Estatística aplicada
Licenciatura em Engenharia Informática

Lino Costa

Departamento de Produção e Sistemas
Escola de Engenharia
lac@dps.uminho.pt

Revisto em 2022/2023

Sumário

1. Regressão linear simples e múltipla
   - modelo populacional e modelo estimado
   - resíduos
   - gráfico de resíduos
   - coeficiente de determinação
   - intervalo de confiança e teste de hipótese para o declive
   - análise de variância para testar o declive
2. Correlação
   - coeficiente de correlação populacional
   - coeficiente de correlação amostral
   - teste de correlação de Pearson

Regressão linear simples

Análise de regressão

A análise de regressão é um técnica estatística para modelar a relação entre duas ou mais variáveis. Pretende-se, usando um modelo, prever a resposta de uma variável para um determinado valor de outra variável (regressão simples) ou variáveis (regressão múltipla).

Modelo populacional de regressão linear simples

Modelo que explica a relação linear entre duas variáveis (x e Y):

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

- β_0 e β_1 são os coeficientes de regressão que representam, respetivamente, a interseção com o eixo das ordenadas e o declive da reta de regressão. O declive β_1 mede a alteração esperada em Y por cada alteração de uma unidade de x.
- x é a variável independente (preditora).
- Y é a variável dependente (resposta) com $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$.
- ε são erros aleatórios que representam a variação de Y relativamente à reta de regressão $\beta_0 + \beta_1 x$ com $\varepsilon \sim N(0, \sigma^2)$.

Regressão linear simples

Modelo estimado de regressão linear simples

Os parâmetros β_0 e β_1 podem ser estimados pelo “Método dos Mínimos Quadrados” a partir dos n pares de obervações (x_i, y_i) com $i = 1, \ldots, n$, obtendo-se o modelo de regressão linear simples estimado:

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

onde

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{e} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Resíduo

Aplicando a reta de regressão linear simples estimada para x_i obtém-se a estimativa $\hat{y}_i$ (o valor médio de Y para $x = x_i$). A diferença entre o valor real y_i e o previsto $\hat{y}_i$ é chamado de resíduo e_i que é uma estimativa do erro aleatório ε_i:

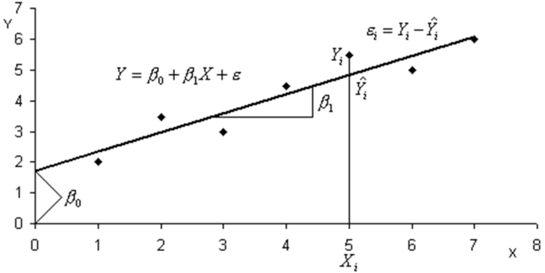
$$e_i = y_i - \hat{y}_i \quad \text{para} \quad i = 1, \ldots, n$$

Regressão linear simples

Variância dos resíduos

A estimativa da variância dos resíduos σ^2 é dada por

$$\hat{\sigma}^2 = s^2 = MQR = \frac{SQR}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$



Regressão linear simples

Exemplo 1

Com base nos dados da tabela, determine o modelo de regressão linear simples para prever a concentração de um anestésico ($\mu L/g$) a partir do tempo de injeção ($min.$).

Tempo ($min.$)	Conc. ($\mu L/g$)
4	106
8	105
12	170
16	240
20	210
24	280
28	310

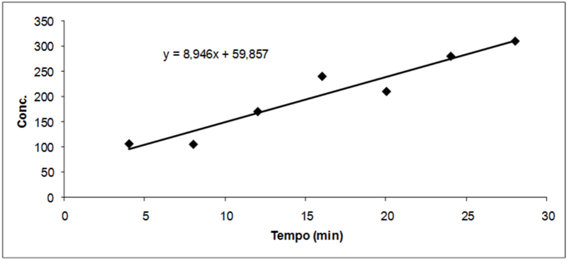
- variável dependente: concentração de um anestésico ($\mu L/g$)
- variável independente: tempo de injeção ($min.$)
- modelo populacional: $Y = \beta_0 + \beta_1 x + \varepsilon$ com $\varepsilon \sim N(0, \sigma^2)$
- modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ com $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ e $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Regressão linear simples

Exemplo 1

Tempo	Conc.	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})y_i$
4	106	-12	144	-1272
8	105	-8	64	-840
12	170	-4	16	-680
16	240	0	0	0
20	210	4	16	840
24	280	8	64	2240
28	310	12	144	3720
112	1421	0	448	4008

- $n = 7$, $\bar{x} = \frac{112}{7} = 16$ e $\bar{y} = \frac{1421}{7} = 203$
- $\hat{\beta}_1 = \frac{4008}{448} = 8.946$ e $\hat{\beta}_0 = 203 - 8.946 \times 16 = 59.857$
- modelo estimado: $\hat{u} = 59.857 + 8.946x$

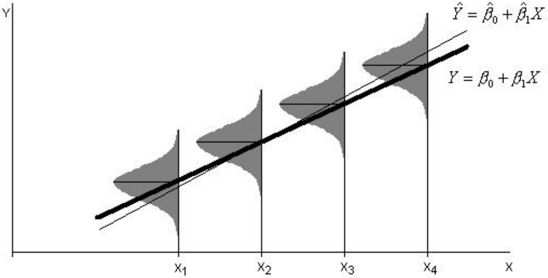


Regressão linear simples

Análise de resíduos

Os resíduos devem ser independentes e com distribuição Normal com média 0 e variância constante σ^2:

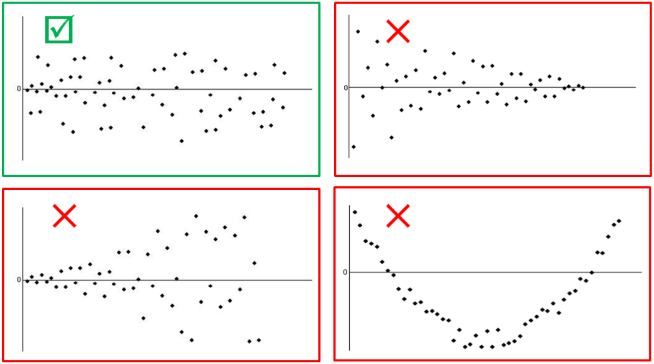
$$\varepsilon \sim N(0, \sigma^2)$$



Regressão linear simples

Gráfico de resíduos

A representação gráfica dos resíduos e_i versus $\hat{y}_i$ ou e_i versus x_i permitem visualizar se são aleatórios e com variância constante. Um padrão de distribuição dos resíduos nestes gráficos pode indicar que o modelo de regressão linear não é adequado.

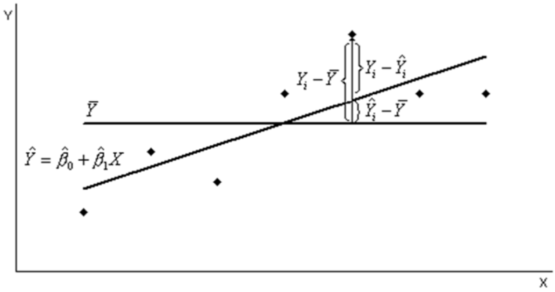


Regressão linear simples

Coeficiente de determinação

O coeficiente de determinação (R^2) indica a proporção da variância da variável dependente que é explicada pelo modelo de regressão. Os valores de R^2 pertencem ao intervalo de 0 a 1 (quanto maior, mais explicativo é o modelo):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$



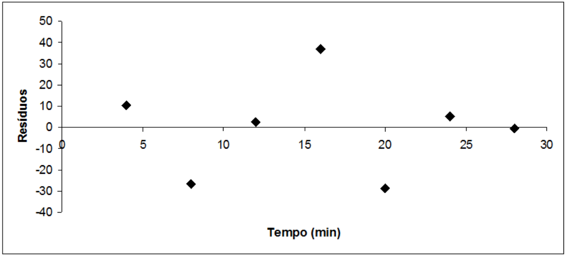
Regressão linear simples

Avaliação dos pressupostos do modelo

- **normalidade** dos resíduos pode ser verificada
  - graficamente: gráfico quantil-quantil
  - teste de hipótese: teste não paramétrico Kolmogorov-Smirnov
- **média nula** dos resíduos pode ser verficada pelo intervalo de confiança para a média dos resíduos
- **variância constante e independência** dos resíduos podem ser verificadas pelo gráfico de resíduos e_i versus $\hat{y}_i$ ou e_i versus x_i

Exemplo 1

- Coeficiente de determinação: $R^2 = 0.922$ pelo que a proporção de variabilidade da variável dependente explicada pelo modelo é de 92.2%.
- Gráfico de resíduos indica que os resíduos são aleatórios e têm variância constante pelo que o modelo se considera adequado.



Regressão linear simples

Intervalo de confiança para o declive (β_1)

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Teste de hipótese para o declive (β_1)

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

$$E.T. : T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim t_{n-2}$$

$$R.R. : |T| > t_{\alpha/2,n-2}$$

Nota: a não rejeição de H_0 pode significar a não existência de relação linear entre a variável dependente e a variável independente (mas pode existir outro tipo de relação).

Regressão linear simples

Análise de variância para testar β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Nota: no caso de um modelo de regressão linear simples, o resultado do teste ANOVA é idêntico ao teste de hipóteses ao declive.

Região de rejeição

$R.R. : F > F_{\alpha,1,n-2}$ onde α é o nível de significância.

Tabela ANOVA e estatística de teste F

Fonte	SQ	gl	MQ	F
Explicada	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$\sum_{i=1}^{n}\hat{\beta}_1(\hat{x}_i - \bar{x})^2$	$F = \frac{\sum_{i=1}^{n}\hat{\beta}_1(\hat{x}_i - \bar{x})^2}{s^2}$
Resíduos	$\sum_{i=1}^{n}(y_i - \hat{y})^2$	$n-2$	s^2	
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

Regressão linear simples

Notas importantes

- **validade prática da regressão**: a relação entre a variável dependente e independente deve ter significado prático
- **validade empírica/teórica da regressão**: o sinal e a magnitude dos coeficientes do modelo devem ser comparados com resultados anteriores (empíricos ou teóricos)
- **associação entre variáveis na regressão**: confirmar a existência de uma relação causa-efeito entre a variável independente e variável dependente
- **limites do modelo**: o modelo só é válido para valores da variável independente na gama de valores dos dados usados na construção do modelo (isto é, não é válido generalizar ou extrapolar o modelo)
- **ajuste do modelo**: analisar a distribuição dos resíduos (e_i) e calcular o coeficiente de determinação (R^2)
- **significância estatística do modelo**: testar a significância do modelo de regressão (teste ao coeficiente β_1 ou análise de variância a β_1)

Regressão linear simples

Exemplo 2

Determine o modelo de regressão linear simples (comprimento alar em cm em função da idade em dias) para os dados relativos a andorinhas.
Estime o comprimento alar para uma idade de 7 dias e para 15 dias.
Analise a qualidade do ajuste e a significância estatística do modelo.

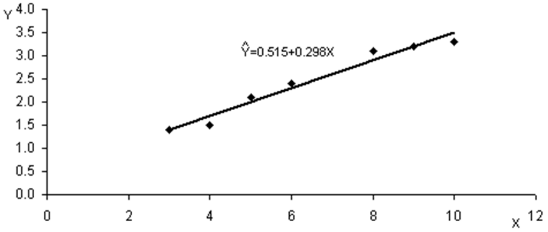
Idade (dias)	Comprimento alar (cm)
3	1.4
4	1.5
5	2.1
6	2.4
8	3.1
9	3.2
10	3.3

- variável dependente: comprimento alar (cm)
- variável independente: idade (dias)
- modelo populacional: $Y = \beta_0 + \beta_1 x + \varepsilon$ com $\varepsilon \sim N(0, \sigma^2)$
- modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ com $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ e $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Regressão linear simples

Exemplo 2

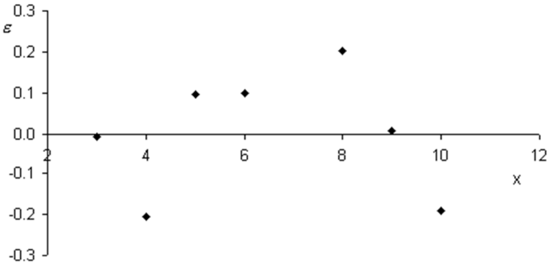
- $n = 7$, $\hat{\beta}_1 = 0.515$ e $\hat{\beta}_0 = 0.298$
- modelo estimado: $\hat{y} = 0.515 + 0.298x$
- valor esperado para Y quando $x = 7$, $E(Y) = 0.515 + 0.298 \times 7 = 2.601$
- valor esperado para Y quando $x = 15$, o modelo não deve ser usado para extrapolar (só deve ser utilizado para valores de x entre 3 dias e 10 dias).



Regressão linear simples

Exemplo 2

- Coeficiente de determinação: $R^2 = 0.964$ pelo que a proporção de variabilidade da variável independente explicada pelo modelo é de 96.4%.
- Gráfico de resíduos indica que os resíduos são aleatórios e têm variância constante pelo que o modelo se considera adequado.



Regressão linear simples

Exemplo 2

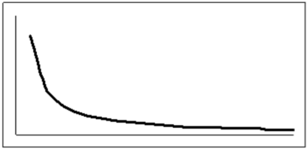
- Intervalo de confiança de 95% para o declive: $0.231 \leq \beta_1 \leq 0.364$ (não inclui o 0 pelo que o modelo linear é significativo).
- Teste de hipótese para o declive ($\alpha = 5\%$): $H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$, $E.T. : T = 11.497$, $R.R. : |T| > t_{0.025,5} = 2.571$ (tabela 6), rejeita-se H_0 para $\alpha = 5\%$ pelo que o modelo linear é significativo.
- Análise de variância para o declive ($\alpha = 5\%$): $H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$
-
Fonte	SQ	gl	MQ	F
Explicada	3.695	1	3.695	$F = 132.174$
Resíduos	0.140	5	0.028	
Total	3.834	6		

- $R.R. : F > F_{0.05,1,5} = 6.61$ (tabela 8), rejeita-se H_0 para $\alpha = 5\%$ pelo que o modelo linear é significativo.

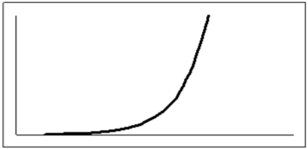
Regressão linear simples

Transformação de modelos não lineares em lineares

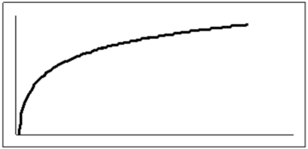
Algumas relações não lineares entre a variável dependente (Y) e a variável independente (x) podem ser transformadas matematicamente num modelo linear.



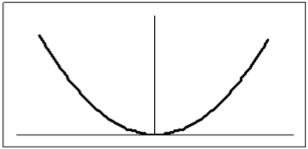
$\hat{Y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{x} \Leftrightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 z$ com $z = \frac{1}{x}$



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^x \Leftrightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 z$ com $z = e^x$



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x \Leftrightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 z$ com $z = \ln x$



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^2 \Leftrightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 z$ com $z = x^2$

Regressão linear múltipla

Modelo populacional de regressão linear múltipla

Modelo que explica a relação linear entre Y e k variáveis x_1 a x_k:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

- β_0 β_1 a β_k são os coeficientes de regressão (regressores) que representam, respetivamente, a interseção com o eixo das ordenadas e os declives dos preditores.
- x_1 a x_k são as k variáveis independentes (preditores).
- Y é a variável dependente com $Y \sim N(\beta_0 + \beta_1 x + \ldots + \beta_k x_k, \sigma^2)$.
- ε são erros aleatórios que representam a variação de Y relativamente à reta de regressão $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ com $\varepsilon \sim N(0, \sigma^2)$.
- os parâmetros β_0 a β_k podem ser estimados pelo "Método dos Mínimos Quadrados" pela resolução de um sistema de equações lineares cuja solução na forma matricial é: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- estimativa da variância: $\hat{\sigma}^2 = s^2 = \dfrac{\mathbf{y}'\mathbf{y} - \hat{\beta}\mathbf{X}'\mathbf{y}}{n-k-1}$

Regressão linear múltipla

Coeficiente de determinação

O coeficiente de determinação (R^2) indica a proporção da variância da variável dependente que é explicada pelo modelo de regressão. Os valores de R^2 pertencem ao intervalo de 0 a 1 (quanto maior, mais explicativo é o modelo). Adicionando mais variáveis independentes (preditoras), aumenta o valor de R^2 sem necessariamente melhorar o modelo.

Coeficiente de determinação ajustado

O coeficiente de determinação ajustado (R^2_{ajustado}) considera o número de regressores ($k+1$) no modelo:

$$R^2_{\text{ajustado}} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

Para comparar modelos com número diferente de regressores é preferível ter em conta o R^2_{ajustado} em vez do R^2.

Regressão linear múltipla

Teste de hipótese para a constante (β_0)

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

$$E.T. : T_0 \sim t_{n-k-1} \qquad R.R. : |T_0| > t_{\alpha/2,n-k-1}$$

Teste de hipótese para os declives (β_j com $j = 1, \ldots, k$)

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

$$E.T. : T_j \sim t_{n-k-1} \qquad R.R. : |T_j| > t_{\alpha/2,n-k-1}$$

Nota: a não rejeição de H_0 pode significar a não existência de relação linear entre a variável dependente e a variável independente x_j.

Regressão linear múltipla

Análise de variância para testar β_1

$H_0 : \beta_1 = \ldots = \beta_k = 0$

$H_1 : \beta_j \neq 0 \;\; \exists j = 1, \ldots, k$

Região de rejeição

$R.R. : F > F_{\alpha,k,n-k-1}$ onde α é o nível de significância.

Tabela ANOVA e estatística de teste F

Fonte	SQ	gl	MQ	F
Explicada	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	$\hat{\beta}\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$	$F = \frac{\hat{\beta}\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}{s^2}$
Resíduos	$\sum_{i=1}^{n}(y_i - \hat{y})^2$	$n-k-1$	s^2	
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

Regressão linear múltipla

Exemplo 3

Determine a relação existente entre o calor envolvido no endurecimento, representado pela variável Y e os pesos de duas substâncias X_1 e X_2, tendo em consideração os seguintes valores obtidos numa experiência:

Y	X_1	X_2
78.5	7	26
74.3	1	29
104.3	11	59
87.6	11	31
95.6	7	52
109.2	11	55
102.7	3	71
72.5	1	31
93.1	2	54
115.9	21	47

- modelo populacional: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ com $\varepsilon \sim N(0, \sigma^2)$
- modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 51.2691 + 1.5173 x_1 + 0.6752 x_2$
- Coeficiente de determinação: $R^2 = 0.977$ pelo que a proporção de variabilidade da variável independente explicada pelo modelo é de 97.7%.
- Coeficiente de determinação ajustado: $R^2_{\text{ajustado}} = 0.9705$ (este é o valor que deve ser considerado na comparação de modelos com número de regressores diferente).

Correlação

Coeficiente de correlação da população

Se X e Y forem variáveis aleatórias e n observações (X_i, Y_i) com $i = 1, \ldots, n$ são obtidas a partir de uma distribuição Normal bivariada, então o coeficiente de correlação populacional é ρ e representa a relação linear normalizada entre X e Y.

Coeficiente de correlação amostral

O estimador deρ é o coeficiente de correlação amostral (R):

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

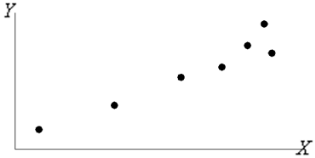
- $-1 \leq R \leq 1$ mede o grau da correlação ou associação entre duas variáveis quantitativas (de escala proporcional ou intervalar)
- $R = 1$ - correlação positiva perfeita entre as duas variáveis (uma aumenta e a outra também aumenta linearmente)
- $R = -1$ - correlação negativa perfeita entre as duas variáveis (uma aumenta, a outra diminui)
- $R = 0$ - as duas variáveis não estão associadas linearmente (no entanto, pode existir uma associação não linear)

Correlação

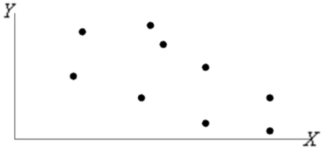
Exemplo 4

Analise cada um dos seguintes gráficos de dispersão de duas variáveis aleatórias X e Y.

- $R = 0.93$ - correlação positiva forte (uma aumenta e a outra também aumenta linearmente)
- $R = -0.62$ - correlação negativa moderada (uma aumenta e a outra diminui linearmente)
- $R = 0$ - correlação nula (não estão associadas linearmente)
- $R = 0$ - correlação nula (não estão associadas linearmente, mas existe associação não linear)

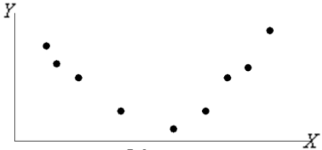


R=0.93



R=-0.62

R=0



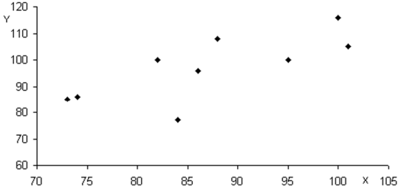
R=0

Correlação

Exemplo 5

O Índice de Desenvolvimento de Griffiths (IDR)de crianças é dado pelo coeficiente de correlação para as avaliações motora e intelectual.
Represente o gráfico de dispersão e determine o IDR para a seguinte amostra de 9 crianças com a idade de 4 anos.

Motor	Intelectual
84	77
73	85
101	105
74	86
88	108
100	116
86	96
95	100
82	100



Correlação

Exemplo 5

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
84	77	-3	-20	9	400	60
73	85	-14	-12	196	144	168
101	105	14	8	196	64	112
74	86	-13	-11	169	121	143
88	108	1	11	1	121	11
100	116	13	19	169	361	247
86	96	-1	-1	1	1	1
95	100	8	3	64	9	24
82	100	-5	3	25	9	-15
783	873	0	0	830	1230	751

- $n = 9$, $\bar{X} = 87$ e $\bar{Y} = 97$
- $R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{751}{\sqrt{830 \times 1230}} = 0.743$
- o coeficiente de correlação amostral é positivo (associação positiva linear)
- o IDR é de 74.3%

Correlação

Teste de correlação de Pearson

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

$$E.T. : T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}$$

$$R.R. : |T| > t_{\alpha/2,n-2}$$

Exemplo 5

Assumindo a normalidade, verifique se a correlação obtida é significativa ($\alpha = 5\%$).

- Teste de correlação de Pearson ($\alpha = 5\%$): $H_0 : \rho = 0; H_1 : \rho \neq 0$, $E.T. : T = 2.939$, $R.R. : |T| > t_{0.025,7} = 2.365$ (tabela 6), rejeita-se H_0 para $\alpha = 5\%$ pelo que a correlação é significativa.