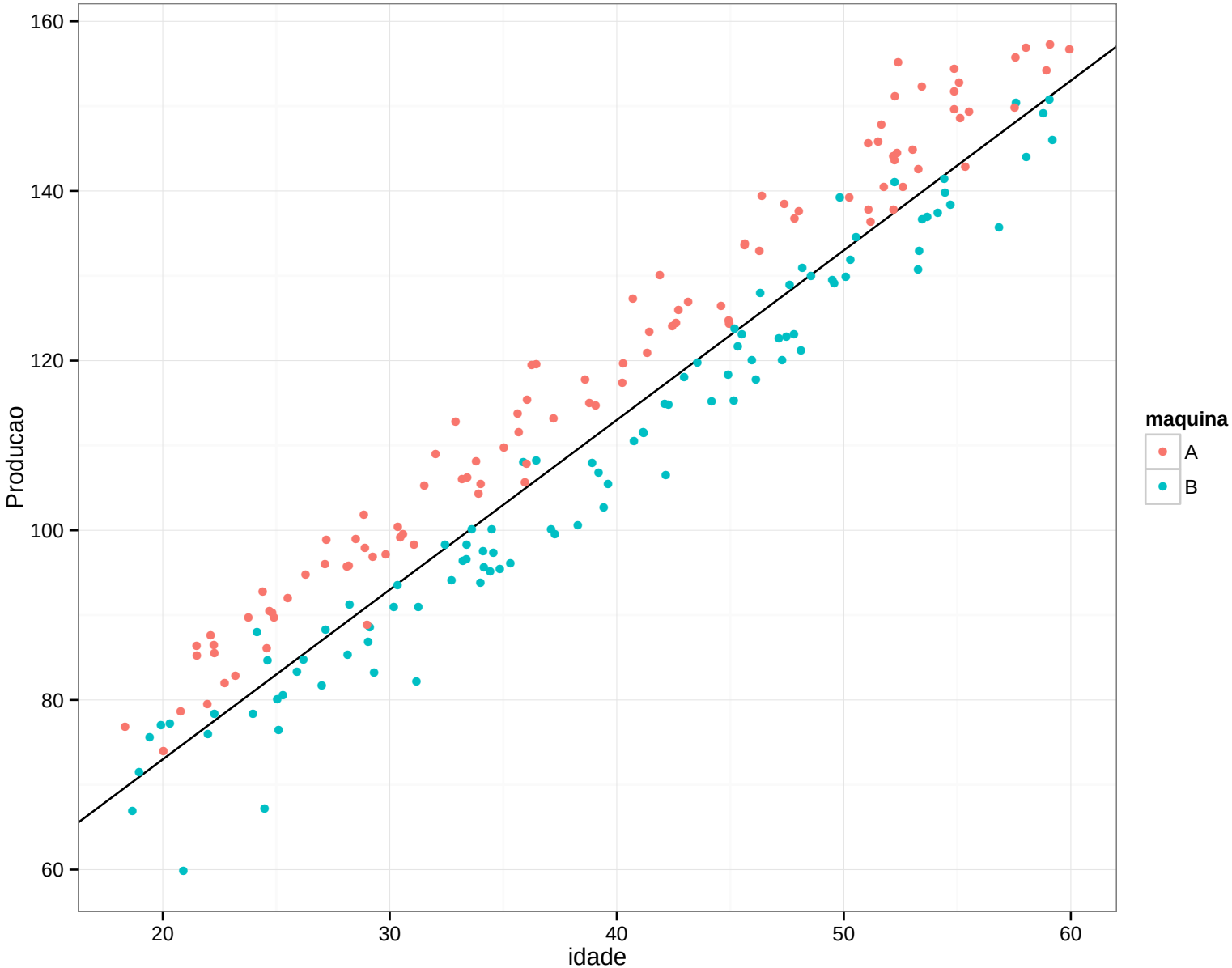
Ex: Y - Produção, X - idade do operário, W - máquina A e B

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

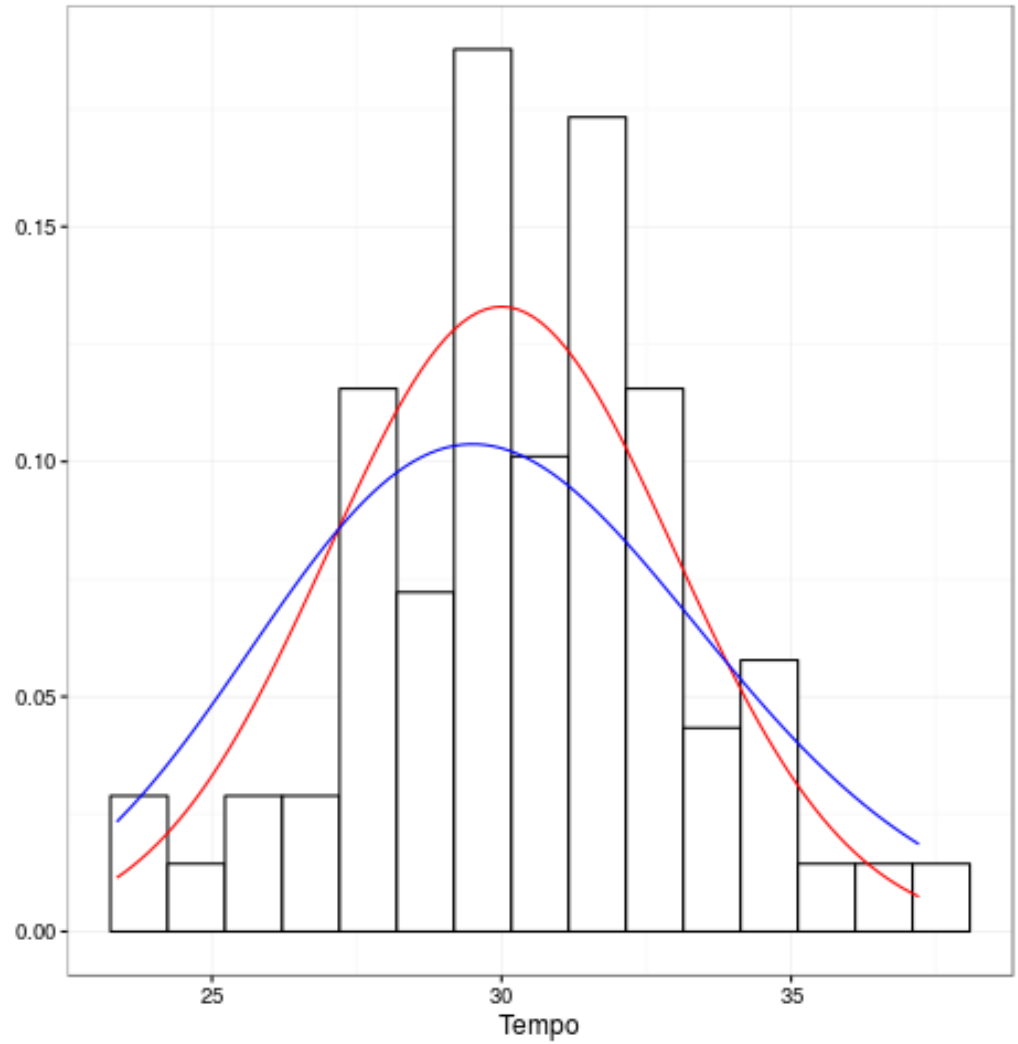


Introdução da variável independente máquina pode

Exemplo variável omitida

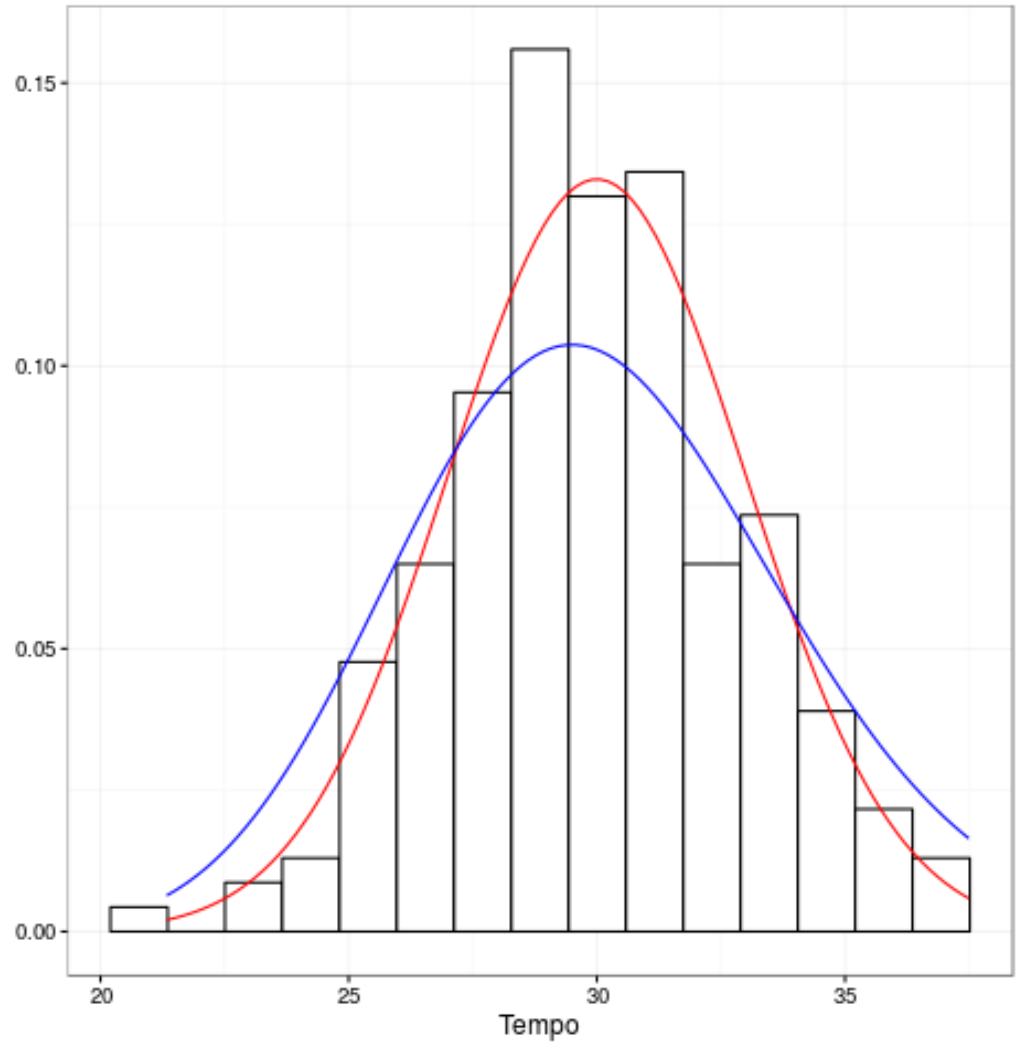


Ideia de aderência



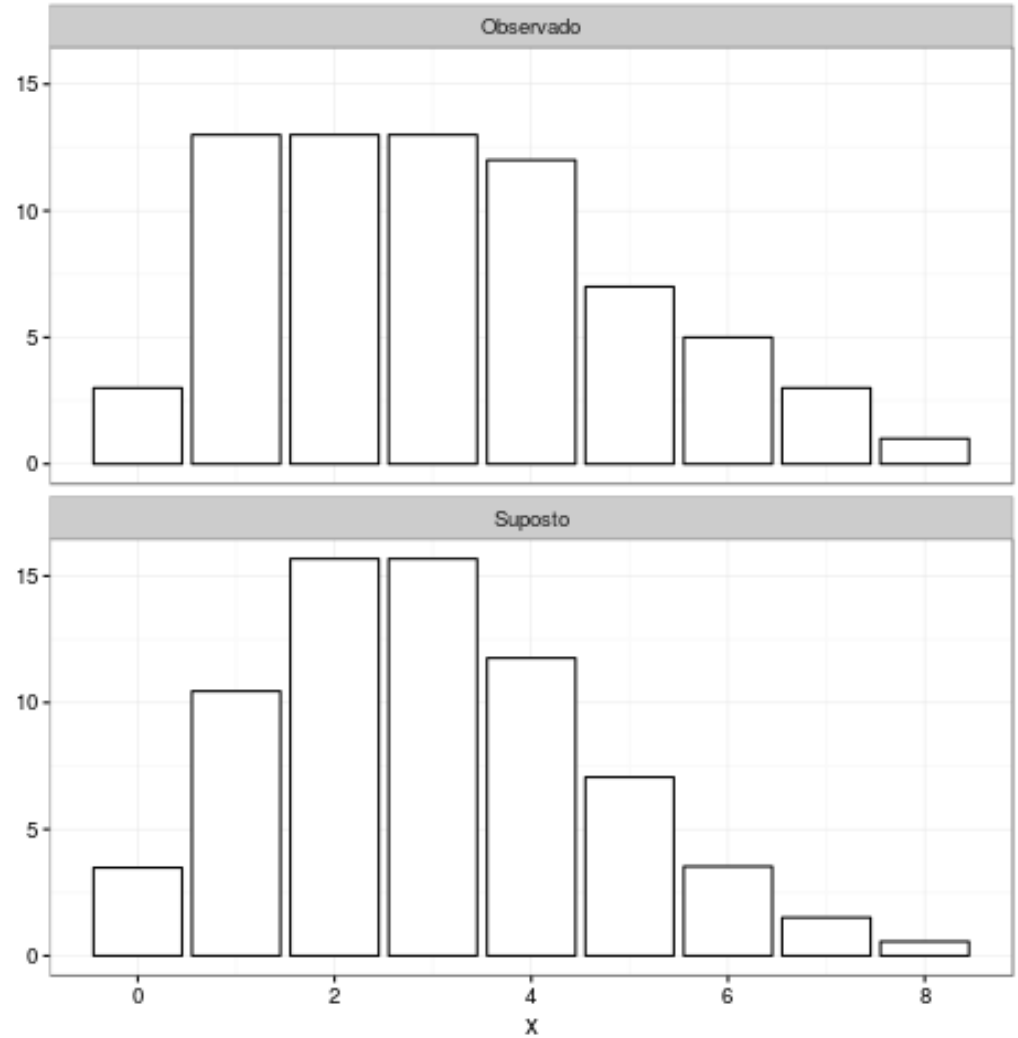
Alguma das curvas teóricas está adequada aos dados?

Ideia de aderência



E quando aumentamos o tamanho da amostra?

Caso discreto



A distribuição observada se aproxima da distribuição teórica?

Teste de aderência

- Na análise estatística são sempre feitas suposições.
- Após observar os dados podemos checar essas suposições.
- Até o momento, fizemos testes de hipóteses para os parâmetros de uma distribuição de probabilidade.
- O teste de aderência permite testar se o modelo estatístico proposto é razoável.
- Para isso são comparadas os valores esperados pelo modelo proposto e os valores observados na amostra.

Teste Qui-quadrado de aderência

- Compara as frequências observadas na amostra com as frequências caso o modelo proposto fosse verdadeiro.
- Considere n observações de uma variável aleatória X com função de distribuição de probabilidade não especificada.
- Cada observação é classificada em uma dentre k categorias.

Variável	Cat. 1	Cat. 2	$\cdots$	Cat. k
Frequência observada	O_1	O_2	$\cdots$	O_k

Estatística de teste

- H_0: A variável X segue o modelo proposto.
- H_1: A variável X não segue o modelo proposto.

A estatística de teste é dada por

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\nu}$$

- k é o numero de categorias;
- O_i frequência observada na i-ésima categoria;
- E_i frequência esperada na i-ésima categoria: np_i;
- p_i probabilidade da i-ésima categoria;
- $\nu = k - 1$ se as frequências esperadas puderem ser calculadas sem estimação dos parâmetros da distribuição.

Regra de decisão

- Para um dado nível de significância α:
  - Rejeita-se H_0 se $\chi^2_{cal} > \chi^2_{1-\alpha}$, em que

$$P(\chi^2_\nu > \chi^2_{1-\alpha}) = \alpha$$

- Também rejeita-se H_0 quando o p-valor calculado é menor que o nível de significância, α.

- Note que o teste rejeita a hipótese nula quando as diferenças entre os valores esperados e os observados são *grandes*.

- A utilização da distribuição Qui-quadrado é um resultado aproximado, por isso alguns cuidados devem ser tomados.

Cuidados com o uso desse teste

- Quando o número de categorias for igual a dois ($k = 2$) as frequências esperadas dentro de cada categoria devem ser iguais ou superiores a 5.
- Quando $k > 2$, não deve ter mais de 20% das categorias com frequências esperadas menores que 5 e nenhuma frequência esperada igual a zero.
- Quando as categorias apresentarem pequenas frequências esperadas elas podem ser combinadas com outras categorias, de tal forma que o sentido do trabalho seja conservado.
- Se houver estimação de algum parâmetro no teste, então $\nu = k - m - 1$, em que m é o número de parâmetros estimados.

Exemplo 15.1 - pg 28

- X: porcentagem de cinzas contidas em carvão
- Afirmação: $P \sim N(\mu, \sigma^2)$.

i	Cinzas (em %)	# de observações
1	09,5|–10,5	2
2	10,5|–11,5	5
3	11,5|–12,5	16
4	12,5|–13,5	42
5	13,5|–14,5	69
6	14,5|–15,5	51
7	15,5|–16,5	32
8	16,5|–17,5	23
9	17,5|–18,5	9
10	18,5|–19,5	1

Exemplo - continuação

- Média e variância (μ, σ^2) são desconhecidos.
- Melhores estimadores: $\bar{X}$ e S^2, respectivamente.

$$\bar{X} = \frac{\sum_{i=1}^{10} x_i f_i}{\sum_{i=1}^{10} f_i} = \frac{10 * 2 + 11 * 5 + \cdots + 19 * 1}{2 + 5 + \cdots + 1} \approx 14,5$$

$$S^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{X})^2 f_i}{(\sum_{i=1}^{10} f_i) - 1} = 2,7$$

- $\hat{X} \sim N(14,5; 2,7)$

Exemplo - continuação

Hipóteses de interesse:

- H_0 : X têm distribuição normal
- H_1 : X não têm distribuição normal

- A distribuição normal é definida no intervalo $(-\infty, \infty)$.
- Temos que calcular as frequências esperdas em cada um dos intervalos propostos.

$$E_1 = 250 * P(X < 10,5) = 250 * P\left(Z < \frac{10,5 - 14,5}{\sqrt{2,7}}\right)$$

$$= 250 * P(Z < -2,43) = 1,875.$$

Exemplo - continuação

$$
\begin{aligned}
E_2 &= 250 * P(10,5 \leq X < 11,5) \\
&= 250 * P\left(\frac{10,5-14,5}{\sqrt{2,7}} \leq Z < \frac{11,5-14,5}{\sqrt{2,7}}\right) \\
&= 250 * P(-2,43 \leq Z < -1,83) = 6,525.
\end{aligned}
$$

Um cálculo similar deve ser feito para as categorias de 3 a 9.

Por último,

$$
\begin{aligned}
E_{10} &= 250 * P(X \geq 18,5) \\
&= 250 * P\left(Z \geq \frac{18,5-14,5}{\sqrt{2,7}}\right) \\
&= 250 * P(Z \geq 2,43) = 1,875.
\end{aligned}
$$

Exemplo - continuação

A tabela com valores observados e esperados ficaria da seguinte forma

Categorias	Freq. observada	Freq. esperada
1	2	1,875
2	5	6,525
3	16	19,4
4	42	39,925
5	69	57,275
6	51	57,275
7	32	39,925
8	23	19,4
9	9	6,525
10	1	1,875

Cálculo da estatística de teste

Com isso, podemos calcular a estatística de teste

$$\begin{aligned}\chi^2_{cal} &= \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(2-1,875)^2}{1,875} + \frac{(5-6,525)^2}{6,525} + \cdots + \frac{(1-1,875)^2}{1,875} \\ &= 7,74\end{aligned}$$

Considerando $\nu = 10-2-1 = 7$ e nível de significância 2,5%, devemos obter o quantil de ordem 97,5% da χ^2_7, que é 16,01.

Conclusão: Não rejeitamos H_0, pois 7,74 é menor que 16,01, então aceitamos que os dados são distribuídos normalmente.

Exercícios de fixação 1

- X: Número de acidentes sofridos por um grupo de mineiros durante um trabalho numa mina de carvão.
- Investigar se a distribuição de X segue o modelo Poisson, com $\lambda = 1,45$.

Número de acidentes	0	1	2	3	4	5
Número de mineiros	35	47	39	20	5	2

Lembrando que, se $X \sim P(\lambda)$, então

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- $H_0 : X \sim P(1,45)$ contra $H_1 : X \not\sim P(1,45)$

Exercícios de fixação 1 - Cont.

Obtendo as frequências esperadas:

$$\begin{aligned} E_0 &= 148 * P(X = 0) \\ &= 148 * \frac{e^{-1,45} 1,45^0}{0!} \\ &= 148 * 0,22 = 34,7164 \end{aligned}$$

Assim por diante até

$$\begin{aligned} E_5 &= 148 * P(X = 5) \\ &= 148 * \frac{e^{-1,45} 1,45^5}{5!} \\ &= 148 * 0,0125 = 1,8544 \end{aligned}$$

Cálculo da estatística de teste

Cat.	0	1	2	3	4	
O_i	35	47	39	20	5	
E_i	34,7164	50,3388	36,4956	17,6396	6,3943	1,8

A estatística de teste fica dada por

$$\chi^2_{cal} = \sum_{i=0}^{5} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(35-34,7164)^2}{34,7164} + \frac{(47-50,3388)^2}{50,3388} + \cdots + \frac{(2-1,8544)}{1,8544}$$

$$= 1,027$$

A região crítica é $[11,0705; \infty)$, logo não rejeitamos H_0.

Definições

Alguns termos que serão utilizados:

- **Fator e nível**
  - Fator é a variável independente num estudo
  - Nível são as formas particulares do fator

- Efeito de diferentes açúcares no crescimento de bactérias
  - Acúcar = fator
    - Glicose, sacarose e frutose são os níveis do fator

- Efeito da concentração de madeira de lei em polpa sobre a resistência à tração das sacolas fabricadas da polpa
  - Concentração de madeira de lei em polpa = fator
    - 5%, 10%, 15% e 20% são os níveis do fator

Definições

- **Tratamento**
  - Objeto que se deseja medir ou avaliar em um experimento
  - Nível do fator sobre análise
  - Combinação de fator e nível em estudos com dois ou mais fatores

- Efeitos de cinco marcas de gasolina na eficiência operacional
  - Fator é a marca
  - Cada marco constitui um tratamento

- Efeitos de horário e fábrica na fabricação de um certo produto
  - Fatores: fábrica e horário
  - Níveis: {A, B} - {Manhã, Noite}
  - Tratamento: {(A - Manhã), (A - Noite), (B - Manhã), (B - Noite)}

Definições

- **Unidade experimental**
  - Unidade que recebe o tratamento
  - Fornece os dados para análise
    - Motor
    - Pessoa
    - Animal
    - Planta
- **Repetição**
  - Número de vezes que aparece um tratamento
  - Depende dos recursos disponíveis
  - Deveria depender da variabilidade do experimento
  - Existem metodologias para estimar um número satisfatório

Análise de variância

- Suponha um experimento com k tratamentos (ou populações)
- A variável resposta de cada unidade experimental em cada tratamento é uma variável aleatória

Tratamento	Observações				Total	Média
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn}	$y_{k.}$	$\bar{y}_{n.}$
					$y_{..}$	$\bar{y}_{..}$

$$y_{i.} = \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{i.} = \frac{y_{i.}}{n}, \quad i = 1, \ldots, k.$$

$$y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{..} = \frac{y_{..}}{N}, \quad N = n \times k.$$

Modelo estatístico

Um modelo para descrever os dados é:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \, j = 1, \ldots, n.$$

- y_{ij}: observação do i-ésimo tratamento na j-ésima unidade
- μ_i: média do i-ésimo tratamento, valor fixo e desconhecido
- ϵ_{ij}: erro aleatório associado ao i-ésimo tratamento na j-ésima unidade experimental
- $\epsilon_{ij} \sim N(0, \sigma^2)$, independentes.
- $y_{ij} \sim N(\mu_i, \sigma^2)$
- μ_i parte sistemática que representa a média da população i, que é fixa
- ϵ_{ij} é a parte aleatória, informação referente a outras informações que podem influenciar o resultado

Representação alternativa

Podemos representar o modelo anterior de uma maneira diferente:

$$\mu_i = \mu + \tau_i$$

O modelo anterior fica dado como

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \, j = 1, \ldots, n.$$

- μ : média global, parâmetro comum a todos tratamento
- τ_i : efeito de tratamento, parâmetro do i-ésimo tratamento

Análise de variância

Esses dois modelos são denominados Análise de Variância (ANOVA) de fator único.

- É necessário que a alocação do material experimental às diversas condições experimentais seja aleatória
- E que o meio em que os tratamentos sejam aplicados (chamado de unidades experimentais) seja tão uniforme quanto possível
- O planejamento experimental é denominado de completamente aleatorizado
- O objetivo será o de testar hipóteses apropriadas sobre as médias dos tratamentos

Análise de um modelo com efeitos fixos

- Considere um experimento completamente aleatorizado
- A análise de variância será para um único fator com efeitos fixos
- O interesse é testar a igualdade média dos tratamentos.

As hipóteses apropriadas para isso são

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_1 : \mu_i \neq \mu_j$ para algum i e j tal que $i \neq j$

A hipótese nula supõe que as observações amostrais dentro de cada tratamento podem ser vistas como provenientes de populações com médias iguais

Representação alternativa

Se reescrevermos a média de cada tratamento como

$$\mu_i = \mu + \tau_i$$

Então, a média global pode ser escrita como

$$\mu = \frac{\sum_{i=1}^{k} \mu_i}{k}$$

Implicando que $\sum_{i=1}^{k} \tau_i = 0$. Logo podemos reescrever as hipóteses de interesse como

- $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$
- $H_1 : \tau_i \neq 0$ para algum i

Suposições do modelo

A idéia básica é de que existe uma distribuição de probabilidade para a variável resposta Y_{ij} em cada nível do fator.

Nesse caso, é necessário assumir que:

i) Y_{ij} são variáveis aleatórias independentes

ii) Y_{ij} tem distribuição normal com média μ_i

iii) $\text{Var}(Y_{ij}) = \sigma^2$, ou seja, todas as k populações devem ter var. homogêneas
$(\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2)$

- A última propriedade também é conhecida como homocedasticidade.
- Em outras palavras, a variância σ^2 deve ser constante para todos nos níveis de fator.

Decomposição da soma de quadrados

- O termo análise de variância pode induzir a um equívoco
- Investigar diferenças entre médias dos tratamentos
- E não diferenças significativas entre as variâncias dos grupos
- Vamos analisar os componentes da variância dos dados para concluir sobre as médias

A soma de quadrado total é dada por

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2$$

Que pode ser decomposta em

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

Considere novamente

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

Podemos definir essa igualdade como

$$SST = SSTrat + SSE$$

- SST: Variabilidade total dos dados
- $SSTrat$: Variabilidade entre os tratamentos
- SSE: Variabilidade entre as observações do mesmo tratamento

Considerações sobre essas somas de quadrados

Considere o segundo termo do lado direito da expressão

$$\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]$$

A soma dentro do colchete dividido por (n-1) é a variância amostral do *i*-ésimo tratamento

$$S_i^2=\frac{\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}{n-1}$$

Combinando essa variância amostral pra todos os tratamento, temos

$$\frac{(n-1)S_1^2+(n-1)S_2^2+\cdots+(n-1)S_k^2}{(n-1)+(n-1)+\cdots+(n-1)}=\frac{\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]}{\sum_{i=1}^{k}(n-1)}=\frac{SSE}{N-k}$$

Sobre estimadores para σ^2

Logo, $\frac{SSE}{N-k}$ é um estimador para σ^2.

De maneira similiar, analisando o primeiro termo de SST, temos que

$$\frac{SSTrat}{k-1} = \frac{n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2}{k-1}$$

também é um estimador para σ^2, se não existe diferença entre os tratamentos.

Assim, temos que

$$\frac{SSE}{N-k} \text{ e } \frac{SSTrat}{k-1}$$

são estimadores para σ^2 quando as médias dos tratamentos são iguais.

ANOVA

A ANOVA para testar as hipóteses:

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_1 : \mu_i \neq \mu_j$ para algum i e j tal que $i \neq j$

Pode ser resumida como

ANOVA:

Fonte de variação	GL	SQ	QM	F
Entre Tratamentos	k-1	$SSTrat$	$QMTrat = \frac{SSTrat}{k-1}$	$F_{calc} = \frac{QMTrat}{QME}$
Dentre Tratamentos	N-k	SSE	$QME = \frac{SSE}{N-k}$	
Total	N-1	SQ_{Tot}		

A estatística de teste F tem distribuição Fisher-Snedecor, com $k-1$ graus de liberdade no numerador e $N-k$ graus de liberdade no denominador.

Regras de decisão

Temos que

$$F_{calc} \sim F_{k-1;N-k}$$

- A hipótese nula deve ser rejeitada para valores grandes de F_{calc}.

O critério do teste é o seguinte então:

- Rejeita-se H_0 se $F_{calc} > F_{k-1,N-k}(\alpha)$
- $F_{k-1,N-k}(\alpha)$ é o quantil $1-\alpha$ da dist. Fisher-Snedecor$_{(k-1,N-k)}$
- Caso contrário, não rejeita-se a hipótese.

Cálculo manual dessas quantidades

Os seguintes valores podem ser utilizados para facilitar o cálculo manual dessas quantidades

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SSTrat = \frac{1}{n}\sum_{i=1}^{k} \bar{y}_{i.}^2 - \frac{y_{..}^2}{N}$$

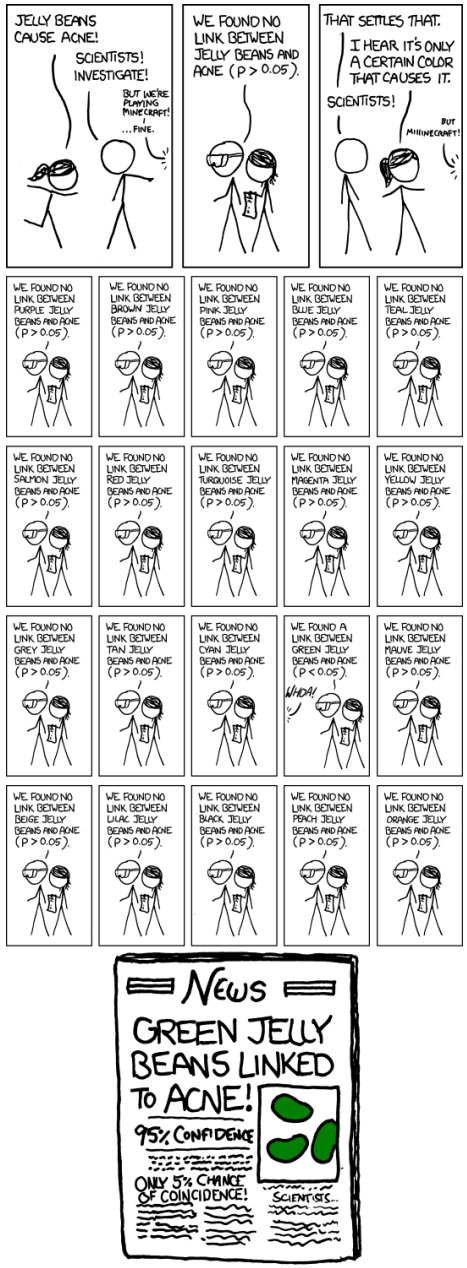
Além disso, SSE pode ser obtida com

$$SSE = SST - SSTrat$$

Teste de Tukey

- Ao se rejeitar H_0 é interessante apontar quais médias podem ser consideradas diferentes.
- Para isso é necessário fazer vários testes de hipótese de forma simultânea.
- Porém, deve-se tomar cuidado com comparações múltiplas

XKCD


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Teste de Tukey

- O teste de Tukey é uma dentre outras alternativas de se controlar essas comparações múltiplas.
- Esse teste permite testar qualquer contraste, sempre, entre duas médias de tratamentos
- Nesse caso, as hipóteses estatísticas são:
  - $H_0 : \mu_i = \mu_j$
  - $H_1 : \mu_i \neq \mu_j$,

  para todo $i \neq j$.

Formulação do teste

- O teste proposto por Tukey baseia-se na diferença significante HSD=Δ.
- HSD = Honestly Significant Difference
- Essa diferença para dados balanceados é

$$\Delta_\alpha = q_\alpha(k; f)\sqrt{\frac{QME}{n}}$$

- Duas médias, μ_i e μ_j são consideradas significativamente diferentes quando

$$|\bar{y}_{i.} - \bar{y}_{j.}| > \Delta_\alpha$$

Fazer exemplo da apostila

Análise de variância

- Suponha um experimento com k tratamentos (ou populações)
- A variável resposta de cada unidade experimental em cada tratamento é uma variável aleatória

Tratamento	Observações				Total	Média
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn}	$y_{k.}$	$\bar{y}_{n.}$
					$y_{..}$	$\bar{y}_{..}$

$$y_{i.} = \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{i.} = \frac{y_{i.}}{n}, \quad i = 1, \ldots, k.$$

$$y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{..} = \frac{y_{..}}{N}, \quad N = n \times k.$$

Modelo estatístico

Um modelo para descrever os dados é:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \; j = 1, \ldots, n.$$

- y_{ij}: observação do i-ésimo tratamento na j-ésima unidade
- μ_i: média do i-ésimo tratamento, valor fixo e desconhecido
- ϵ_{ij}: erro aleatório associado ao i-ésimo tratamento na j-ésima unidade experimental
- $\epsilon_{ij} \sim N(0, \sigma^2)$, independentes.
- $y_{ij} \sim N(\mu_i, \sigma^2)$

Análise de um modelo com efeitos fixos

- Considere um experimento completamente aleatorizado
- A análise de variância será para um único fator com efeitos fixos
- O interesse é testar a igualdade média dos tratamentos.

As hipóteses apropriadas para isso são

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_1 : \mu_i \neq \mu_j$ para algum i e j tal que $i \neq j$

A hipótese nula supõe que as observações amostrais dentro de cada tratamento podem ser vistas como provenientes de populações com médias iguais

Suposições do modelo

A idéia básica é de que existe uma distribuição de probabilidade para a variável resposta Y_{ij} em cada nível do fator.

Nesse caso, é necessário assumir que:

i) Y_{ij} são variáveis aleatórias independentes

ii) Y_{ij} tem distribuição normal com média μ_i

iii) $\text{Var}(Y_{ij}) = \sigma^2$, ou seja, todas as k populações devem ter var. homogêneas
$(\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2)$

- A última propriedade também é conhecida como homocedasticidade.
- Em outras palavras, a variância σ^2 deve ser constante para todos nos níveis de fator.

Considere novamente

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

Podemos definir essa igualdade como

$$SST = SSTrat + SSE$$

- SST: Variabilidade total dos dados
- $SSTrat$: Variabilidade entre os tratamentos
- SSE: Variabilidade entre as observações do mesmo tratamento

ANOVA

A ANOVA para testar as hipóteses:

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_1 : \mu_i \neq \mu_j$ para algum i e j tal que $i \neq j$

Pode ser resumida como

ANOVA:

Fonte de variação	GL	SQ	QM	F
Entre Tratamentos	k-1	$SSTrat$	$QMTrat = \frac{SSTrat}{k-1}$	$F_{calc} = \frac{QMTrat}{QME}$
Dentre Tratamentos	N-k	SSE	$QME = \frac{SSE}{N-k}$	
Total	N-1	SQ_{Tot}		

A estatística de teste F tem distribuição Fisher-Snedecor, com $k-1$ graus de liberdade no numerador e $N-k$ graus de liberdade no denominador.

Regras de decisão

Temos que

$$F_{calc} \sim F_{k-1;N-k}$$

- A hipótese nula deve ser rejeitada para valores grandes de F_{calc}.

O critério do teste é o seguinte então:

- Rejeita-se H_0 se $F_{calc} > F_{k-1,N-k}(\alpha)$
- $F_{k-1,N-k}(\alpha)$ é o quantil $1-\alpha$ da dist. Fisher-Snedecor$_{(k-1,N-k)}$
- Caso contrário, não rejeita-se a hipótese.

Análise de diagnóstico

Precisamos verificar se o modelo

$$y_{ij} = \mu_i + \epsilon_{ij}$$

é adequado. Para isso, devemos analisar o resíduo

$$e_{ij} = y_{ij} - \hat{y}_{ij}.$$

O valor predito é obtido como

$$\hat{y}_{ij} = \hat{\mu}_i = \bar{y}_{i.}.$$

Algumas violações do modelo podem ser observadas pelos resíduos.

Gráfico de probabilidade normal

- Histograma dos resíduos (se n grande)
- $z_i = \dfrac{e_i}{\sqrt{\text{QME}}} \approx N(0,1)$ resíduo padronizado. ± 95% dos z_i's devem estar no intervalo (-1,96;1,96).
- Um teste estatístico também pode ser usado.

Gráfico de probabilidade normal

$$z_i = \frac{e_i}{\sqrt{\text{QME}}},$$

$$F_i = \frac{\#(Z \leqslant z_i)}{n},$$

$$\Phi(x) = P(Z \leqslant x),\ Z \sim N(0,1).$$

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	

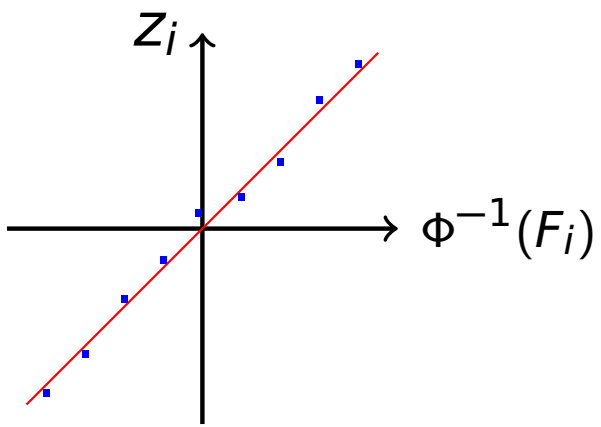


Gráfico de Resíduos Contra Ordem das Observações Coletadas

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. - - - - + + + + não é esperado.
2. + - + - + - + - + - também não é razoável.

Gráficos dos Resíduos (e_{ij}) contra os Valores Preditos (y_{ij})

Suposição verificada:

- Homogeneidade das variâncias dos erros em todos os níveis do fator:
  - dispersão dos resíduos não pode depender dos valores preditos.