Covariância, Correlação e Regressão

Covariância

- A covariância mede o grau de relação entre duas variáveis;

DEFINIÇÃO:

Dados n pares de valores $(x_1, y_1), \dots, (x_n, y_n)$ define-se variância entre as duas variáveis X e Y como:

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

ou seja, é a média dos valores centrados das variáveis.

Covariância

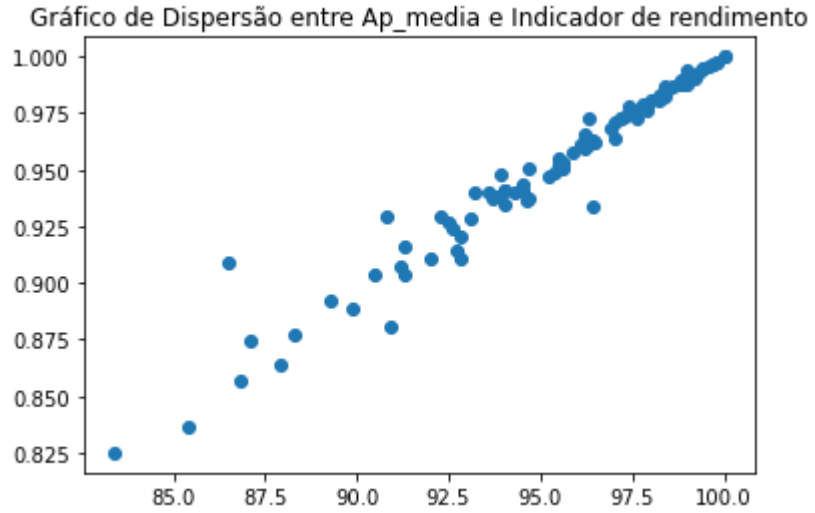
IMPORTANTE:

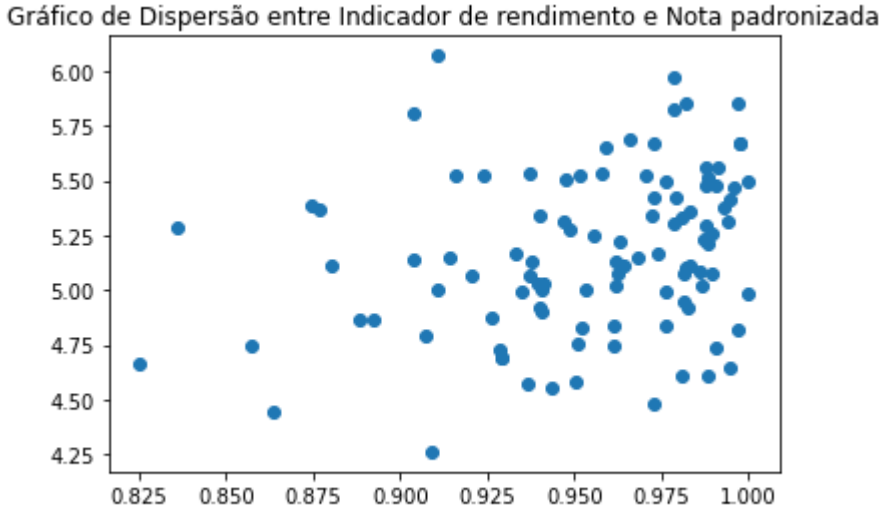
1. A covariância não tem escala bem definida;
2. Caso as variáveis sejam independentes a covariância é **nula**;
3. A **covariância positiva** significa que as variáveis tem o mesmo comportamento (de crescimento ou decrescimento)
4. A **covariância negativa** indica que as variáveis tem comportamentos distintos (se uma aumenta a outra diminui ou vice-versa)

Correlação

- Existem situações em que há o interesse/necessidade em estudar o comportamento conjunto de uma ou mais variáveis;

- Uma boa ferramenta para verificar este comportamento são os gráficos de dispersão.





Correlação

- Para desenhar um diagrama de dispersão, que é relativamente simples, basta indicar a variável do eixo das abscissas (eixo x) e a outra variável do eixo das ordenadas (eixo y)

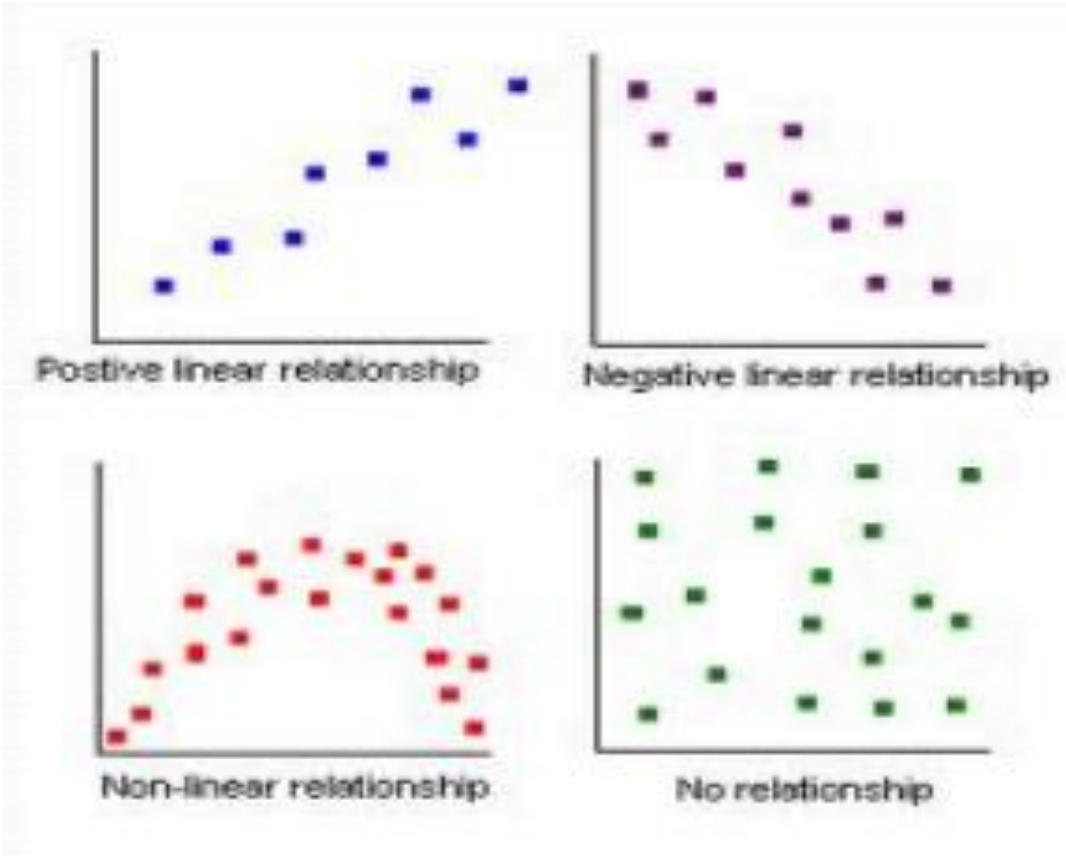
- Os pontos serão plotados de acordo com o par de valores correspondente

Correlação

- **Correlação** resume o grau de relacionamento entre duas variáveis

- Caso os pontos das variáveis, representados no gráfico de dispersão estejam distribuídos como em uma reta imaginária, dizemos que os dados apresentam **correlação linear**

- Entretanto, outros modelos de correlação podem existir como as polinomiais e as exponenciais.



Coeficiente de Correlação de Pearson

- O **Coeficiente de correlação de Pearson** é muito utilizado pela sua simplicidade

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

- O valor de r sempre estará em uma escala de -1 a 1.

idp

Outros coeficientes de correlação

- **Coeficiente de Correlação de Matthews**. Fornece uma medida de precisão entre as variáveis. Muito utilizada em Data Mining para construção da matriz de confusão;

- **Coeficiente de Correlação de Kendall.** Verifica a semelhança de dados ordinais. Muito utilizado em psicologia e aprendizado de máquina.

- **Coeficiente de Correlação de Wilcoxon.** Utilizado quando não há certeza sobre a normalidade das variáveis (estatística não-paramétrica)

- **Coeficiente de correlação de Spearman**. Utiliza-se da ordem das observações e não seus valores. Pode ser utilizado para qualquer tipo de relação entre variáveis, não somente a linear.

idp

Regressão

- A análise de regressão, parte do pressuposto que existe correlação entre as variáveis;

- Dito isso, é possível encontrar uma relação matemática (equação) entre as variáveis;

- A mais simples é a equação de uma reta do tipo

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

onde ε_i é o erro aleatório para i-ésima observação.

idp