Correlação e Regressão

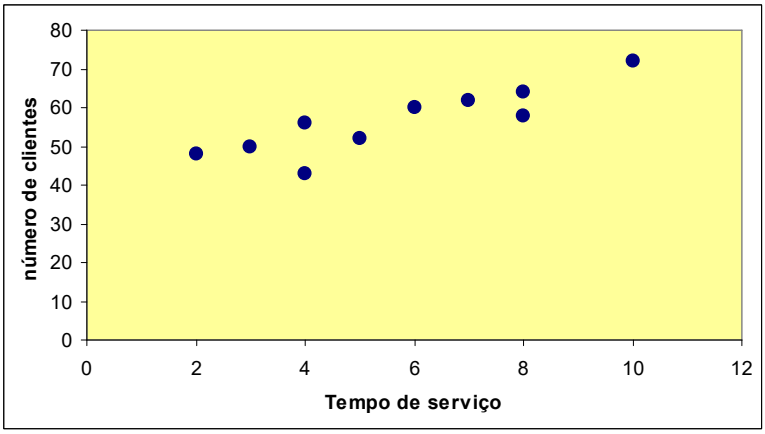
Vamos começar com um exemplo:

Temos abaixo uma amostra do tempo de serviço de 10 funcionários de uma companhia de seguros e o número de clientes que cada um possui.

Será que existe uma relação entre a variável número de clientes e tempo de serviço?

Anos de serviço	2	3	4	5	4	6	7	8	8	10
Nº de clientes	48	50	56	52	43	60	62	58	64	72

Vamos fazer um diagrama de dispersão



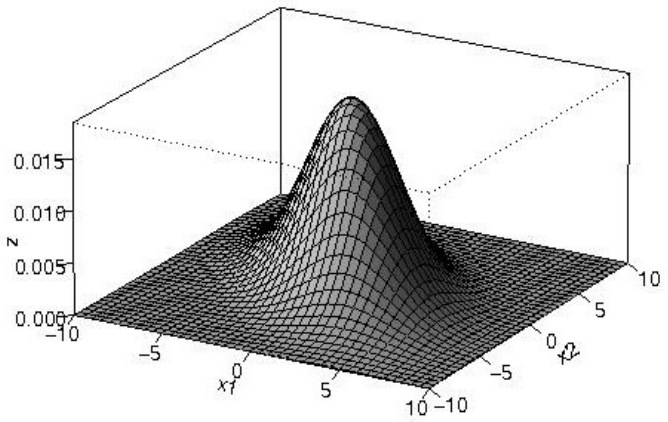
Parece haver uma relação linear entre número de clientes (y) e tempo de serviço (x).

Correlação: Existe uma correlação entre duas variáveis quando uma delas está relacionada com a outra.

Hipóteses:

Amostra aleatória de pares de dados (*x*,*y*).

Os pares (*x*,*y*) tem uma distribuição normal bivariada.



$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\sigma_{11}\,\sigma_{22}\,(1-\rho^2)}}\ \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_{11}} - 2\rho\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}}\right]\right\}$$

Coeficiente de correlação

Coeficiente de correlação (r): Mede o grau de relacionamento linear entre valores emparelhados x e y de uma amostra.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Que também pode ser escrito como:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} = \frac{\left(\sum xy\right) - n\bar{x}\bar{y}}{\sqrt{\left(\sum x^2\right) - n\bar{x}^2}\sqrt{\left(\sum y^2\right) - n\bar{y}^2}}$$

Onde n é o número de pares (x,y)

r : calculado para dados amostrais, ou seja, é uma estatística amostral.

ρ : coeficiente de correlação populacional, ou seja, se tivéssemos todos os valores (x,y) da população.

Propriedades do coeficiente de Correlação Linear r

1. O valor de r é limitado entre -1 e 1, isto é,

$$-1 \leq r \leq 1$$

2. O valor de r não varia se todos os valores de qualquer das variáveis são convertidos para uma escala diferente.

3. O valor de r não é afetado pela escolha da variável nomeada x ou y.

4. r mede a intensidade de um relacionamento linear. Não serve para medir a intensidade de um relacionamento não-linear, ou melhor, um valor pequeno de r não descarta uma relação não linear.

Coeficiente de correlação de Pearson

Procedimento para estudo:

1. Exploração dos dados: **Diagrama de dispersão**.
2. Cálculo do **coeficiente de correlação linear: *r***

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

3. Realizar o **teste de hipótese** para correlação:

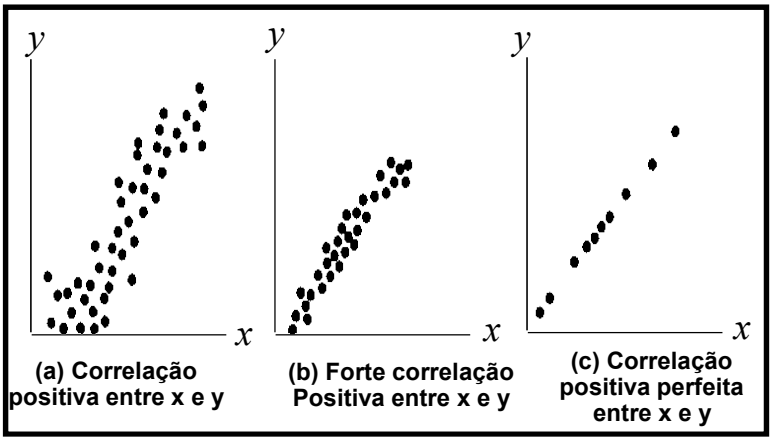
$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Estatística teste: $t_{teste} = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

Valores críticos na tabela da distribuição de Student com gl = n-2.

Correlação linear positiva



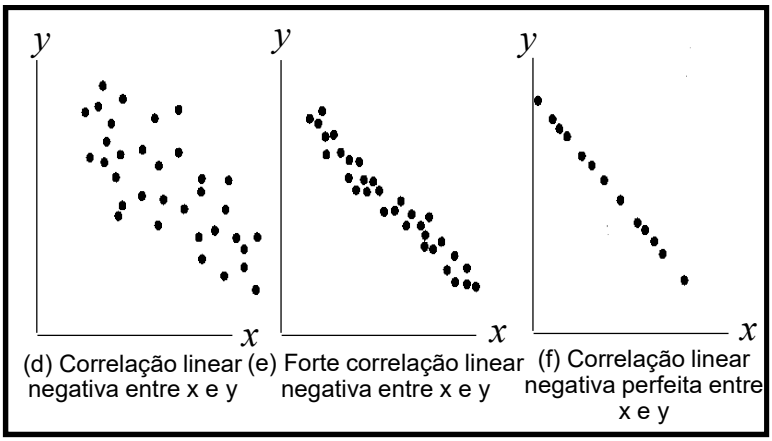
(a) Correlação positiva entre x e y

(b) Forte correlação Positiva entre x e y

(c) Correlação positiva perfeita entre x e y

Diagrama de dispersão

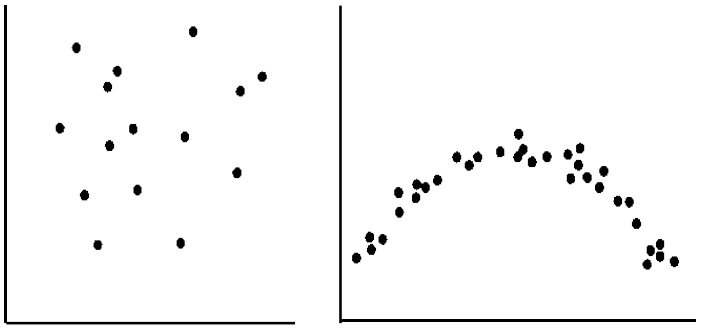
Correlação linear negativa



(d) Correlação linear negativa entre x e y

(e) Forte correlação linear negativa entre x e y

(f) Correlação linear negativa perfeita entre x e y

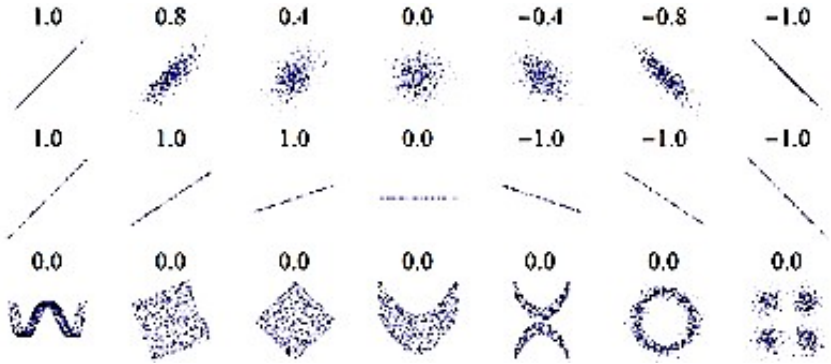


(g) Não há correlação

(h) Correlação não linear

Correlação

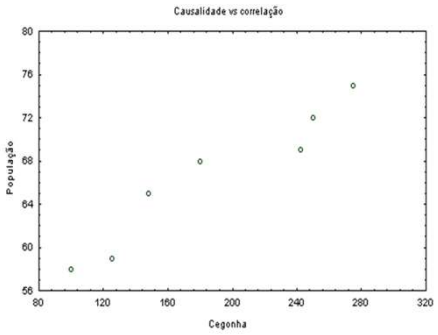
Conjuntos de pontos (x,y) com o coeficiente de correlação linear de Pearson (fonte Wikipedia).



Correlação

A correlação não implica causalidade!

Exemplo: (Box, Hunter & Hunter, Statistics for Experimenters, p.8) O gráfico mostra a população de Oldemberg, Alemanha, no fim de cada um dos 7 anos (*Y*) contra o número de cegonhas (pássaros) naquele ano (*X*).



Interpretação: existe associação entre *X* e *Y*.

Freqüentemente, quando duas *v. X e Y* parecem estar fortemente associadas, pode ser porque *X e Y* estão, de fato, associadas com uma terceira variável, *W.* No exemplo, *X e Y* aumentam com *W* = tempo.

E se os pares (*x*,*y*) não tem uma distribuição normal bivariada???

Use uma versão não-paramétrica baseada em postos.

Equação de regressão linear

Dada uma coleção de dados AMOSTRAIS emparelhados, a equação de regressão linear é dada por

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

A equação de regressão expressa uma relação entre *x* (chamada de variável independente, variável explicativa) e *y* (chamada de variável dependente, ou variável explicada)

O gráfico da equação de regressão é chamado reta de regressão (ou reta de melhor ajuste, ou reta de mínimos quadrados).

$\hat{\beta}_0$ é chamado intercepto (valor de y no qual x=0)

$\hat{\beta}_1$ é o coeficiente angular

Notação para a equação de regressão

	Parâmetro populacional	Estatísticas amostral
Intecepto y da equação de regressão	β_0	$\hat{\beta}_0$
Coeficiente angular da equação de regressão	β_1	$\hat{\beta}_1$
Equação da reta de regressão	$y = \beta_0 + \beta_1 x$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Melhor ajuste: método dos mínimos quadrados:

minimiza–se a soma das distâncias vertical entre um ponto amostral e a reta procurada : $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

Solução:

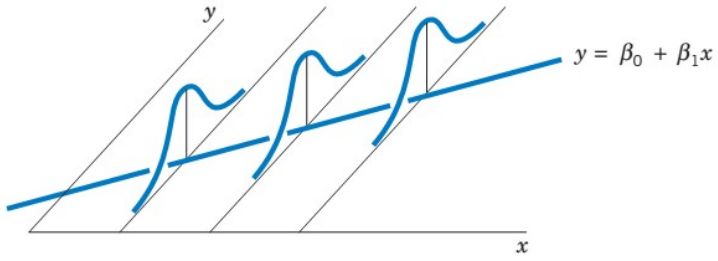
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Pressupostos

1. **Estamos considerando apenas relações lineares.**
2. **Para cada valor de x, y é uma variável aleatória com distribuição normal.**
2. **Todas as distribuições de y tem mesma variância.**
3. **Para um dado x, a distribuição dos valores de y tem uma média sobre a reta de regressão.**



Os parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$ são desconhecidos (da população) devem ser estimados a partir dos dados amostrais pelas estatísticas b_0 e b_1

$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, \ldots, n$$

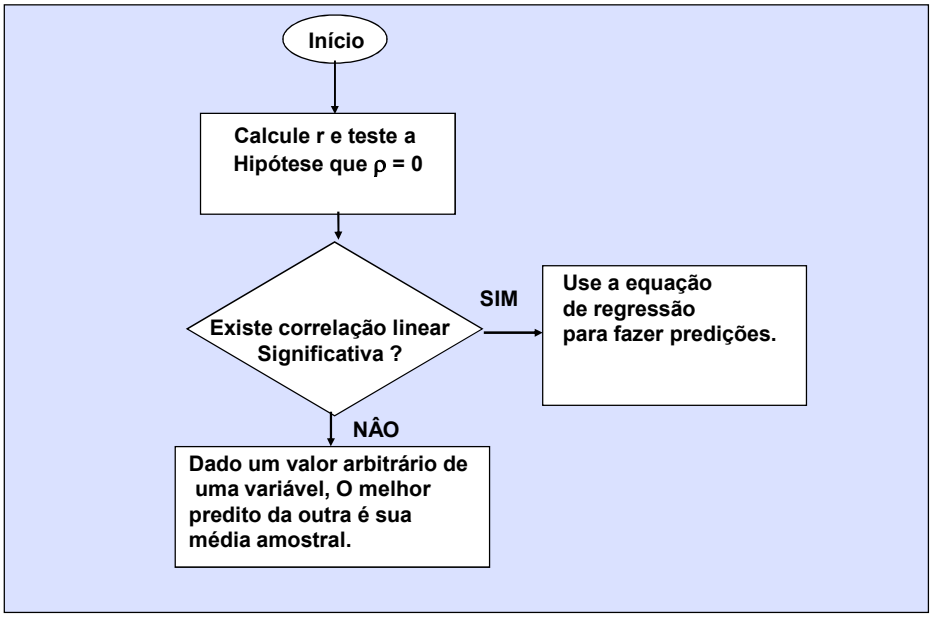
(5) $e_1, e_2, \ldots, e_n$ erros aleatórios independentes e normalmente distribuidos com média zero e desvio-padrão desconhecido

Previsões usando a equação de regressão.

Quando estimar um valor de y para um dado valor de x..

1. **Se não há correlação linear significativa, não utilize a equação de regressão para fazer predições.**
2. **Se existe correlação linear significativa, o melhor valor estimado para y é obtido substituindo-se x na equação de regressão.**
3. **Ao aplicar a equação de regressão para predições, mantenha-se dentro do intervalo dos dados amostrais.**
4. **Não devemos fazer estimativas sobre uma população diferente daquela de onde provém os dados.**
5. **Uma equação de regressão baseada em dados passados não é necessariamente válida hoje ou no futuro.**

Estimativa do valor de uma variável



Exemplo: Novamente a amostra do tempo de serviço de 10 funcionários de uma companhia de seguros e o número de clientes que cada um possui.

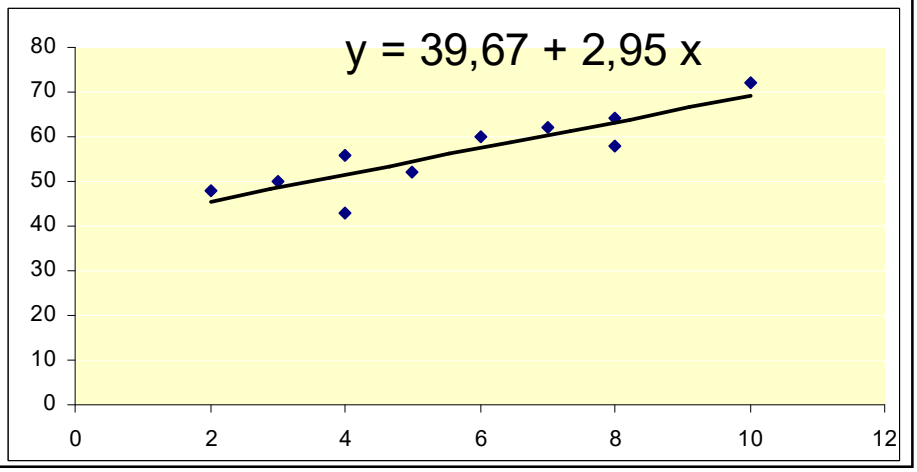
Será que existe uma relação entre a variável número de clientes e tempo de serviço?

Anos de serviço	2	3	4	5	4	6	7	8	8	10
Nº de clientes	48	50	56	52	43	60	62	58	64	72

r = 0,88

$\hat{\beta}_1 = 2,95$

$\hat{\beta}_0 = 39,67$



Para x = 7,5 anos de serviço, qual a estimativa de clientes?

Temos que testar se r =0,88 é significativo.

Ho: $\rho = 0$

H1: $\rho \neq 0$

$$t_{teste} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,88}{\sqrt{\frac{1-0,88^2}{10-2}}} = 5,24$$

Para α = 0,05 e gl = n-2=10-2=8 temos: tc = ± 2,31

Rejeitamos Ho. Logo a correlação é significativa.

Portanto, podemos usar a equação de regressão para estimar o número de clientes. Temos

$$y = 39,67 + 0,88 \cdot 7,5 = 46,27$$

RESÍDUO: é a diferença $(y - \hat{y})$ entre um valor amostral y e um valor estimado $\hat{y}$ a partir da equação de regressão.

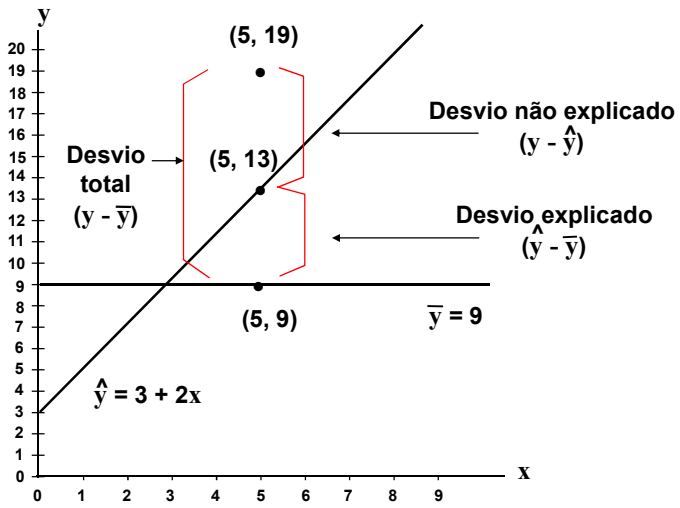
Desvio total (em relação a média): $y - \bar{y}$

Desvio explicado: $\hat{y} - \bar{y}$

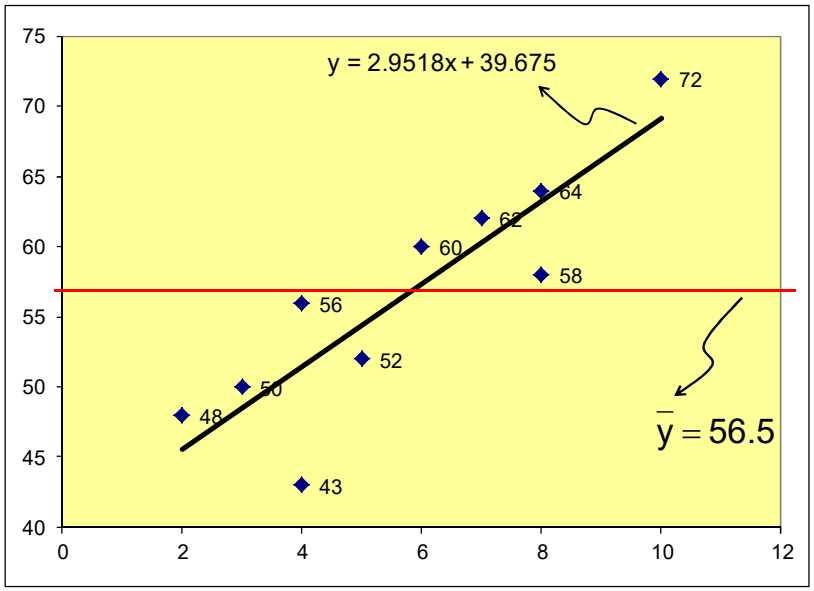
Desvio não explicado (resíduos e_i): $y - \hat{y}$

onde $\bar{y} = \frac{\sum_i y_i}{n}$

Exemplo:



Considere nosso exemplo inicial



para x = 6 temos:

$y = 60$ (ponto amostral)

$\hat{y} = 39{,}67 + 2{,}95.(6) = 57{,}37$ (valor estimado)

$\bar{y} = \dfrac{\sum y}{n} = 56{,}5$

Desvio total : $(y - \bar{y}) = 60 - 56{,}5 = 3{,}5$

Desvio explicado : $(\hat{y} - \bar{y}) = 57{,}37 - 56{,}5 = 0{,}87$

Desvio não explicado (resíduo) : $(y - \hat{y}) = 60 - 57{,}37 = 2{,}63$

Para um ponto (x,y) particular temos:

(desvio total) = (desvio explicado) + (desvio não explicado)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

A variação total será obtida da soma dos quadrados do desvio total, a variação explicada da soma dos quadrados do desvio explicado e a variação não explicada da soma dos quadrados dos resíduos.

(variação total) = (variação explicada) + (variação não explicada)

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

Coeficiente de deteminação (r^2): Valor da variação de y que é explicado pela reta de regressão

$$R^2 = r^2 = \frac{\sum(\hat{y} - \overline{y})^2}{\sum(y - \overline{y})^2} = \frac{\text{variação explicada}}{\text{variação total}} = \frac{S_{xy}^2}{S_{xx}\,S_{yy}}$$

Que é simplesmente o coeficiente de correlação ao quadrado.

Para nosso exemplo inicial temos r = 0,88, e

$r^2 = (0{,}88)^2 = 0{,}7744$

ou seja, 77,44% da variação total de y é explicada pela reta de regressão. Decorre que 22,66% da variação total de y permanece não explicado.

Soma dos quadrados dos erros (Sum of Square due Errors)

$$\text{SSE} = \sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Erro padrão da estimativa é uma medida de quanto os pontos amostrais se afastam da reta de regressão (desvio-padrão)

$$s = \sqrt{\frac{\text{SSE}}{n-2}}$$

Inferência em regressão

(1) Coeficiente angular

Estimador: $\hat{\beta}_1$

Erro padrão da estimativa: $\frac{S}{\sqrt{S_{xx}}}$

Degrees of freedom (graus de liberdade)

Distribuição amostral: student $T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}}$ $\text{d.f.} = n - 2$

Intervalo de (1-α)% de confiança par β_1: $\hat{\beta}_1 \pm t_{\alpha/2} \frac{S}{\sqrt{S_{xx}}}$

Teste de hipótese para β_1: $H_0 : \beta_1 = \beta_{10}$ Não necessariamente zero

Estatística teste: student $T = \frac{\hat{\beta}_1 - \beta_{10}}{S/\sqrt{S_{xx}}}$ $\text{d.f.} = n - 2$

(2) Intercepto

Estimador: $\hat{\beta}_0$

Erro padrão da estimativa: $S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

Distribuição amostral: student $T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$ d.f. $= n - 2$

Intervalo de (1-α)% de confiança para β_0: $\hat{\beta}_0 \pm t_{\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

Teste de hipótese para β_o: $H_0 : \beta_0 = \beta_{00}$

Estatística Teste: $T = \frac{\hat{\beta}_0 - \beta_{00}}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$ d.f. $= n - 2$

(3) Resposta média esperada de y (média de y) para um dado valor de x = x*: $E(Y/x^*) = \beta_0 + \beta_1 x^*$

Estimador: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Erro padrão da estimativa: $S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

Intervalo de (1-α)% de confiança para $E(Y/x^*)$:

$$\hat{y} \pm t_{\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

(4) Previsão para a resposta de um y para um dado valor de x = x*

Estimador: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

Erro padrão da estimativa: $S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

Valor esperado para resposta única y:

$$\hat{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Tabela ANOVA para regressão: a qualidade da regressão estimada pode ser Analisada por meio de uma análise de variância (ANOVA)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Regression	SSR	1	SSR	$\frac{SSR}{s^2}$
Error	SSE	$n-2$	$s^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

SSR: Sum of Squared due Regression (explicada)
SSE: Sum os equare due Errors (resíduos)
SST= SSR+SSE (Total)

Ho: $\beta_1 = 0$
H1: $\beta_1 \neq 0$

Rejeite Ho ao nível de significância α se

$$f = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{s^2} > f_\alpha(1, n-2)$$

Onde o valor crítico ou valor p é obtido da distribuição-F

Voltemos ao exemplo inicial

Resultados do excel: nível de confiança = 95%, α=0,05

Estatísticas t

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	39,6746988	3,542145989	11,200752	3,6186E-06
X	2,95180723	0,572357094	5,15728251	0,00086654

β_0

β_1

$\frac{S}{\sqrt{S_{xx}}}$

$S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$

$T = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}}$

$T = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}}$

y = 2.9518x + 39.675

Soma dos quadrados

média da soma dos quadrados

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	506,2349398	506,23494	26,5975629	0,000866543
Resíduo	8	152,2650602	19,0331325		
Total	9	658,5			

SSR

SSE

SST

SSR/1

SSE/(n-2)

$f = \frac{SSR/1}{SSE/(n-2)}$

Alguns modelos não lineares podem ser linearizados por meio de transformações das variáveis. E o modelo de regressão linear pode ser aplicado às variáveis transformadas.

TABLE 3 Some Nonlinear Models and Their Linearizing Transformations

			Transformed Model
Nonlinear Model	Transformation		$y' = \beta_0 + \beta_1 x'$
(a) $y = ae^{bx}$	$y' = \log_e y$	$x' = x$	$\beta_0 = \log_e a$ $\beta_1 = b$
(b) $y = ax^b$	$y' = \log y$	$x' = \log x$	$\beta_0 = \log a$ $\beta_1 = b$
(c) $y = \frac{1}{a + bx}$	$y' = \frac{1}{y}$	$x' = x$	$\beta_0 = a$ $\beta_1 = b$
(d) $y = a + b\sqrt{x}$	$y' = y$	$x' = \sqrt{x}$	$\beta_0 = a$ $\beta_1 = b$

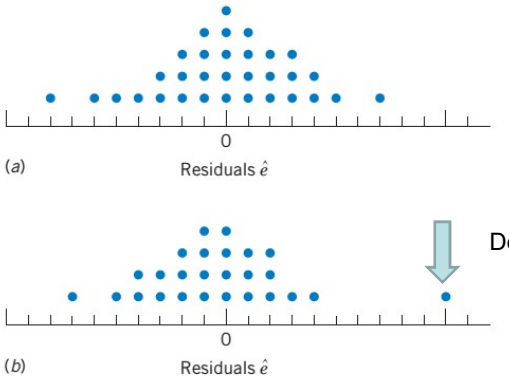
Variáveis transformadas são facilmente obtidas com softwares.

Todos os procedimentos de inferência no modelo de regressão dependem das hipóteses sob o qual o modelo é construído, ou seja:

(1) Relação linear
(2) Independência dos erros
(3) Variância constante
(4) Distribuição normal

Adequação do modelo estatístico: examinar os Resíduos é importante pois ajuda a detectar inconsistências entre os dados e as hipóteses do modelo.

Faça um histograma, diagrama de pontos, ou um Normal-score gráfico dos resíduos para verificar normalidade: O modelo assume distribuição normal



Deve ser investigado

Figure 5 Dot diagram of residuals.
(*a*) Normal pattern. (*b*) One large residual.

Score-normal plot dos resíduos: deve ser aproximadamente linear

	score normal	resíduos (ei)
1	-1.50	-5
2	-1.11	-3
3	-0.84	-2
4	-0.62	-2
5	-0.43	-1
6	-0.25	-1
7	-0.08	0
8	0.08	0
9	0.25	0
10	0.43	1
11	0.62	2
12	0.84	3
13	1.11	3
14	1.50	5

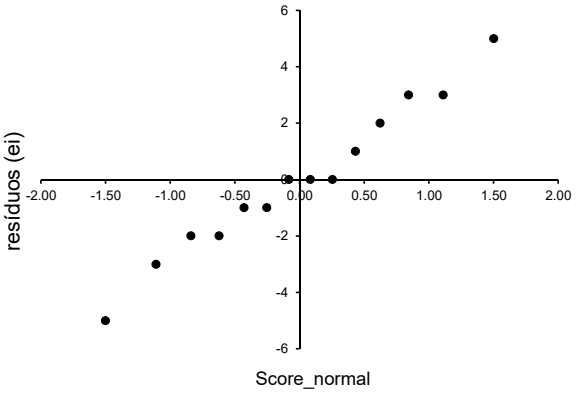
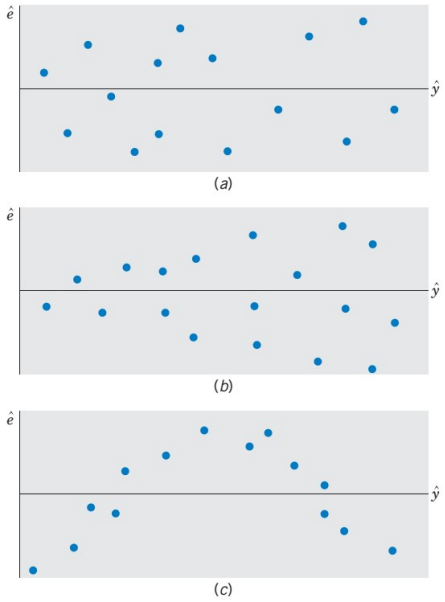


Gráfico: resíduos ($\hat{e}_i$) x valores previstos ($\hat{y}_i$)



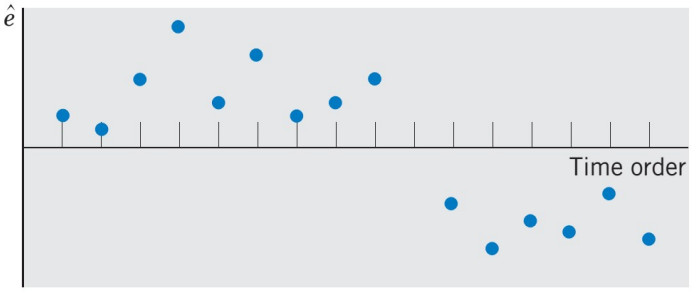
Aleatóriamente distribuido em torno de $\hat{y}_i$ com variância constante: **OK**

Variância não constante: Hipótese do modelo violada

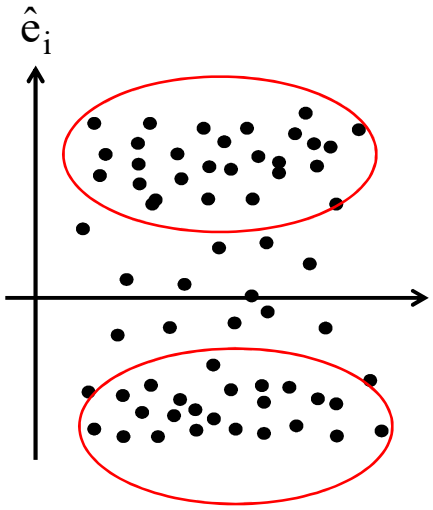
Padrão sistemático: talvez Um modelo não linear seja mais adequado

Gráfico: resíduos ($\hat{e}_i$) x tempo (t_i)

Importante para observações coletadas numa ordem temporal (série temporal)



Padrão indica a a violação de independência



Maior incidência de valores nos extremos indicando uma possível violação da hipótese de normalidade dos resíduos.

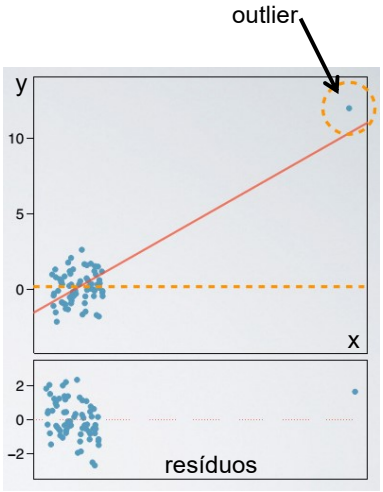
Outliers: são pontos distantes da nuvem da maioria dos pontos.

Na figura, observa-se que sem o outlier não existe relação entre x e y.

Se os valores estimados mudam significativamente quando um outlier é removido, ele é chamado um **ponto influente.**

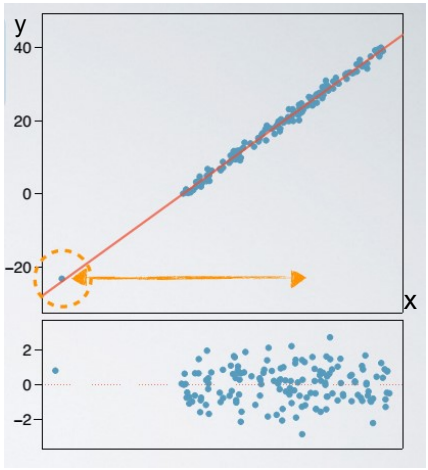
Nem todo outlier é influente.

Sugestão: verifique se os valores discrepantes não são erros de medida.

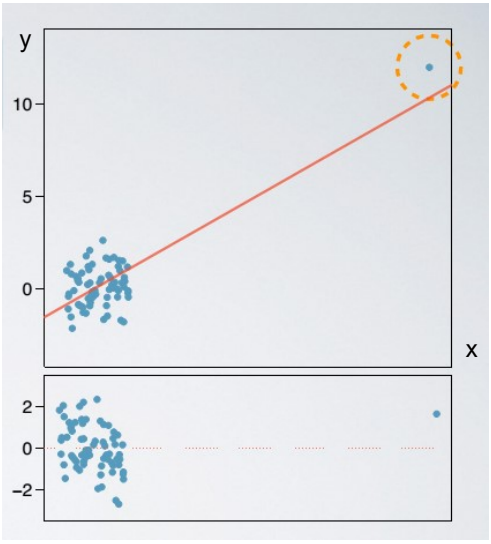


Alavancas (leverage points): têm valor não usual da variável explicativa. Tem potencial de ser influentes
Nem toda alavanca é influente.
Mas um ponto influente é um outlier e/ou uma alavanca.

alavanca



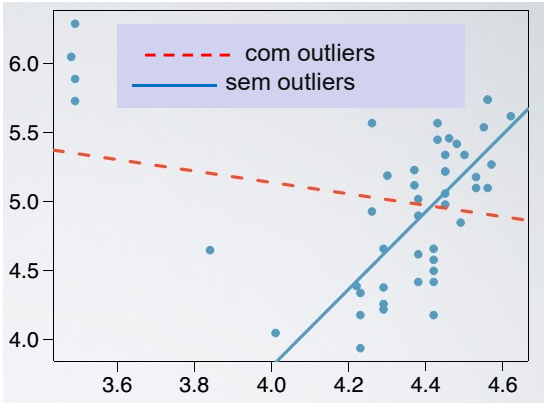
Ponto influente



exemplo:

intensidade luminosa
e temperatura de superfície
no cluster de estrelas
CYG OBI

melhor estudar
os grupos separadamente



Atenção:

- Esteja atento aos dados que você está analisando: *Amostra aleatória, amostra não aleatória, população.*
- Inferência estatística e os valores p resultantes não tem sentido se os dados correspondem a população.
- Se a amostra não é aleatória, os resultados não são confiáveis.