UNIVERSIDADE FEDERAL DA PARAÍBA

Correlação e Regressão

Luiz Medeiros de Araujo Lima Filho
Departamento de Estatística

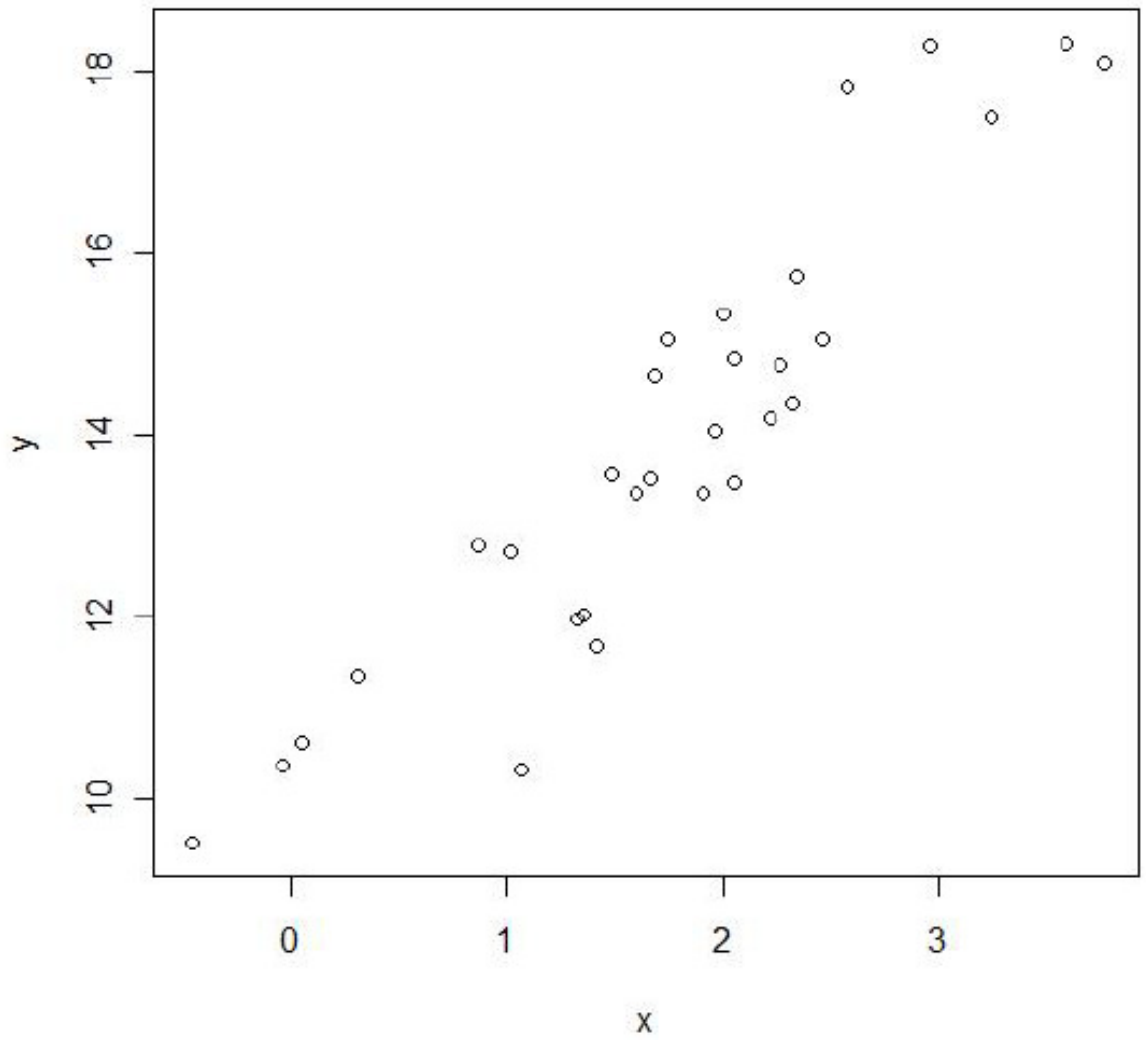
Introdução

• Existem situações nas quais há interesse em estudar o comportamento conjunto de uma ou mais variáveis;

• Em muitos casos, a explicação de um fenômeno de interesse pode estar associado a outros fatores (variáveis) que contribuem de algum modo para a ocorrência deste fenômeno.

• O comportamento conjunto de duas variáveis quantitativas pode ser observado por meio do gráfico de dispersão.

Introdução



Exemplo

Quadro 1: Comprimento (em *cm*) e peso (em *kg*) de cães

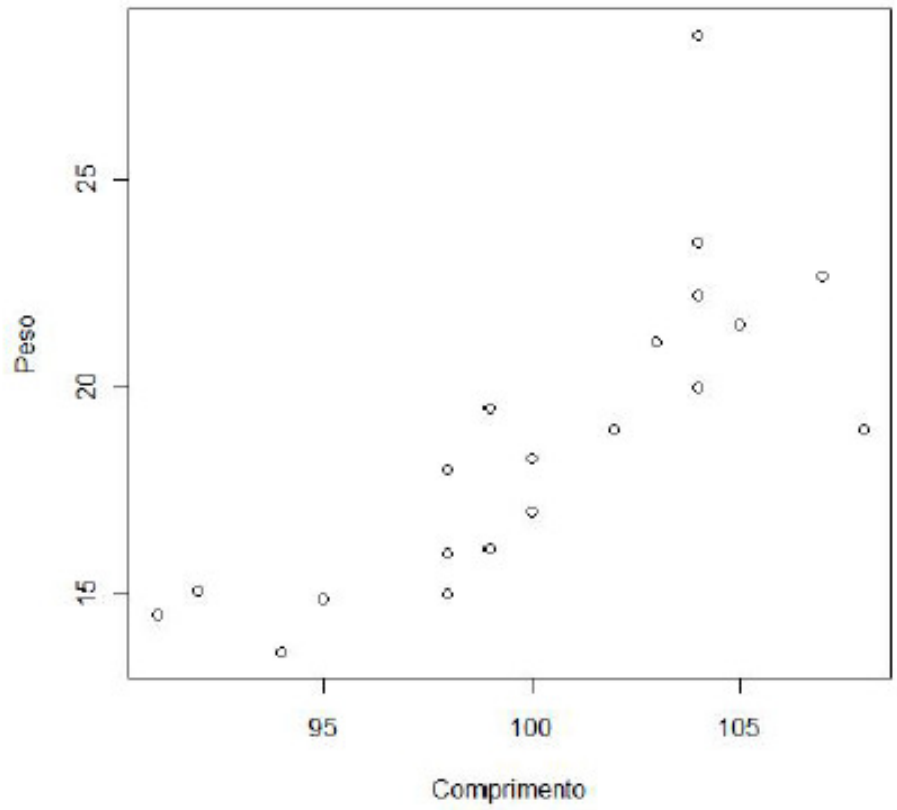
Nº	Comprimento	Peso	Nº	Comprimento	Peso
1	104	23.5	11	98	15.0
2	107	22.7	12	95	14.9
3	103	21.1	13	92	15.1
4	105	21.5	14	104	22.2
5	100	17.0	15	94	13.6
6	104	28.5	16	99	16.1
7	108	19.0	17	98	18.0
8	91	14.5	18	98	16.0
9	102	19.0	19	104	20.0
10	99	19.5	20	100	18.3

• Para desenhar um diagrama de dispersão, é necessário sempre fazer o eixo cartesiano para identificar os pontos das variáveis quantitativas consideradas.

• Representa-se primeiramente uma das variáveis no eixo das abscissas (variável X) e a outra variável no eixo das ordenadas (variável Y).

• Os valores das variáveis são marcados sob os respectivos eixos e assim marca-se um ponto para cada par de valores.

Exemplo



Correlação e Regressão

• São duas técnicas estreitamente relacionadas, que visa estimar uma relação que possa existir entre duas variáveis na população.

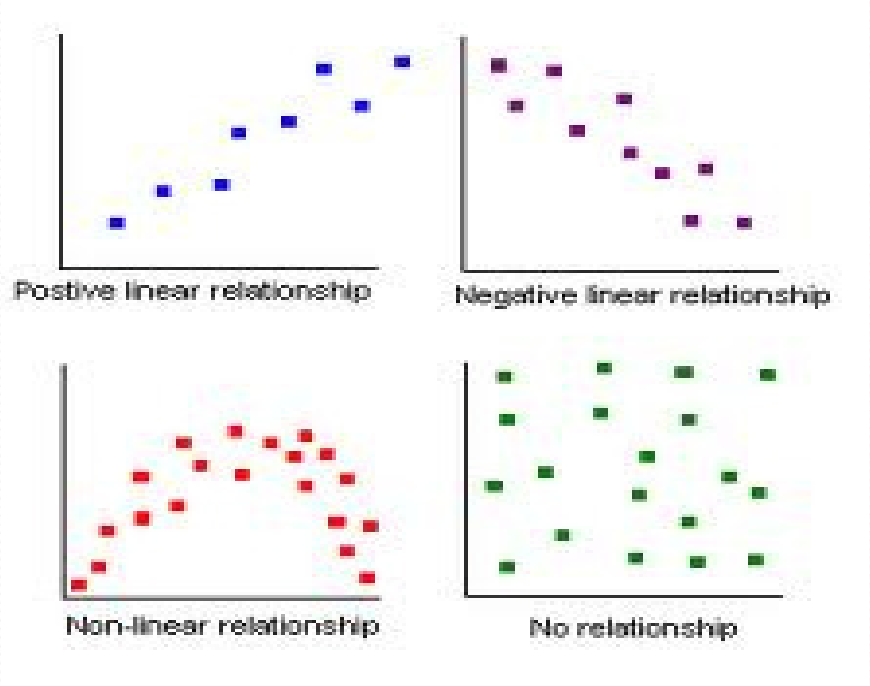
Correlação: resume o grau de relacionamento entre duas variáveis (X e Y, por exemplo).

Regressão: tem como resultado uma equação matemática que descreve o relacionamento entre variáveis.

Correlação

• O objetivo do estudo da correlação é determinar (mensurar) o grau de relacionamento entre duas variáveis.

• Caso os pontos das variáveis, representados num plano cartesiano (X, Y) ou gráfico de dispersão, apresentem uma dispersão ao longo de uma reta imaginária, dizemos que os dados apresentam uma correlação linear.



Coeficiente de correlação linear de Pearson

• Uma medida do grau e do sinal da correlação linear entre duas variáveis (X,Y) é dado pelo **Coeficiente de Correlação Linear de Pearson**, definido por:

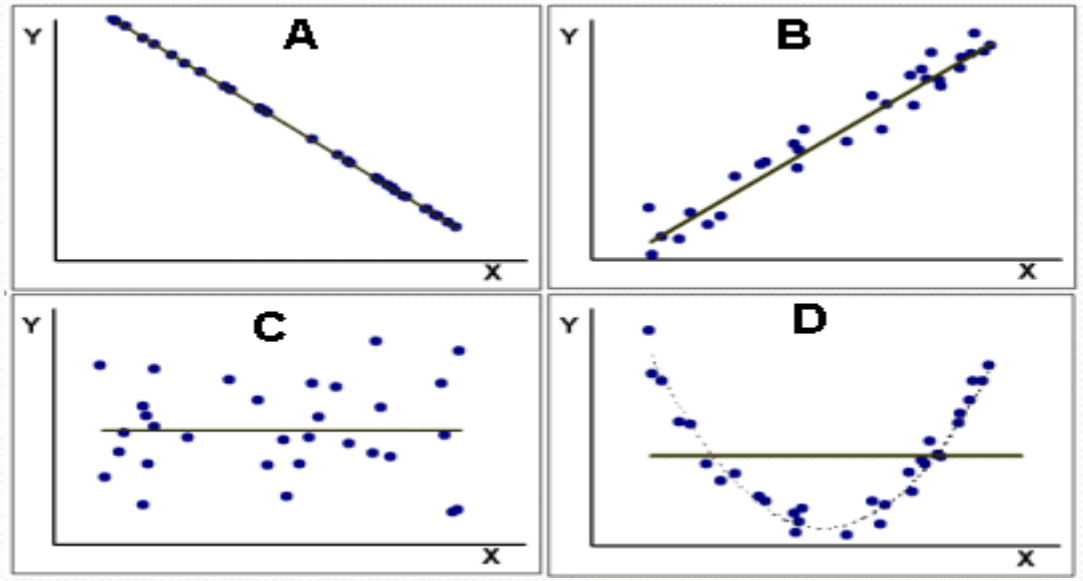
$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

• O valor de "**r**" estará sempre no intervalo de -1 a 1.

Propriedades do Coeficiente de Correlação Linear

• Este coeficiente é adimensional, logo não é afetado pelas unidades de medidas das variáveis X e Y.

• O sinal **positivo** indica que as variáveis são **diretamente proporcionais**, enquanto que o sinal **negativo** indica que a relação entre as variáveis é **inversamente proporcional**.

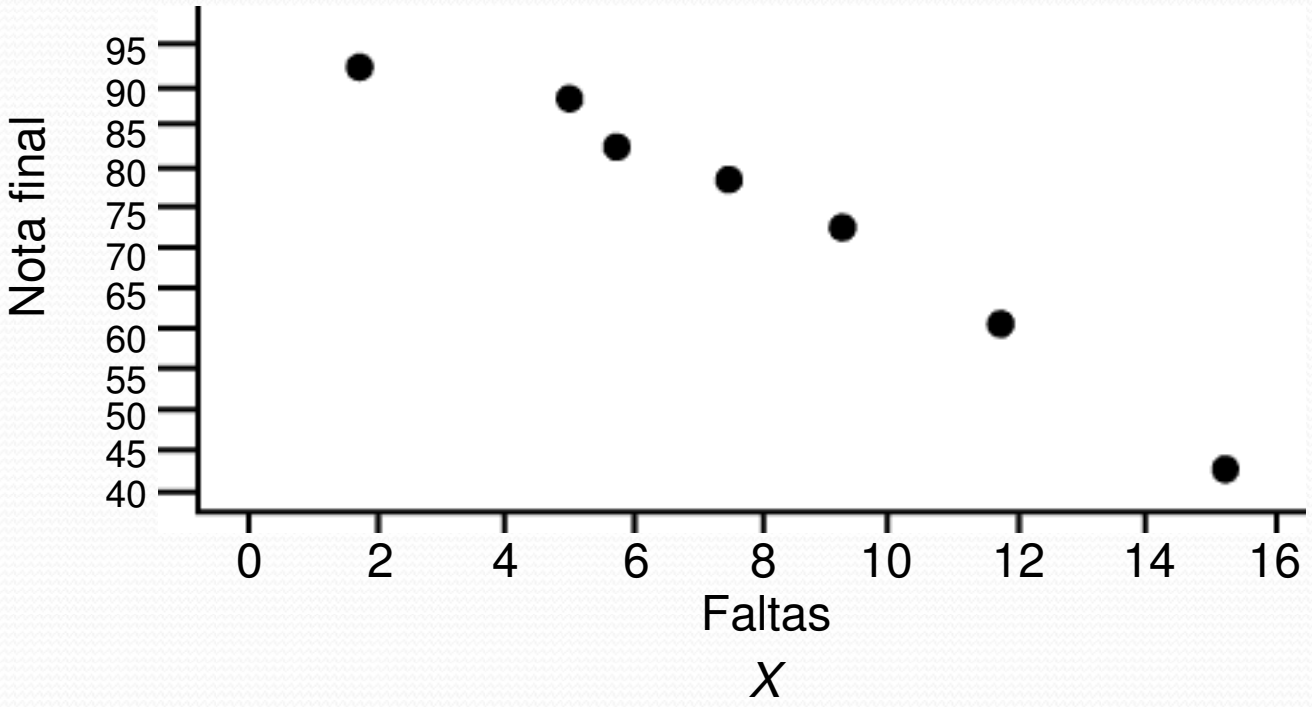


Exemplo 1:

A tabela abaixo apresenta os preços médios das ações e títulos divulgados pela Bolsa de Nova York entre 1950 e 1959. Calcule o coeficiente de correlação de Pearson e interprete o resultado.

Ano	Ações (X)	Títulos (Y)
1950	35,22	102,43
1951	39,87	100,43
1952	41,85	97,43
1953	43,23	97,81
1954	40,06	98,32
1955	53,29	100,07
1956	54,14	97,08
1957	49,12	91,59
1958	40,71	94,85
1959	55,15	94,65
Total (∑)	452,64	974,66

Exemplo 2: Existe correlação entre o número de faltas e a nota final? De que forma?



Faltas x	Nota final y
8	78
2	92
5	90
12	58
15	43
9	74
6	81

REGRESSÃO

• Quando analisamos dados que sugerem a existência de uma relação funcional entre duas variáveis, surge então o problema de se determinar uma função matemática que exprima esse relacionamento, ou seja, uma equação de regressão.

• Ao imaginar uma relação funcional entre duas variáveis, digamos X e Y, estamos interessados numa função que explique grande parte da variação de Y por X. Entretanto, uma parcela da variabilidade de Y não explicada por X será atribuída ao acaso, ou seja, ao erro aleatório.

• Quando se estuda a variação de uma variável Y em função de uma variável X, dizemos que Y é a variável dependente e que X é a variável explanatória (ou independente).

• O modelo em que busca explicar uma **variável Y como uma função linear** de apenas uma **variável X** é denominado de modelo de **regressão linear simples**.

Variável independente, X	Variável dependente, Y
Temperatura do forno (°C)	Resistência mecânica da cerâmica (MPa)
Quantidade de aditivo (%)	Octanagam da gasolina
Renda(R$)	Consumo(R$)
Memória RAM do computador (Gb)	Tempo de resposta do sistema (s)
Área construída do imóvel (m^2)	Preço do imóvel (R$)

X

REGRESSÃO LINEAR SIMPLES

Formalmente, a análise de regressão parte de um conjunto de observações pareadas (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), relativas às variáveis X e Y e considera que podemos escrever a relação entre as duas variáveis, da seguinte maneira:

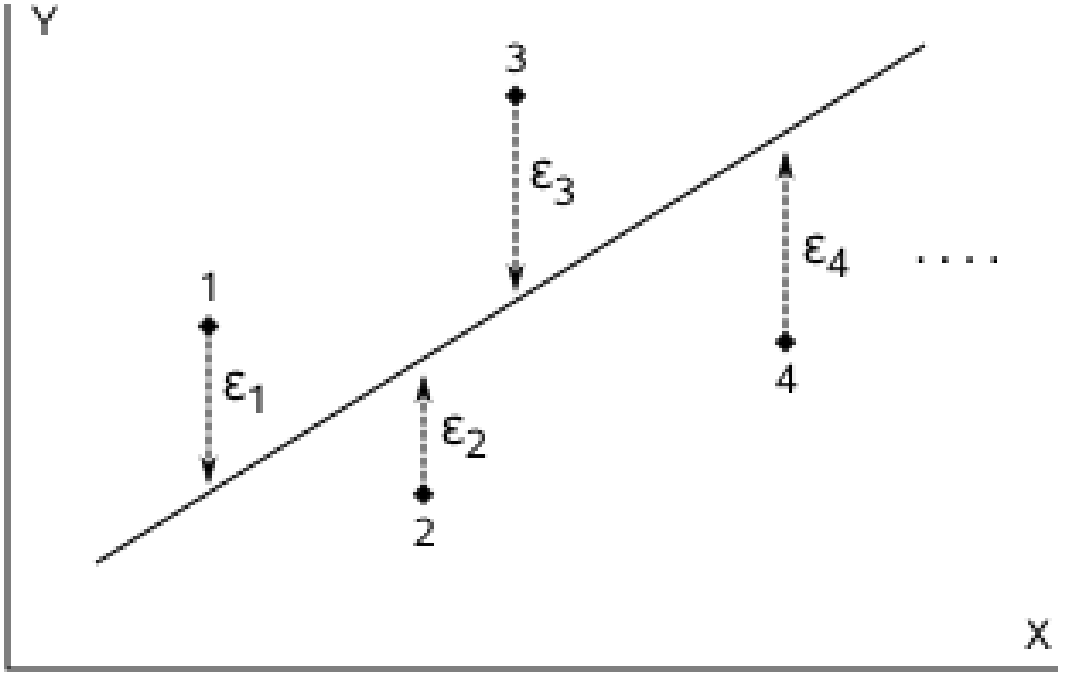
$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

onde:

- y_i **é a variável resposta** associada à i-ésima observação de Y;
- x_i **é a i-ésima observação** do valor fixado para a **variável independente** (e não aleatória) X;
- ε_i **é o erro aleatório** para a i-ésima observação, isto é, o efeito de fatores que estão afetando a observação de Y de forma aleatória. Por suposição, consideramos que $\varepsilon_i \sim N(0,\sigma^2)$;
- α e β são parâmetros que precisam ser estimados.

ESTIMAÇÃO DOS PARÂMETROS

O objetivo é estimar valores para α e β através dos dados fornecidos pela amostra. Além disso, encontrar a reta que passe o **mais próximo possível dos pontos observados segundo um** critério pré-estabelecido.



MÉTODO DOS MÍNIMOS QUADRADOS

É usado para estimar os parâmetros do modelo (α e β) e consiste em fazer com que a soma dos erros quadráticos seja menor possível, ou seja, este método consiste em obter os valores de α e β que minimizam a expressão:

$$S = \sum \varepsilon_i = \sum (Y_i - \alpha - \beta x_i)^2$$

Aplicando-se derivadas parciais à expressão acima, e igualando-se a zero, acharemos as **estimativas para α e β.**

MÉTODO DOS MÍNIMOS QUADRADOS

Após aplicar as derivadas parciais, e igualando-se a zero, é possível obter as seguintes **estimativas para** α **e** β**,** as quais **chamaremos de *a* e *b*, respectivamente**:

$$a = \frac{\sum y_i - b\sum x_i}{n}$$

e

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

A chamada *equação (reta) de regressão* é dada por

$$\hat{y} = a + bx$$

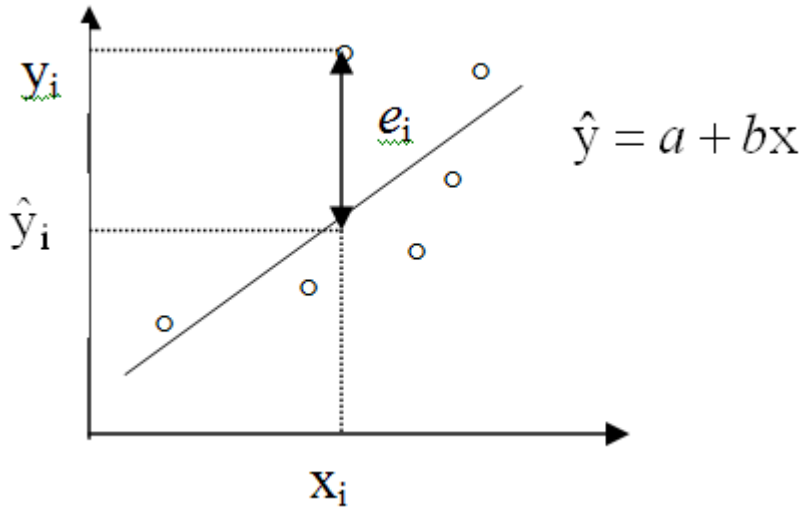
e para cada valor x_i (i = 1, ..., n) temos, pela equação de regressão, o *valor predito*:

$$\hat{y}_i = a + b x_i$$

A diferença entre os valores observados e os preditos será chamada de *resíduo do modelo de regressão,* sendo denotado por:

$$e_i = y_i - \hat{y}_i$$

O resíduo relativo à i-ésima observação (e_i) *pode ser considerado uma estimativa do erro aleatório* (e_i), *como* ilustrado abaixo.



COEFICIENTE DE DETERMINAÇÃO (R^2)

O coeficiente de determinação é uma **medida** descritiva da proporção da **variação de Y que pode ser explicada por variações em X**, segundo o modelo de regressão especificado. Ele é dado pela seguinte razão:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{variação explicada pelo modelo}}{\text{variação total}}$$

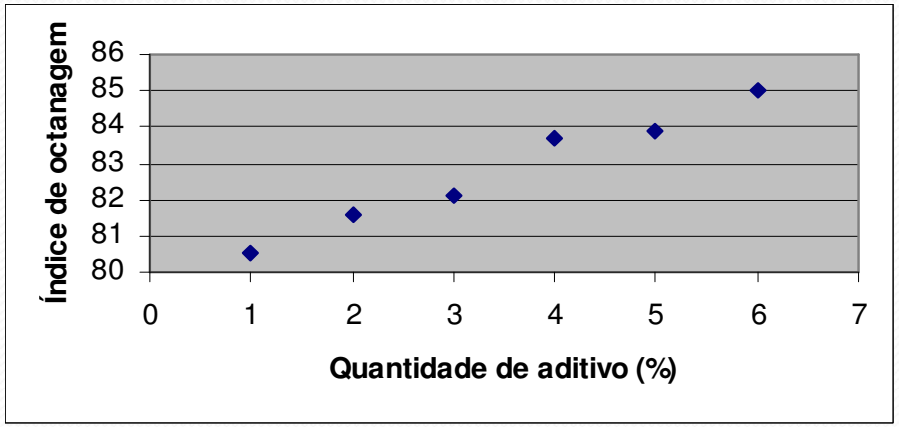
- Quanto mais próximo de 1 estiver o coeficiente de determinação, melhor será o grau de explicação da variação de Y em termos da variável X.

- É uma medida sempre positiva, e é obtida, na regressão linear simples, elevando-se o coeficiente de correlação de pearson ao quadrado.

EXEMPLO 3:

Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um novo aditivo (X). Para isso, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados são mostrados no gráfico de dispersão.

X	Y
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



a) Existe uma relação linear entre a adição de um novo aditivo e a octanagem da gasolina? Qual o grau dessa relação?
b) Determine a reta de regressão que explica a octanagem da gasolina em função da adição do novo aditivo. Calcule o coeficiente de determinação do modelo.
c) Se adicionarmos 5,5% de aditivo, qual o índice de octanagem esperado?
d) Calcule o erro de estimação para cada valor de X.

EXEMPLO 4:

Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo decorrido após sua administração.

Tempo (minutos)	Quantidade hidrolisada
2	3.5
3	5.7
5	9.9
8	16.3
10	19.3
12	25.7
14	28.2
15	32.6

Fonte: AVENS e FOLDES(1951)

a) Existe uma relação linear entre a quantidade de procaína e o tempo decorrido após sua administração? Qual o grau dessa relação?

b) Determine a reta de regressão que explica a quantidade de procaína em função do tempo. Calcule o coeficiente de determinação do modelo.

c) Qual a quantidade de procaína hidrolisada após 6 minutos de sua administração? E após 13 minutos?

d) Calcule o erro de estimação para cada valor de X.

Exemplo 5:

A tabela abaixo apresenta os preços médios das ações e títulos divulgados pela Bolsa de Nova York entre 1950 e 1959. Calcule o coeficiente de correlação de Pearson e interprete o resultado.

Ano	Ações (X)	Títulos (Y)
1950	35,22	102,43
1951	39,87	100,43
1952	41,85	97,43
1953	43,23	97,81
1954	40,06	98,32
1955	53,29	100,07
1956	54,14	97,08
1957	49,12	91,59
1958	40,71	94,85
1959	55,15	94,65
Total (∑)	452,64	974,66

a) Determine a reta de regressão que explique os títulos divulgados em função do preço médio das ações. Calcule o coeficiente de determinação do modelo.

b) Qual o número de títulos divulgados para um preço médio da ação de 45,00? E para um preço médio de 50,00?

c) Calcule o erro de estimação para cada valor de X.

REGRESSÃO LINEAR MÚLTIPLA

Em algumas situações o interesse é estudar o comportamento de uma variável dependente Y em função de duas ou mais variáveis independentes X_i.

Variável Dependente (Y)	Variáveis Independentes($X_1,\ldots,X_k$)
Y= Preço de um imóvel(R$)	X_1=Área do imóvel(m2)
	X_2=Custo do m2(R$)
	X_3=Localização
Y= Tempo de resposta de um sistema computacional(seg)	X_1=Memória RAM(Gb)
	X_2=Sistema Operacional
	X_3=Tipo de processador
Y= Valor de revenda de um carro seminovo(R$)	X_1=Valor modelo novo (R$)
	X_2=Kilometragem
	X_3=Idade do veículo(anos)
	X_4=Estado de conservação
	X_5=Opcionais

Os métodos para tratar com o problema de prever uma variável por meio de diversas outras são semelhantes àqueles para uma variável independente.