Medidas Estatísticas

ASSOCIAÇÃO: COVARIÂNCIA E CORRELAÇÃO LINEAR SIMPLES

Análise de Correlação e medidas de associação

CET083 – PROBABILIDADE E ESTATÍSTICA

PROFESSOR JOSÉ CLÁUDIO FARIA

SETEMBRO DE 2014

IAGO FARIAS

Roteiro

Análise exploratória de dados

Medidas estatísticas

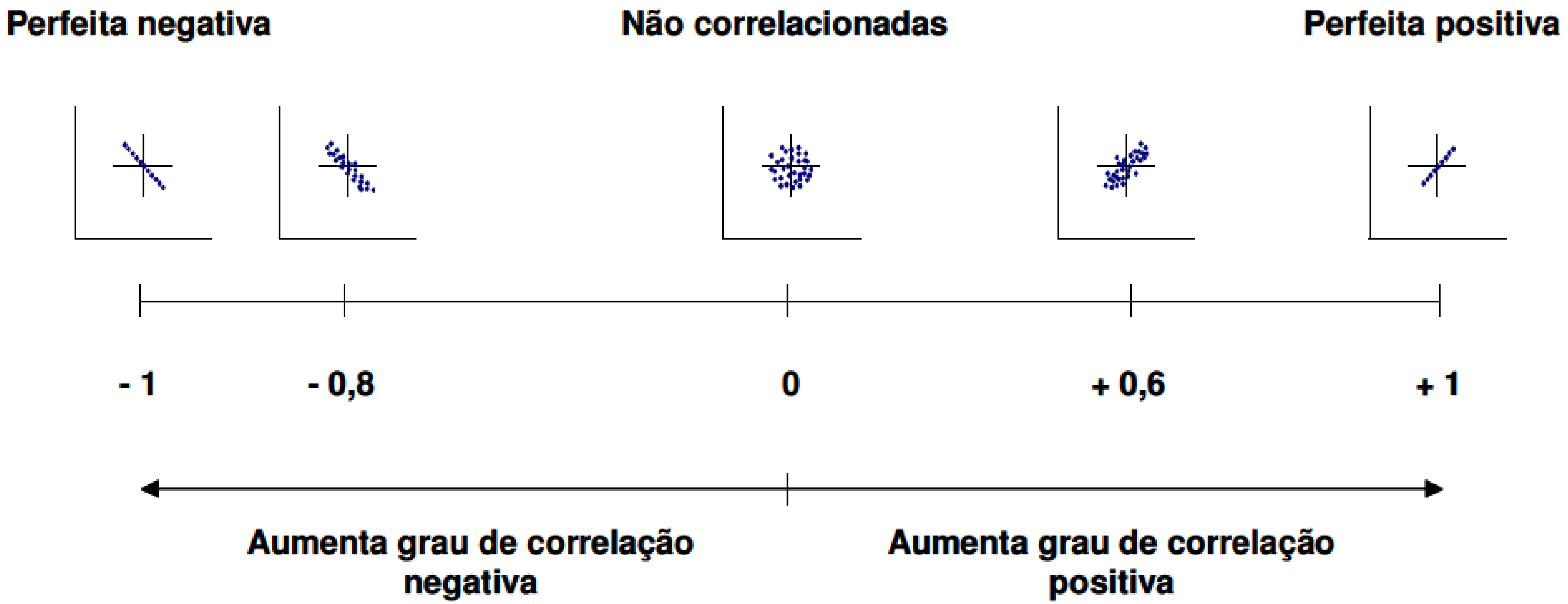
Associação

Intro

Avaliar o grau de relacionamento linear entre duas ou mais variáveis

- Quanto uma variável interfere na outra? Qual a sua dependência?
  - Técnicas de correlação

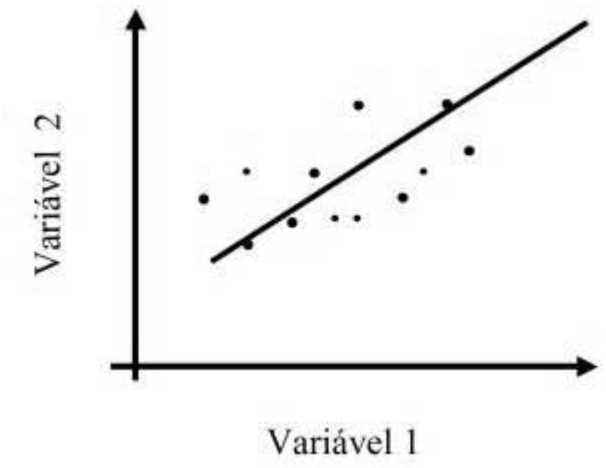
Grau de associação linear



Diagramas de Dispersão

Coleção de pontos em que as coordenadas cartesianas (x,y) são valores de cada membro do par de dados

- Qual a necessidade de um diagrama?
  - Análise de tendências
  - Mudanças de espalhamento de uma variável em relação à outra
  - Análise de valores discrepantes



Covariância

Medida de relação linear entre duas variáveis:

Cov(X,Y) = E[(X-E(X))(Y-E(Y))];

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1},$$

Compreensão a partir de um exemplo

Duas variáveis aleatórias:

- M : rendimento acadêmico em matemática
- L : rendimento acadêmico em línguas

Tabela 1 – Rendimento acadêmico em matemática (M) e línguas (L) do curso X da Universidade Y - 2014

Obs	1	2	3	4	5	6	7	8
M	36	80	50	58	72	60	56	08
L	35	65	60	39	48	44	48	61

- $\sum m = 480$
- $m(M) = 60$
- $s(M) = 13{,}65$

- $\sum l = 400$
- $m(L) = 50$
- $s(L) = 10{,}93$

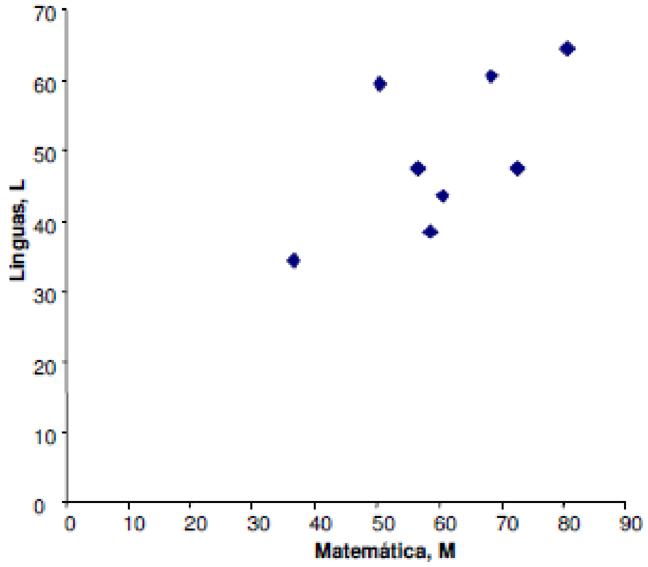


Figura 1 – Gráfico de dispersão entre M e L.

- Novo gráfico, com os eixos das médias m e l sobreposto

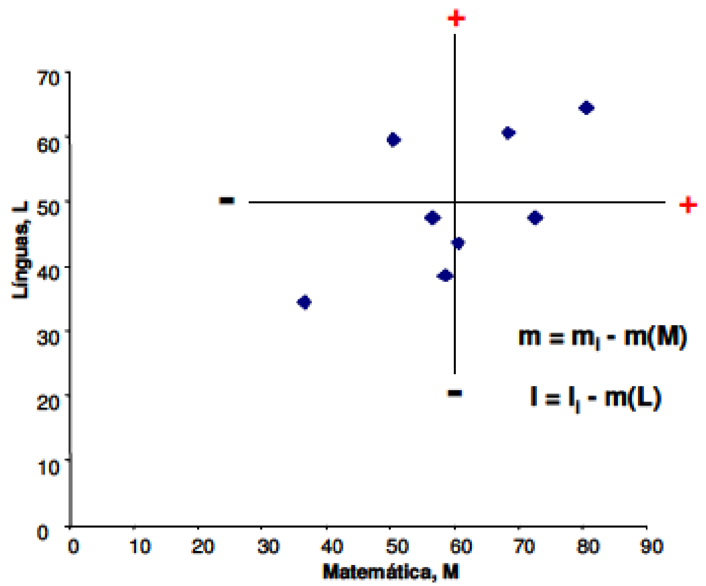


Figura 2 – Gráfico de dispersão entre M e L com médias transladadas.

- Grau de associação entre as duas variáveis aleatórias?
  - $\sum ml$
    - Onde $m = m_i - \mathrm{m(M)}$ e $l = l_i - m(L)$

Tabela 2 – Cálculo do índice $\sum ml$

Obs	M	L	m = (M_i – m(M))	l = (L_i – m(L))	m.l
1	36	35	- 24	- 15	360
2	80	65	20	15	300
3	50	60	- 10	10	- 100
4	58	39	- 2	- 11	22
5	72	48	12	- 2	- 24
6	60	44	0	- 6	0
7	56	48	- 4	- 2	8
8	68	61	8	11	88
	m(M) = 60 s(M) = 13,65	m(L) = 50 s(L) = 10,93			Σml = 654

Interpretação de resultados

- Observações sobre o resultado de ml.
  - $ml > 0$
  - $ml < 0$
  - $ml \cong 0$
- Observações sobre o resultado de $\sum ml$.
  - $\sum ml > 0$
  - $\sum ml < 0$
  - $\sum ml \cong 0$

- Sinal representa a associação corretamente!
- E se a amostra tivesse o dobro do tamanho?
  - $2 * \sum ml$? A tendência também dobra?

- $\frac{\sum ml}{n-1} = \frac{1}{n-1}[\sum(M_i - m(M))\text{x}(L_i - m(L)]$
  - Divisão pelo tamanho da amostra

- Nova medida: Correlação.
  - Unidade de medidas das variáveis envolvidas? (pés e polegadas, metros e milhas)
    - Padronização de unidades -> Dividir m e l pelos seus respectivos desvios-padrões s(M) e s(L)
- $\frac{1}{n-1}\sum(\frac{m}{s(M)})(\frac{l}{s(L)}) = \frac{1}{n-1}\left[\sum\left(\frac{M_i - m(M)[un]}{s(M)[un]}\right)\left(\frac{L_i - m(L)[un]}{s(L)[un]}\right)\right],$

  onde $s(M) = \sqrt{\sum_{i=1}^{n}\frac{(x_i - \overline{x})^2}{n-1}}$ e $s(L) = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \overline{y})^2}{n-1}}$

- $\frac{\sum ml}{n-1} = \frac{1}{n-1}[\sum(M_i - m(M))\text{x}(L_i - m(L)]$
  - Divisão pelo tamanho da amostra
- Nova medida: Correlação.
  - Unidade de medidas das variáveis envolvidas? (pés e polegadas, metros e milhas)
    - Padronização de unidades -> Dividir m e l pelos seus respectivos desvios-padrões s(M) e s(L)
- $\frac{1}{n-1}\sum(\frac{m}{s(M)})(\frac{l}{s(L)}) = \frac{1}{n-1}\left[\sum\left(\frac{M_i - m(M)[\cancel{un}]}{s(M)[\cancel{un}]}\right)\left(\frac{L_i - m(L)[\cancel{un}]}{s(L)[\cancel{un}]}\right)\right],$

  onde $s(M) = \sqrt{\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}}$ e $s(L) = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{n-1}}$
- $\boxed{r = \frac{cov(M,L)}{s(M)s(L)}}$ -> Correlação de Pearson

- Covariância:
  - Não é influenciado pelo tamanho da amostra, entretanto influenciado pelas unidades de medida das variáveis
- Coeficiente de correlação:
  - Não é influenciado nem pelo tamanho, nem pelas unidades de medida das variáveis

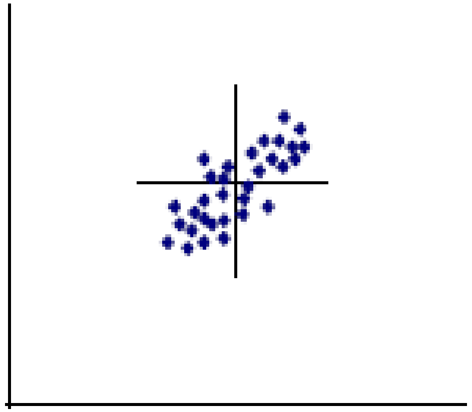
- Pressupõe-se da correlação que:
  - Relacionamento linear
  - Variáveis aleatórias e intervalares ou proporcionais (nunca categóricas ou nominais)
  - Distribuição normal bivariada

- Teorema
  - Se X e Y forem independentes, então não são correlacionadas, isto é,

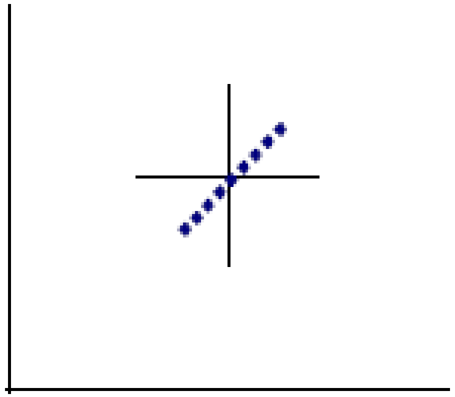
$$p_{(x,y)} \rightarrow r_{(x,y)} = 0$$

- $cov(M,L) = \sum \frac{(M_i - m(M))(L_i - m(L))}{n-1} = \frac{654}{7} = 93{,}43$
- $r(M,L) = \frac{cov(M,L)}{s(M)s(L)} = \frac{93{,}43}{13{,}65 * 10{,}93} = 0{,}63$
  - Obs.: $-1 \leq r \leq 1$
- $r^2 = 0{,}63^2 = 0{,}3922$
- $r^2 = 39{,}22\%$
  - A variação observada em M é explicada pela variação em L, e vice-versa.
    - Interpretação dos resultados.

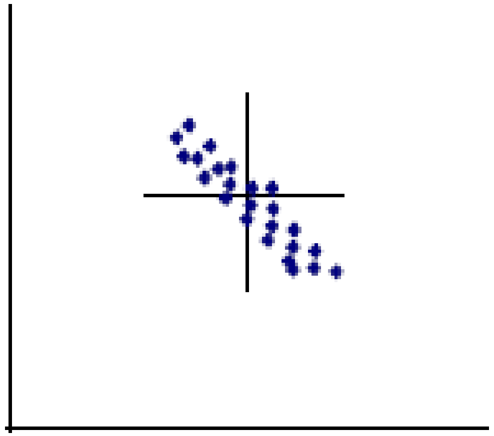
Coeficiente de Correlação	Correlação
$r = 1$	Perfeita positiva.
$0,8 <= r < 1$	Forte positiva
$0,5 <= r < 0,8$	Moderada positiva
$0,1 <= r < 0,5$	Fraca positiva
$0 <= r < 0,1$	Íntima positiva
$r = 0$	Nula
$0 <= r < -0,1$	Íntima negativa
$-0,1 <= r < -0,5$	Fraca negativa
$-0,5 <= r < -0,8$	Moderada negativa
$-0,8 <= r < -1$	Forte negativa.
$r = -1$	Perfeita negativa.



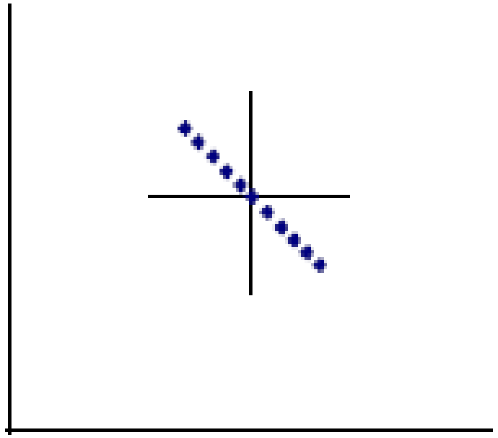
$r = 0{,}6$



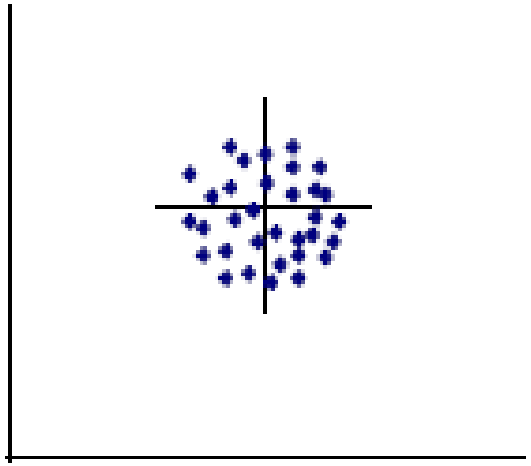
$r = 1$



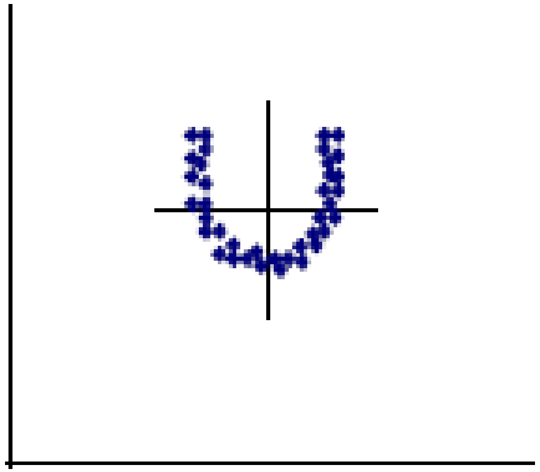
$r = -0{,}8$



$r = -1$



$r = 0$



$r = 0$

Obs.: existe relação, mas não é linear.

Funções em R

- cov(x, y, na.rm, use, method, V)

x	a numeric vector, matrix or data frame.
y	NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).
na.rm	logical. Should missing values be removed?
Use	an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings"everything", "all.obs", "complete.obs", "na.or.complete", or"pairwise.complete.obs".
Method	a character string indicating which correlation coefficient (or covariance) is to be computed. One of "pearson" (default), "kendall", or "spearman", can be abbreviated.
V	symmetric numeric matrix, usually positive definite such as a covariance matrix.

- cor(x, y, na.rm, use, method, V)

x	a numeric vector, matrix or data frame.
y	NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).
na.rm	logical. Should missing values be removed?
use	an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings"everything", "all.obs", "complete.obs", "na.or.complete", or"pairwise.complete.obs".
method	a character string indicating which correlation coefficient (or covariance) is to be computed. One of "pearson" (default), "kendall", or "spearman", can be abbreviated.
V	symmetric numeric matrix, usually positive definite such as a covariance matrix.

Considerações

- Predição e análise exploratória
- Pressupõe-se da correlação:
  - Relacionamento linear
  - Variáveis aleatórias medidas nas escalas intervalar ou proporcional (nunca categórica ou nominal)
  - Distribuição normal bivariada
- Análise da concordância, porém não estabelece relação causa-efeito, nem permite previsões
- Covariância fortemente influenciado por *outliers*
- Correlação é uma técnica menos poderosa que a análise de regressão

Referências

- GUIMARÃES, Paulo Ricardo B. ***Análise de Correlação e medidas de associação***. DEST/UFPR, 2013.

- BUSSAB, Wilton O & MORETTIN, Pedro A. **Estatística Básica**. São Paulo, Saraiva, 5 ed. 2004.

- Slides referentes a apresentação do grupo de 2014.1

- FARIA, José Cláudio. **Notas de aulas expandidas** – Ilhéus, UESC/DCET, 10 ed. 2009.