Covariância e Coeficiente de Correlação

Guilherme de Alencar Barreto

gbarreto@ufc.br

Grupo de Aprendizado de Máquinas – GRAMA
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
http://lattes.cnpq.br/8902002461422112

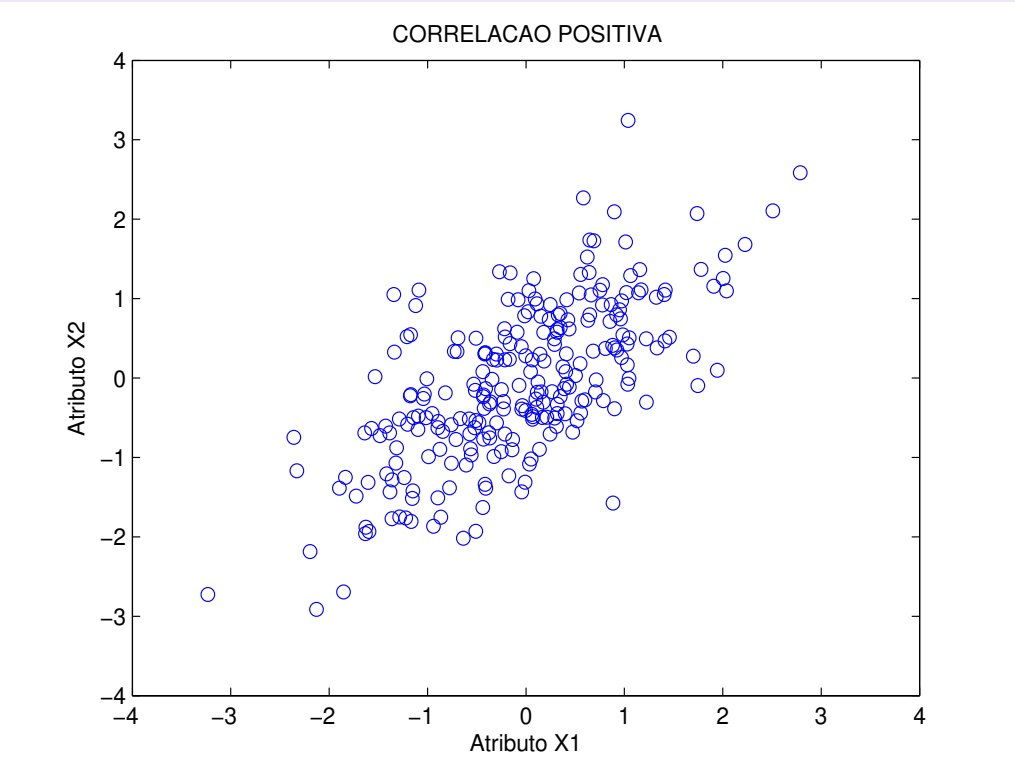
Objetivos

Objetivo: Entender como duas variáveis estão interrelacionadas do ponto de vista estatístico.

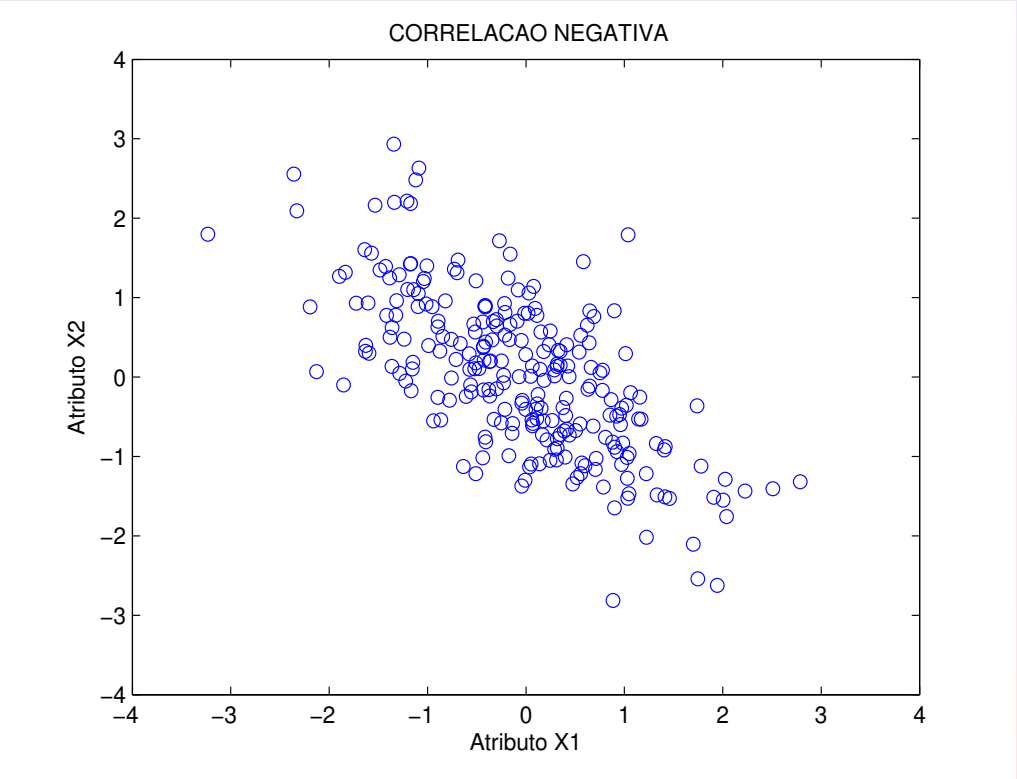
- **Método 1:** Gráfico de dispersão (qualitativo).
- **Método 2:** Coeficiente de Correlação (quantitativo).

Objetivo: Discutir relação como modelo de regressão linear simples.

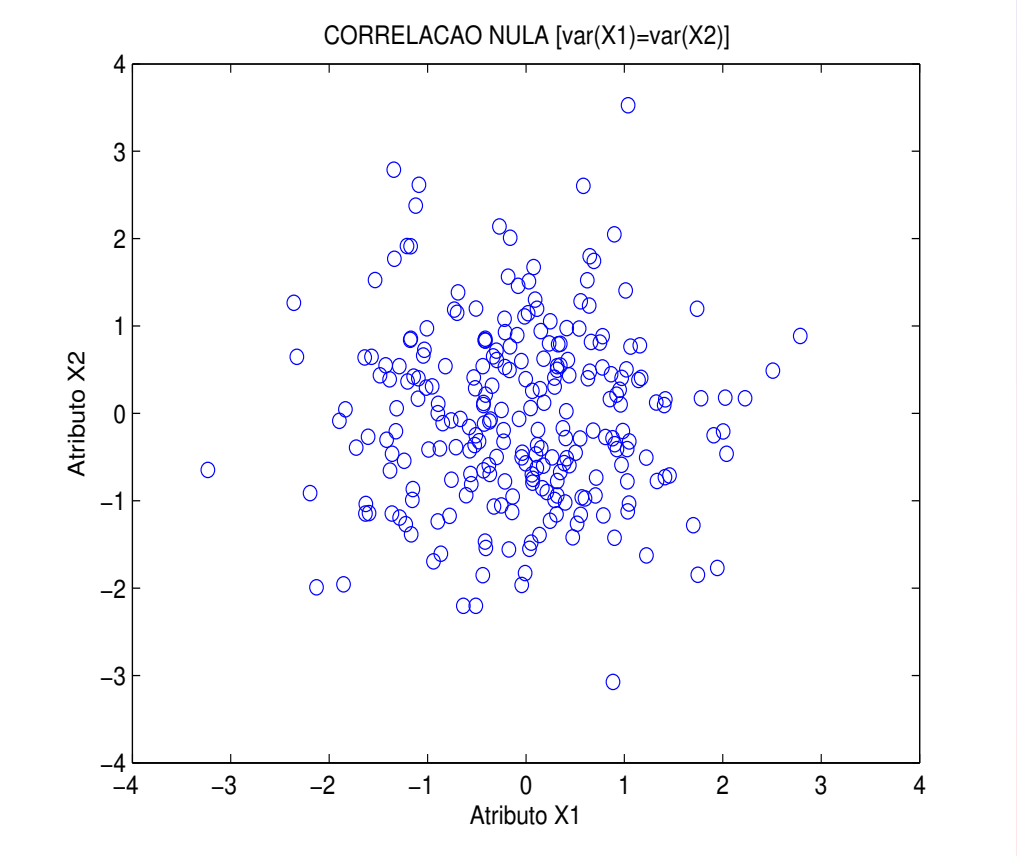
- **Correlação Positiva**: quando os valores de uma das variáveis aumentam/diminuem, os valores da outra variável tendem a aumentar/diminuir.



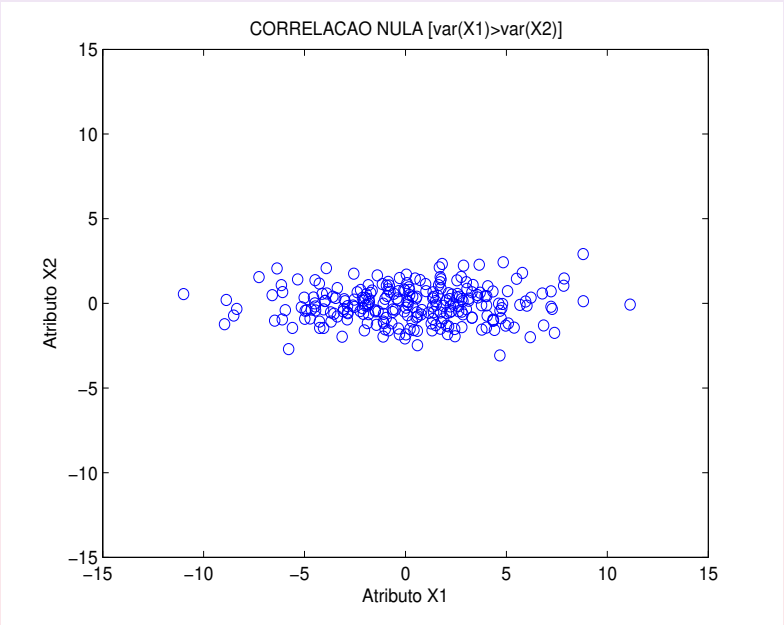
- **Correlação Negativa**: quando os valores de uma das variáveis aumentam/diminuem, os valores da outra variável seguem uma tendência contrária.

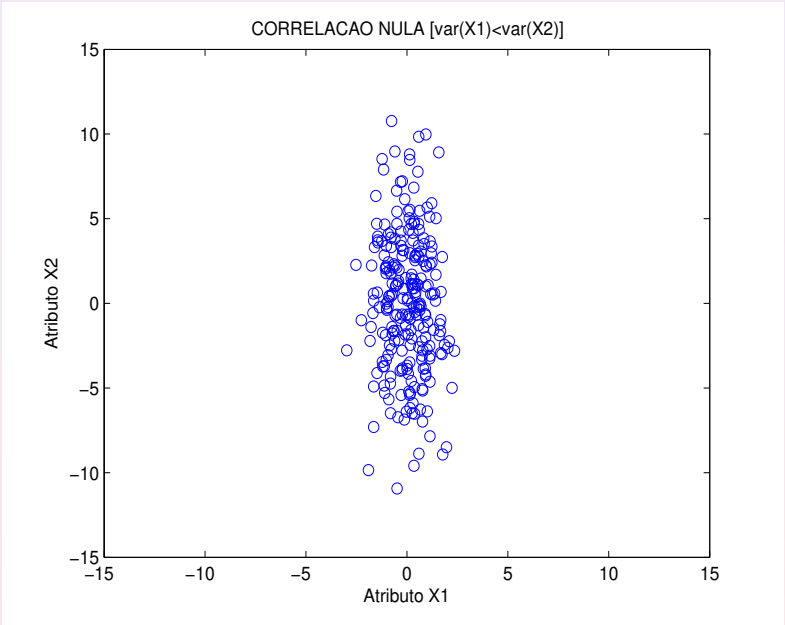


- **Correlação Nula [Caso 1: var(X1)=var(X2)]**: não há um padrão definido de tendência.



- **Correlação Nula [Caso 2: var(X1) $\neq$ var(X2)]**: também ocorre quando ao aumentar um atributo não há mudança significativa nos valores do outro atributo.





- A variância de uma variável aleatória é uma medida de sua dispersão (i.e. espalhamento) em torno de seu valor médio.
- Para um conjunto de N observações de X qualquer, a variância amostral pode ser calculada por

$$\boxed{\hat{\sigma}_x^2 = \frac{\sum_{n=1}^{N}(x(n) - \bar{x})^2}{N-1}} \quad \text{ou} \quad \boxed{\hat{\sigma}_x^2 = \frac{\sum_{n=1}^{N}(x(n) - \bar{x})^2}{N}}, \tag{1}$$

  em que $x(n)$ é a n-ésima observação de X e $\bar{x}$ é a média amostral de X calculada como $\bar{x} = \sum_{n=1}^{N} x(n)/N$.
- O "chapéu" ($\wedge$) no símbolo da variância amostral indicada se tratar de uma estimativa. A primeira expressão é preferível para pequenas amostras (i.e., N pequeno).
- O desvio padrão de X é definido como $\sigma_x = \sqrt{\sigma_x^2}$.

- Para N observações de (X_1, X_2), a covariância amostral entre X_1 e X_2 é dada por

$$\boxed{\hat{\sigma}_{12} = \frac{\sum_{n=1}^{N}(x_1(n) - \bar{x}_1)(x_2(n) - \bar{x}_2)}{N}} \quad (2)$$

em que $x_1(n)$ e $x_2(n)$ denotam as observações conjuntas de X_1 e X_2, respectivamente, enquanto $\bar{x}_1$ e $\bar{x}_2$ são as médias amostrais correspondentes.

- Seja $d_i = x_i - \bar{x}_i$ O desvio de um atributo em relação à sua média. Assim, a covariância entre duas variáveis aleatórias nada mais é do que a média dos produtos dos seus respectivos desvios:

$$\boxed{\hat{\sigma}_{12} = \frac{\sum_{n=1}^{N} d_1(n) \cdot d_2(n)}{N}} \quad (3)$$

- Se $X_1 = X_2$ ou $X_1 = -X_2$, então $\hat{\sigma}_{12} = \hat{\sigma}_1^2 = \hat{\sigma}_2^2$. Ou seja, a covariância amostral reduz-se à variância amostral.

- Assim como no caso da variância amostral, pode-se usar $N - 1$ no denominador da Eq. (2) em vez de N.

- Dá-se preferência a $N - 1$ para pequenas amostras, ou seja, para N pequeno.

Conjunto de dados: Câncer de Pulmão

Amostra	País	Cigarros per capita	Morte por milhão
1	Australia	480	180
2	Canada	500	150
3	Denmark	380	170
4	Finland	1100	350
5	Great Britain	1100	460
6	Iceland	230	60
7	Netherlands	490	240
8	Norway	250	90
9	Sweden	300	110
10	Switzerland	510	250

Tabela: Consumo per capita de cigarros em vários países em 1930 e mortes por milhão devido a câncer de pulmão em 1950

Fonte: D. Freedman, R. Pisani & R. Purves (2007), "Statistics", 4a. edição.

Conjunto de dados: Câncer de Pulmão

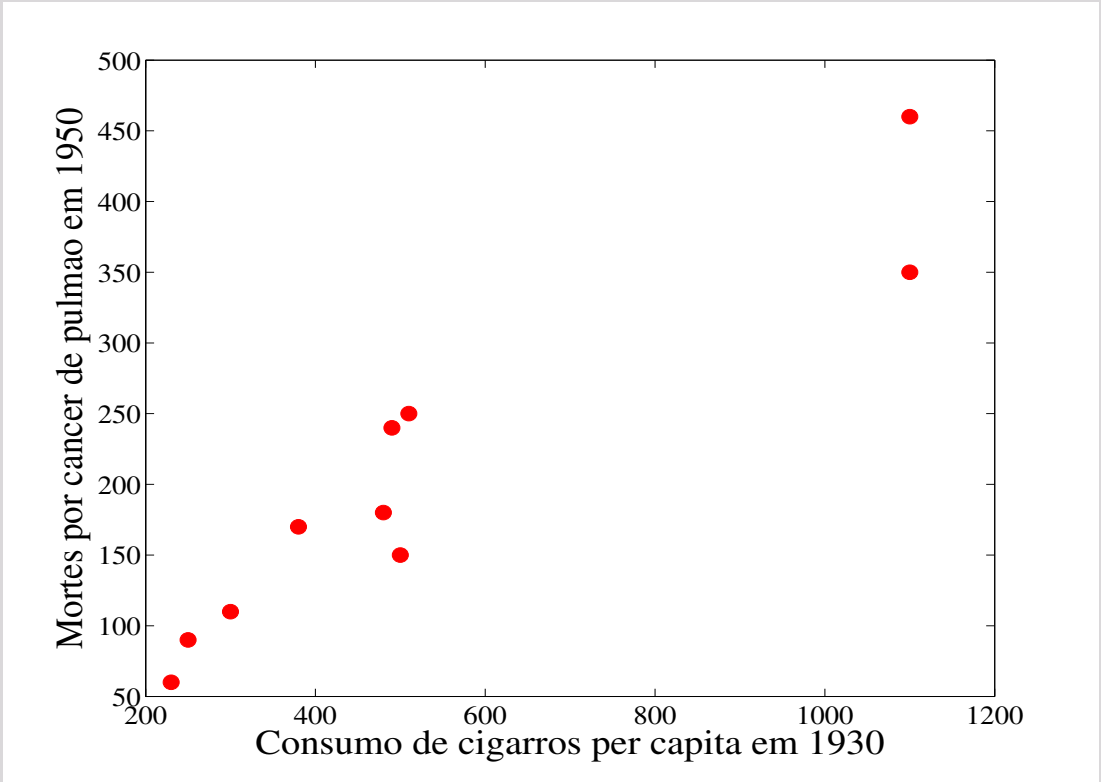


Figura: Scatterplot dos dados.

Conjunto de dados: Câncer de Pulmão

X_1	X_2	d_1	d_2	$d_1 \times d_2$
480	180	-54	-26	+1404
500	150	-34	-56	+1904
380	170	-154	-36	+5544
1100	350	+566	+144	+81504
1100	460	+566	+254	+143764
230	60	-304	-146	+44384
490	240	-44	+34	-1496
250	90	-284	-116	+32944
300	110	-234	-96	+22464
510	250	-24	+44	-1056
$\bar{x}_1 = 534$	$\bar{x}_2 = 206$	$--$	$--$	$\sum_{n=1}^{10} d_1(n) \times d_2(n) = 331.360$

$$\hat{\sigma}_{12} = \frac{\sum_{n=1}^{10} d_1(n) \times d_2(n)}{10-1} = 331.360/9 = 36.817{,}78$$

$$\hat{\sigma}_{12} = \frac{\sum_{n=1}^{10} d_1(n) \times d_2(n)}{10} = 331.360/10 = 33.136{,}00$$

Tabela: Tabela para cálculo passo-a-passo da covariância amostral.

Conjunto de dados: Câncer de Pulmão

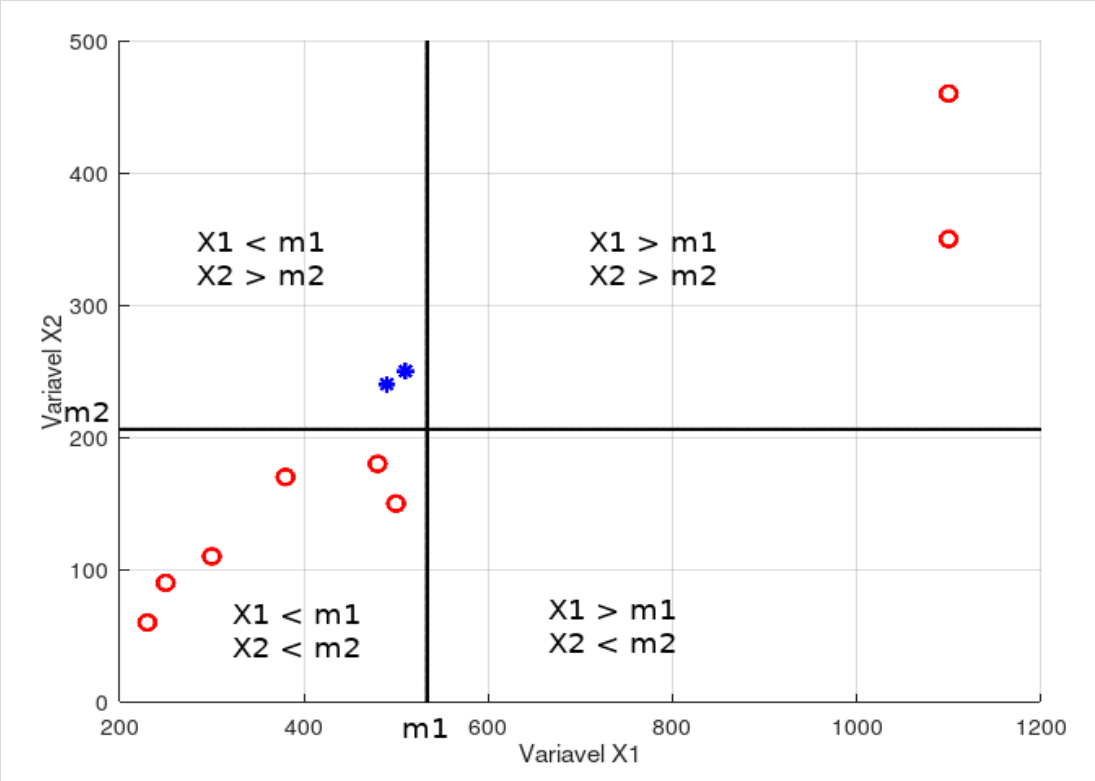


Figura: Scatterplot dos dados indicando posições dos pares (X_1, X_2) com desvios positivos (vermelho) e negativos (azuis) em relação às respectivas médias.

Implementação

- Dadas N observações conjuntas de (X_1, X_2), a covariância entre X_1 e X_2 e os desvios-padrão de X_1 e X_2 podem ser facilmente calculados em diferentes softwares.
  - No Matlab/Octave, usar os comandos `cov` e `std`.
  - No Excel[a], usar os comandos `COVAR` e `DESVPAD`.
  - No LibreOffice Calc, usar os comandos `COVAR` e `DESVPAD`.

[a] www.bertolo.pro.br/matematica/Disciplinas/3ano/Estatistica/Bimestre2/EstatisticaAplicada3.pdf

Definição

- O coeficiente de correlação entre duas variáveis aleatórias quaisquer é dado pela seguinte expressão:

$$\boxed{\hat{\rho}_{12} = \frac{\text{cov}(X_1, X_2)}{\text{dp}(X_1) \cdot \text{dp}(X_2)} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \cdot \hat{\sigma}_2}} \qquad (4)$$

em que

$$\begin{aligned} \hat{\sigma}_{12} &= \text{cov}(X1, X2) = \text{covariância amostral entre } X_1 \text{ e } X_2. \\ \hat{\sigma}_1 &= \text{dp}(X_1) = \text{desvio-padrão amostral de } X_1 = \sqrt{\hat{\sigma}_1} \\ \hat{\sigma}_2 &= \text{dp}(X_2) = \text{desvio-padrão amostral de } X_2 = \sqrt{\hat{\sigma}_2} \end{aligned}$$

- Note que: $-1 \leq \hat{\rho}_{12} \leq 1$.

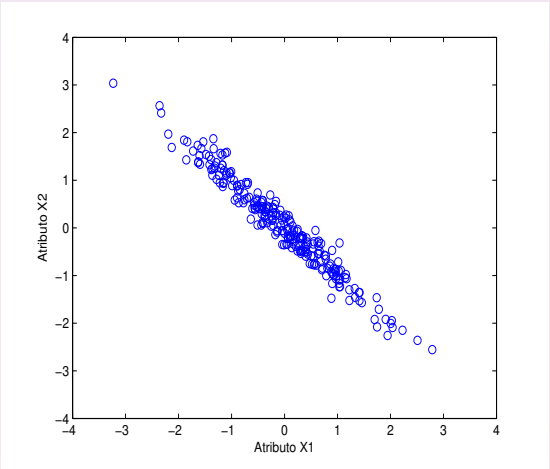
- Vamos analisar as situações em que o coeficiente de correlação assume seus valores máximo ($\hat{\rho}_{12} = +1$) e mínimo ($\hat{\rho}_{12} = -1$).
- Para $X_2 = X_1 = X$, temos $\bar{x}_2 = \bar{x}_1 = \bar{x}$ e $\hat{\sigma}_1 = \hat{\sigma}_2 = \hat{\sigma}$. Logo, chegamos a

$$\hat{\rho}_{12} = \frac{cov(X, X)}{\hat{\sigma} \cdot \hat{\sigma}} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2} = +1. \tag{5}$$

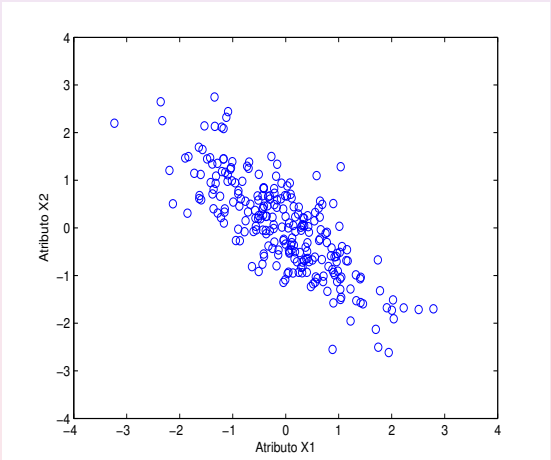
- Para $X_2 = -X_1 = X$, temos $\bar{x}_2 = -\bar{x}_1 = -\bar{x}$ e $\hat{\sigma}_1 = \hat{\sigma}_2 = \hat{\sigma}$. Logo, chegamos a

$$\hat{\rho}_{12} = \frac{cov(-X, X)}{\hat{\sigma} \cdot \hat{\sigma}} = \frac{-\hat{\sigma}^2}{\hat{\sigma}^2} = -1. \tag{6}$$

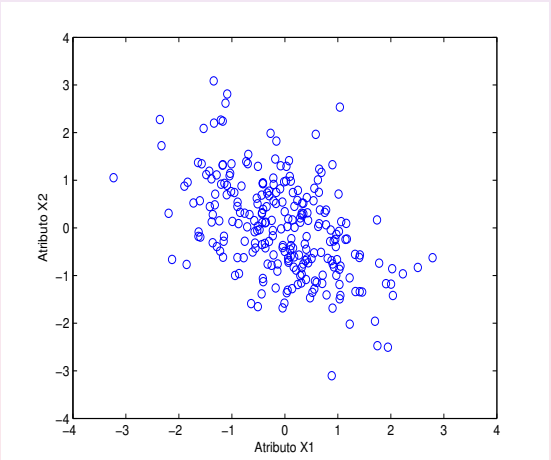
- Se $-1 < \hat{\rho}_{12} < 0$, então X_1 e X_2 estão **negativamente** correlacionadas.
- Quanto mais próximo $\hat{\rho}_{12}$ estiver de -1, maior é a correlação negativa entre X_1 e X_2.



$\hat{\rho}_{12} = -0.98$

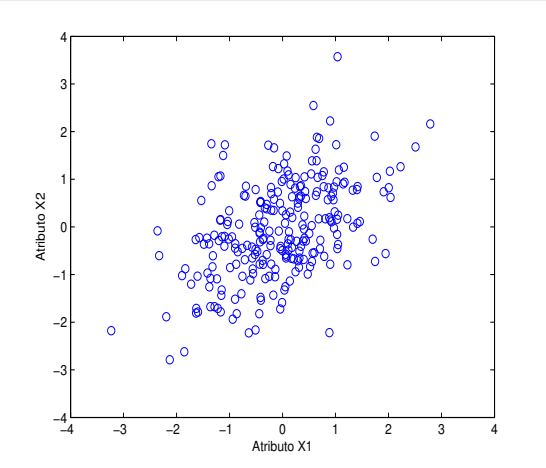


$\hat{\rho}_{12} = -0.8$

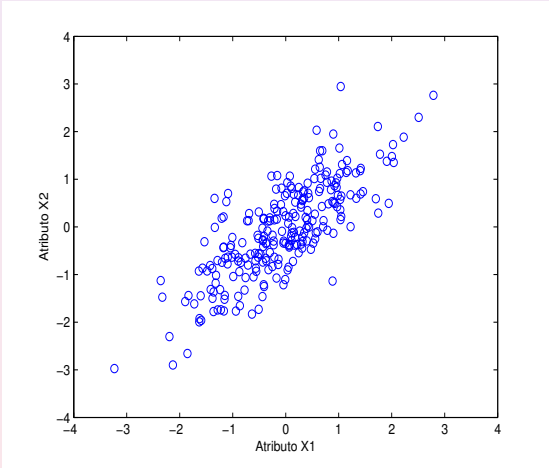


$\hat{\rho}_{12} = -0.5$

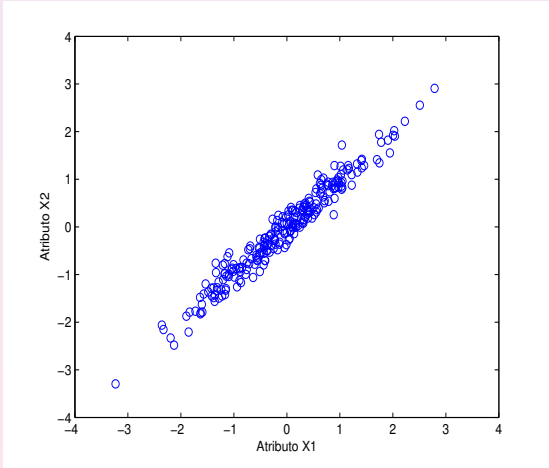
- Se $0 < \hat{\rho}_{12} \leq 1$, então X_1 e X_2 estão **positivamente** correlacionadas.
- Quanto mais próximo $\hat{\rho}_{12}$ estiver de 1, maior é a correlação positiva entre X_1 e X_2.



$\hat{\rho}_{12} = 0.5$



$\hat{\rho}_{12} = 0.8$



$\hat{\rho}_{12} = 0.98$

Implementação

- Dadas N observações conjuntas de (X_1, X_2), o coeficiente de correlação $(\hat{\rho}_{12})$ também pode ser calculado diretamente em vários pacotes de software.
  - No Matlab/Octave, usar os comandos `corrcoef` e `corr`.
  - No Excel[a], usar o comando `CORREL`.
  - No LibreOffice Calc, usar o comando `CORREL`.

[a] www.bertolo.pro.br/matematica/Disciplinas/3ano/Estatistica/Bimestre2/EstatisticaAplicada3.pdf

- Existe uma interessante relação teórica entre o coeficiente de correlação e a inclinação da reta de regressão linear (ou linha de tendência)

$$y = ax + b \qquad (\text{ou } x_2 = ax_1 + b), \tag{7}$$

ajustada aos dados pelo método dos mínimos quadrados.

- As constantes a e b são chamadas, respectivamente, de inclinação e intercepto da reta de tendência, podem ser estimadas como

$$\boxed{\hat{a} = \left(\frac{\hat{\sigma}_2}{\hat{\sigma}_1}\right) \cdot \hat{\rho}_{12} = \left(\frac{\hat{\sigma}_2}{\hat{\sigma}_1}\right) \cdot \left(\frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2}\right) = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1^2}} \tag{8}$$

e

$$\boxed{\hat{b} = \bar{x}_2 - \hat{a}\bar{x}_1}. \tag{9}$$

- Note que no slide anterior escrevemos a inclinação da reta de regressão como sendo diretamente proporcional ao coeficiente de correlação das variáveis X_1 e X_2, ou seja

$$\boxed{\hat{a} = k \cdot \hat{\rho}_{12}}, \tag{10}$$

em que a constante de proporcionalidade é dada pela razão entre os desvios-padrão de X_2 e X_1:

$$\boxed{k = \frac{\hat{\sigma}_2}{\hat{\sigma}_1}}. \tag{11}$$

- Se os desvios-padrão forem o mesmo para as duas variáveis, temos que a inclinação será igual ao coeficiente de correlação: $\hat{a} = \hat{\rho}_{12}$, se $\hat{\sigma}_1 = \hat{\sigma}_2$.

- O gráfico de dispersão pode indicar a adequação de um modelo de regressão linear a um conjunto de medidas de (X, Y).
- Se este for o caso, devemos estimar os parâmetros a e b da reta de regressão $\hat{y}_i = \hat{a}x_i + \hat{b}$.
- Em seguida, calculamos os resíduos, ou seja, os valores preditos da variável de saída para os dados de entrada observados:

$$e_i = y_i - \hat{y}_i, \tag{12}$$

em que y_i é o i-ésimo valor observado e $\hat{y}_i$ é o valor predito pelo modelo de regressão.

- Espera-se para os resíduos uma distribuição aproximadamente simétrica e próxima da gaussiana.

Observações sobre Análise dos Resíduos

- O histograma dos resíduos deve ser semelhante ao esperado para dados com uma distribuição gaussiana. No Octave/Matlab, pode-se usar o comando `histfit()` para visualizar a similaridade com a distribuição gaussiana.
- Alguns autores recomendam que observaçoes atípicas (*outliers*) sejam descartados.
- Outros autores acham que *outliers* fornecem informação importante sobre circunstâncias não usuais (e.g. falhas), de interesse para o experimentador, e não devem ser descartados.

Passos Básicos da Análise de Resíduos

(1) Construir um histograma de freqüência dos resíduos.

(2) Normalizar os resíduos, calculando-se

$$d_i = \frac{e_i}{\hat{\sigma}_e}, \quad i = 1, \ldots, n$$

(3) Se os resíduos normalizados d_i forem $N(0, 1,$ então aproximadamente 99% dos seus valores devem estar no intervalo $(-3, +3)$.

(4) Resíduos muito fora do intervalo $(-3, +3)$ podem indicar a presença de um *outlier*, isto, é uma observação atípica em relação ao resto dos dados.

Definição - Coeficiente de Determinação

- O coeficiente de determinação R^2 é usado para avaliar modelos de regressão.

$$R^2 = 1 - \frac{SEQ}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{13}$$

  tal que que $0 \leq R^2 \leq 1$.

- SEQ é a soma dos erros quadráticos produzida pelo modelo de regressão linear $\hat{y}_i = \hat{a}x_i + \hat{b}$.
- S_{yy} é a soma dos erros quadráticos produzida quando o modelo de regressão é a média da variável de saída, ou seja, $\hat{y}_i = \bar{y}$.
- Quanto mais próximo R^2 está de 1, melhor é o modelo.

Interpretando o Coeficiente de Determinação

- O índice R^2 fornece uma medida do quanto da variabilidade da variável dependente y, conforme observada na amostra coletada, está sendo explicada (ou capturada) pelo modelo de regressão escolhido; no caso, o modelo linear.
- Por exemplo, suponha que obtemos $R^2 = 0{,}87$. Este valor indica que 87% da variabilidade da variável dependente, presente no conjunto de medidas coletado, foi capturada adequadamente, enquanto 13% da variabilidade remanescente ainda carece de explicação.
- Pode-se mostrar que, para o modelo de regressão linear, o coeficiente de determinação R^2 é igaul ao quadrado do coeficiente de correlação ρ_{12}. Fica como exercício.

Implementação

- Dadas N observações conjuntas de (X_1, X_2), a inclinação da linha de tendência ajustada a este conjunto de observações pode ser facilmente calculada em planilhas numéricas, bem como o valor de R^2 correspondente.
  - No Excel e no LibreOffice Calc, usar o comando `INCLINAÇÃO`.
  - No Excel e no LibreOffice Calc, usar o comando `RQUAD`.