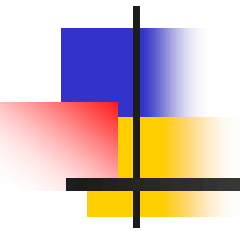


Regressão Linear Simples e Múltipla

Modelos de Regressão Linear

- Técnica Estatística que permite obter uma equação que explique satisfatoriamente a relação funcional entre uma variável resposta e uma ou mais variáveis explicativas.
- Classificada como um método de dependência, pois uma variável depende de uma ou mais variáveis (chamadas independentes).
- São usados para Predição de uma variável de interesse.
- Usadas em Computação, Administração, Engenharias, Biologia, Agronomia, Saúde, Sociologia, etc.

Exemplos

- Vendas de um Produto X Investimento em Propaganda
- Gastos da Família X (Renda Familiar ou Nro Pessoas no Domicílio)
- Demanda de um Produto X Preço
- Renda Per Capita X Crescimento Populacional
- Renda X Idade
- População de Ratos X População de Cobras
- Peso X (Alimentação e Atividade Física)
- Altura dos Pais X Altura dos Filhos

Relação Funcional entre as Variáveis

- Variável Dependente

  - Será expressa em função de uma ou mais variáveis independentes.
    - Exemplo: (1) Renda Futura
      (2) Inadimplência
      (3) Valor da Ação
  - É possível projetar os seus valores futuros.

- Variável(is) Independente(s), ou Explicativa(s), ou Regressora(s)

  - Utilizadas para Compreensão do Comportamento da Variável Dependente
    - Exemplo: (1)
      (2)
      (3)

Relação Funcional entre as Variáveis

Relação de Causa e Efeito

- Variável(is) Independente(s): Causa
  - Quanto mais variáveis, mas se reduz o erro aleatório (os resíduos)
  - Objetivo de Maior Explicação com Menor Número de Variáveis

- Variável Dependente: Efeito

Coeficiente de Correlação de Pearson

- Medida do Grau de Relacionamento entre 2 variáveis
- Escala das Variáveis: Intervalar ou Razão
- Natureza da Relação entre as Variáveis: Sinal e Magnitude

Coeficiente de Correlação de Pearson

Sejam X e Y duas variáveis aleatórias

$$r = \frac{n.\sum X_i.Y_i - \left(\sum X_i\right).\left(\sum Y_i\right)}{\sqrt{\left[n.\sum X_i^2 - \left(\sum X_i\right)^2\right].\left[n.\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}}$$

Características de r

- Número Adimensional: Independe das unidades de medida de X e Y
- Intervalo de variação de r : [-1, 1]
- $r_{(X, X)} = r_{(Y, Y)} = 1$
- $r_{(X, Y)} = r_{(Y, X)}$

Coeficiente de Correlação de Pearson

- Sinal
  - Positivo: Crescimento no Mesmo Sentido
  - Negativo: Crescimento em Sentidos Opostos
  - Nulo: Não há relação entre as variáveis

- Exemplos
  - ALTA CORRELAÇÃO POSITIVA
    - comissão de um vendedor x vendas
  - ALTA CORRELAÇÃO NEGATIVA
    - demanda x preço de um produto
  - BAIXA CORRELAÇÃO POSITIVA
    - nº médicos x nº habitantes de um local
  - BAIXA CORRELAÇÃO NEGATIVA
    - demanda de uma marca x nº marcas concorrentes
  - CORRELAÇÃO NULA
    - estatura x graus de miopia

Coeficiente de Correlação de Pearson

Magnitude dos coeficientes (Hair Jr.; Babin, Money e Samouel, 2006)

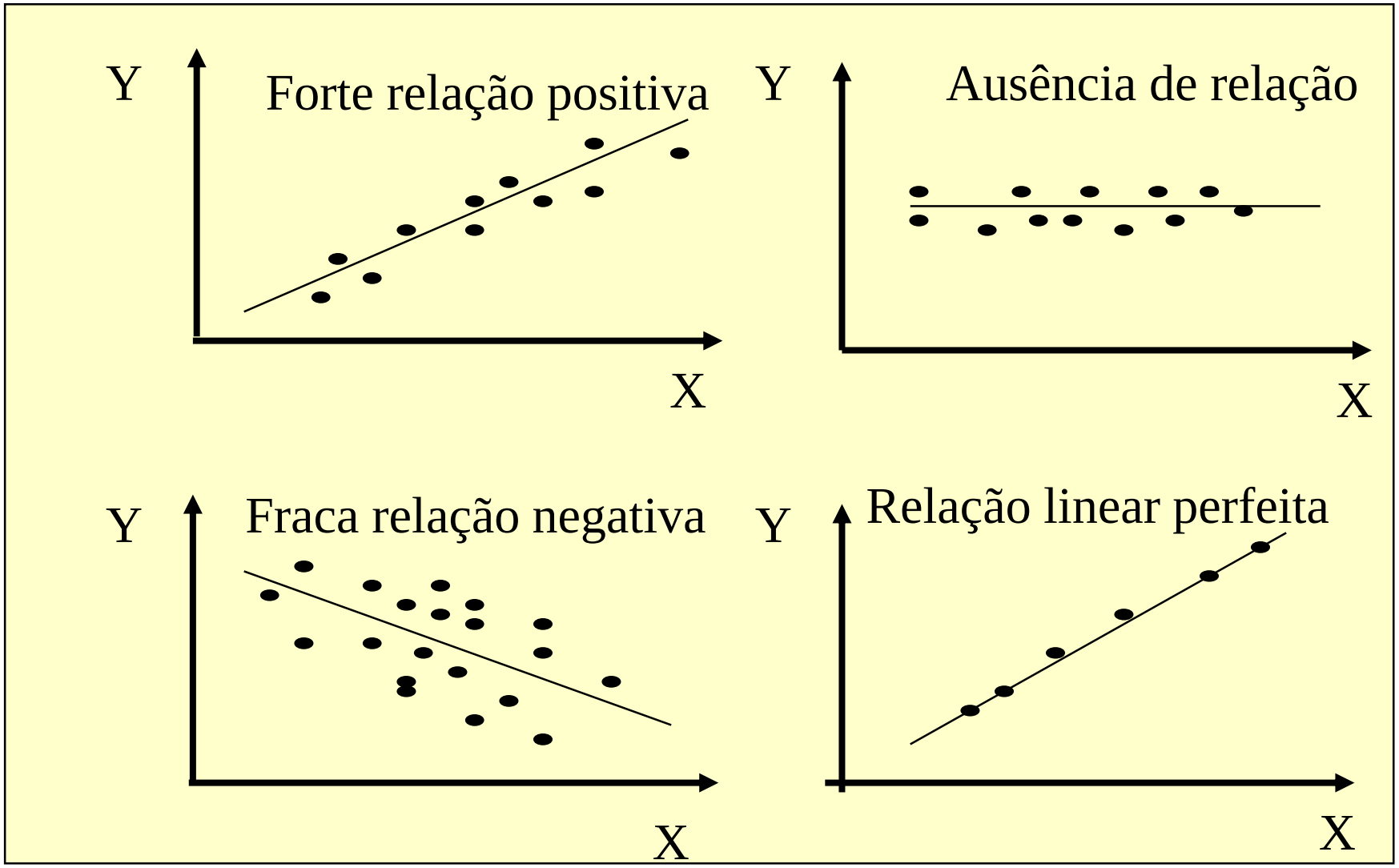
Leve, quase imperceptível	→ 0,01 a 0,20
Pequena, mas definida	→ 0,21 a 0,40
Moderada	→ 0,41 a 0,70
Alta	→ 0,71 a 0,90
Muito forte	→ 0,91 a 1,00

Coeficiente de Correlação de Pearson

- DIAGRAMA DE DISPERSÃO: Representação gráfica para observar se a correlação é:

  - Positiva, Negativa ou Nula
  - Linear, Não Linear
  - De grau alto, moderado, baixo

Coeficiente de Correlação de Pearson



Valores de r igual ou próximos de 1 ou –1 indica que exige uma forte relação entre as variáveis: no primeiro caso a relação é direta, enquanto que no segundo a relação é inversa. Valores próximos de Zero, significa que existe pouco relacionamento entre as variáveis.

Cuidados na Interpretação do Coeficiente de Correlação

- Correlação não implica Relação de Causa de Efeito

  - Exemplo1: Alto Número de Consumo de Refrigerantes associados ao Alto Número de Casos de Internações por Desidratação.
    Questão: Consumo de Refrigerantes causa desidratação?

Cuidados na Interpretação do Coeficiente de Correlação

- Correlação não implica Relação de Causa de Efeito

  - Exemplo1: Alto Número de Consumo de Refrigerantes associados ao Alto Número de Casos de Internações por Desidratação.
    Questão: Consumo de Refrigerantes causa desidratação?

  - Exemplo2: A correlação entre o número de filmes feitos por Nicolas Cage em um ano e a quantidade de gente que morre em acidentes de helicóptero nos EUA foi calculada e indicou uma correlação de -0,82
  - Questão: Nicolas Cage deve fazer mais filmes para evitar acidentes de Helicópero?

Cuidados na Interpretação do Coeficiente de Correlação

- Então, por quê existe correlação ?
- Há três possíveis explicações:
  - Existe, de fato, relação de causa e efeito
  - Ambas as variáveis estão relacionadas com uma terceira
  - A correlação deve-se ao acaso

Exemplo

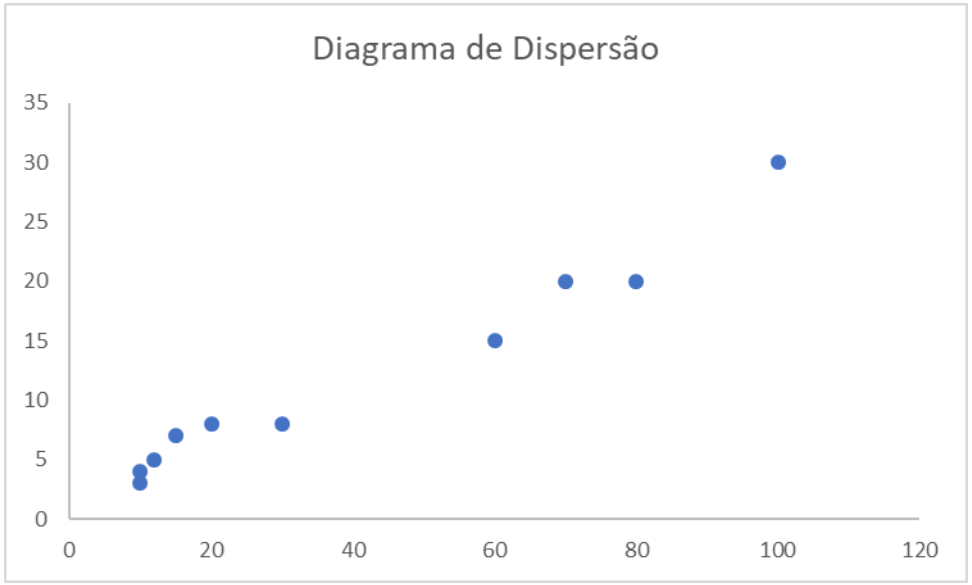
Dados sobre a renda e poupança (em milhares de reais) de 10 clientes bancários.

Renda X	Poupança Y
10	4
15	7
12	5
70	20
80	20
100	30
20	8
30	8
10	3
60	15

Exemplo

Dados sobre a renda e poupança (em milhares de reais) de 10 clientes bancários.

Renda X	Poupança Y
10	4
15	7
12	5
70	20
80	20
100	30
20	8
30	8
10	3
60	15



Exemplo

Dados sobre a renda e poupança (em milhares de reais) de 10 clientes bancários.

Renda X	Poupança Y	X . Y	X^2	Y^2
10	4	40	100	16
15	7	105	225	49
12	5	60	144	25
70	20	1400	4900	400
80	20	1600	6400	400
100	30	3000	10000	900
20	8	160	400	64
30	8	240	900	64
10	3	30	100	9
60	15	900	3600	225
407	120	7535	26769	2152

$$r = \frac{n.\sum X_i.Y_i - \left(\sum X_i\right).\left(\sum Y_i\right)}{\sqrt{\left[n.\sum X_i^2 - \left(\sum X_i\right)^2\right].\left[n.\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}}$$

$$r = \frac{10.7535 - 407.120}{\sqrt{\left(10.26769 - 407^2\right).\left(10.2152 - 120^2\right)}}$$

$$r = 0{,}984$$

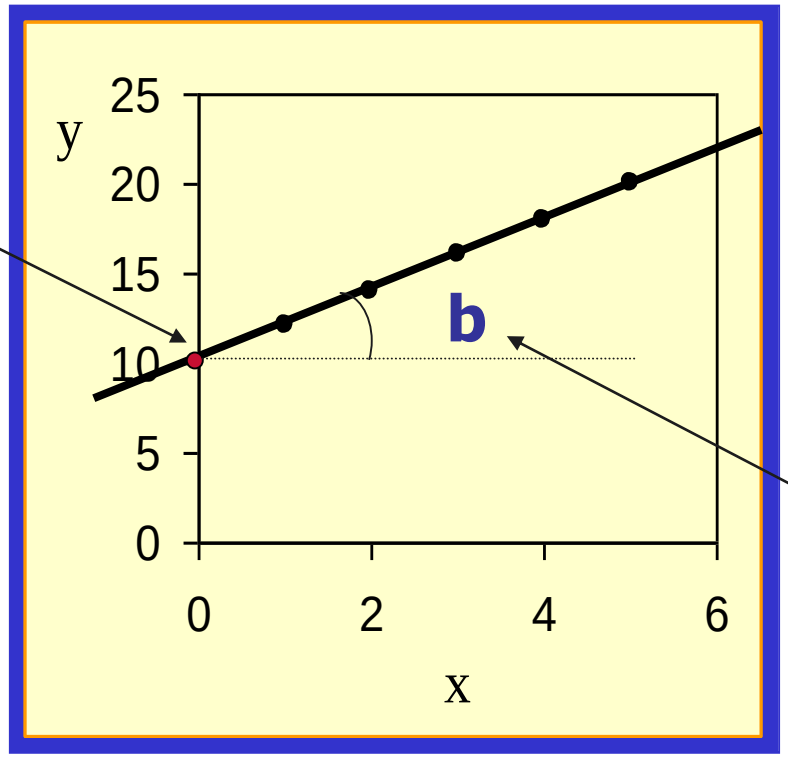
Existe alta correlação positiva entre poupança e renda.

Regressão Linear Simples

Na Regressão Linear Simples, o objetivo é construir um modelo explicativo (uma reta) baseado em apenas uma variável independente.

$$Y = a + b\,X$$

a



a e b - parâmetros da reta

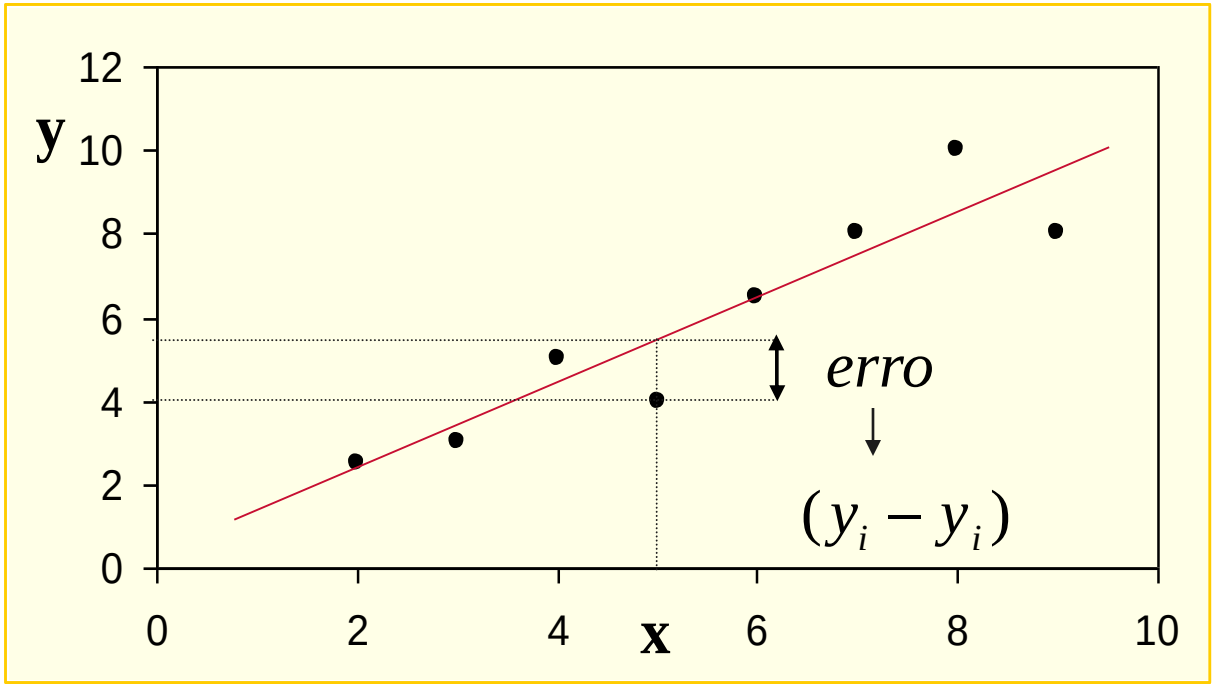
a é chamado coeficiente linear

b é chamado coeficiente angular

Inclinação da reta

Regressão Linear Simples

Método dos Mínimos Quadrados



O objetivo é minimizar a soma do quadrado dos erros (resíduos):

$$SQR = \sum \left(y - \hat{y} \right)^2$$

Regressão Linear Simples

Método dos Mínimos Quadrados

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

Podemos utilizar a reta de regressão para estimar os valores de y.

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Regressão Linear Simples

- Reflete o poder explicativo do modelo e a qualidade do ajuste

- Indica a proporção da variação total da variável dependente Y explicada pela equação de regressão

$$R^2 = \frac{\text{Variância de Y explicada pela regressão}}{\text{Variância total de Y}}$$

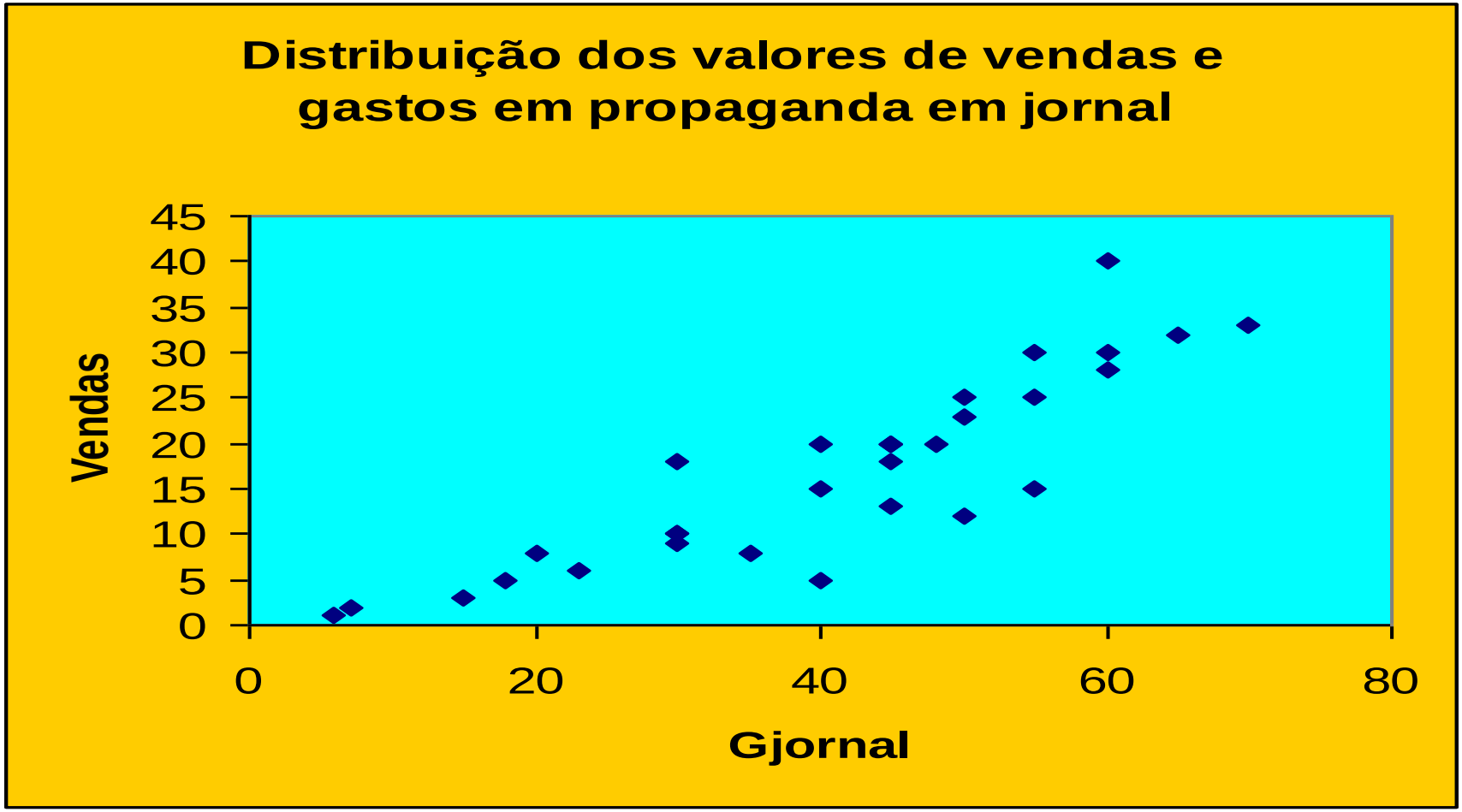
$$0 \leq R^2 \leq 1$$

Dados para Desenvolvimento de um Modelo

Projetar (Estimar) as Vendas de um modelo específico de automóvel em função do Valor gasto em propaganda na mídia impressa jornal

Vendas (em milhares)	Gjornal (Em milhões)
6	1
7	2
15	3
18	5
20	8
23	6
30	10
35	8
40	5
30	9
50	12
55	15
40	20
45	13
30	18
50	23
55	25
60	30
30	10
40	15
45	18
45	20
50	25
45	20
60	28
48	20
65	32
70	33
55	30
60	40

Diagrama de Dispersão dos Dados



Resultados do Ajuste do Modelo de Regressão Linear Simples

RESUMO DOS RESULTADOS	
Estatística de regressão	
R-Quadrado	0,77
Observações	30

Correlação de Pearson
r = 0,88

Ajuste do Modelo de Modelo Regressão Linear Simples

	Coeficientes
Interseção	16,633
Gjornal	1,435

$$Y = 16{,}633 + 1{,}435 \cdot Gjornal$$

Ajuste do Modelo de Modelo Regressão Linear Simples

	Coeficientes
Interseção	16,633
Gjornal	1,435

$$Y = 16{,}633 + 1{,}435 \,.\, Gjornal$$

Questão:

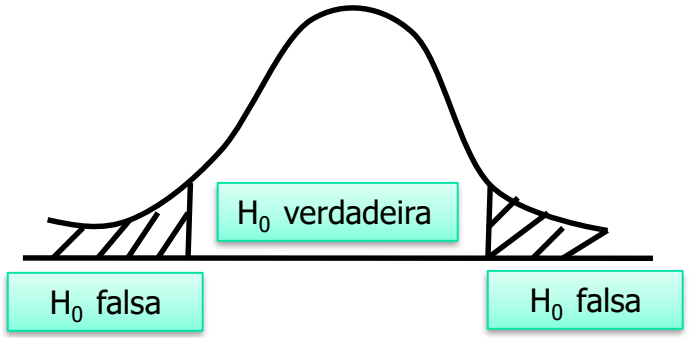
Estes coeficientes são, de fato, relevantes, ou seja, são estatisticamente diferentes de zero?

Teste de Hipóteses

- HIPÓTESES:

H_0: o coeficiente linear é igual a zero (não vale a pena incluir o coef linear)
H_1: o coeficiente linear é diferente de zero (vale a pena incluir o coef linear)

H_0: o coeficiente angular é igual a zero (não vale a pena incluir Gjornal)
H_1: o coeficiente angular é diferente de zero (vale a pena incluir Gjornal)



Interpretação dos Resultados Modelo Regressão Linear Simples

	Coeficientes	valor-P
Interseção	16,633	0,000033159479
Gjornal	1,435	0,000000001499

INTERPRETAÇÃO

H_0: o coeficiente linear é igual a zero
H_1: o coeficiente linear é diferente de zero
Área = 0,0000033 é menor do que 0,05
Decisão: rejeitar H_0

H_0: o coeficiente angular é igual a zero
H_1: o coeficiente angular é diferente de zero
Área = 0,000000001499 é menor do que 0,05
Decisão: rejeitar H_0

Previsão

- Qual a Previsão de Vendas referente a Gastos com Propaganda em Jornal de 65?

Previsão

- Qual a Previsão de Vendas referente a Gastos com Propaganda em Jornal de 65?

$$Y = 16{,}633 + 1{,}435 \cdot Gjornal$$

Previsão

- Qual a Previsão de Vendas referente a Gastos com Propaganda em Jornal de 65?

$$Y = 16{,}633 + 1{,}435 \cdot Gjornal$$

$$Y = 16{,}633 + 1{,}435 \cdot 65$$

$$Y = 109{,}908$$

Regressão Linear Múltipla

- Reta Estimada

$$Y_i = a + b_1 . X_{1i} + b_2 . X_{2i} + ... + b_k . X_{ki} + e_i$$

a - COEFIC. LINEAR DA FÓRMULA
Valor do eixo vertical interceptado pela reta (Intercepto do Eixo y).

b_i - COEFIC. ANGULAR DA i-ésima VARIÁVEL

i – Número de Variáveis Independentes

Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis resulta um plano, e um problema de k variáveis resulta um hiperplano.

- PASSOS DA ANÁLISE DE REGRESSÃO

  - Uso da **Correlação de Pearson**
  - Seleção de Variáveis Independentes
  - Estimação dos Coeficientes da Reta

Método dos Mínimos Quadrados: Obter a reta que melhor se ajusta aos dados

  - Previsão

Dados Adicionais para desenvolvimento de um Modelo

Vendas	Gjornal	Gtv	Gmdireta
6	1	3	4
7	2	4	4
15	3	8	16
18	5	8	8
20	8	10	30
23	6	11	2
30	10	15	10
35	8	18	20
40	5	20	10
30	9	15	10
50	12	13	5
55	15	18	4
40	20	15	23
45	13	20	30
30	18	15	20
50	23	18	10
55	25	25	25
60	30	15	10
30	10	15	34
40	15	18	32
45	18	24	25
45	20	30	18
50	25	25	30
45	20	20	25
60	28	25	35
48	20	15	28
65	32	21	30
70	33	20	26
55	30	17	21
60	40	25	29

Resultados do Ajuste do Modelo de Regressão Linear Múltipla

	Vendas	*Gjornal*	*Gtv*	*Gmdireta*
Vendas	1			
Gjornal	0,87993	1		
Gtv	0,76978	0,68186	1	
Gmdireta	0,42428	0,51339	0,5454	1

Estatística de regressão	
R-Quadrado	**0,840**
Observações	**30**

Ajuste do Modelo de Regressão Linear Múltipla

	Coeficientes
Interseção	9,371
Gjornal	1,140
Gtv	0,970
Gmdireta	-0,217

Y = 9,371 + 1,140 . Gjornal +

0,970 . Gtv - 0,217 . Gmdireta

Interpretação dos Resultados do Modelo

	Coeficientes	valor-P
Interseção	9,371	0,021115
Gjornal	1,140	1,01E-06
Gtv	0,970	0,00319
Gmdireta	-0,217	0,173598

INTERPRETAÇÃO

H_0: o coeficiente linear é igual a zero
H_1: o coeficiente linear é diferente de zero
Área = 0,021115 é menor do que 0,05
Decisão: rejeitar H_0

H_0: o coeficiente de Gjornal é igual a zero
H_1: o coeficiente de Gjornal é diferente de zero
Área = 0,00000101 é menor do que 0,05
Decisão: rejeitar H_0

Problema: Gmdireta – Solução Refazer o modelo sem essa variável

Ajuste do Modelo de Regressão Linear Múltipla

NOVA REGRESSÃO MÚLTIPLA	
Estatística de regressão	
R-Quadrado	0,828
Observações	30

	Coeficientes
Interseção	8,384
Gjornal	1,082
Gtv	0,840

Y = 8,384 + 1,082 . Gjornal + 0,84 . Gtv

Interpretação dos Resultados do Modelo

	Coeficientes	valor-P
Interseção	8,384	0,03688
Gjornal	1,082	1,7E-06
Gtv	0,840	0,007165

INTERPRETAÇÃO

H_0: o coeficiente linear é igual a zero
H_1: o coeficiente linear é diferente de zero
Área = 0,03688 é menor do que 0,05
Decisão: rejeitar H_0

H_0: o coeficiente de Gjornal é igual a zero
H_1: o coeficiente de Gjornal é diferente de zero
Área = 0,0000017 é menor do que 0,05
Decisão: rejeitar H_0

H_0: o coeficiente de Gtv é igual a zero
H_1: o coeficiente de Gtv é diferente de zero
Área = 0,007165 é menor do que 0,05
Decisão: rejeitar H_0

Interpretação dos Resultados do Modelo

F	F de significação
65,05973	4,72396E-11

INTERPRETAÇÃO

H_0: $b_1 = b_2 = ... = b_k = 0$

H_1: existe pelo menos um coeficiente b_i diferente de 0

Este teste mede o desempenho global da reta de regressão e procura detectar se, de um modo geral, a regressão foi relevante.

Área = 0,000000000472396 é menor do que 0,05

Decisão: rejeitar H_0

Logo, a regressão é relevante.

Previsão

- Qual a Previsão de Vendas Referente a Gastos com Propaganda em Jornal de 65 e Gastos com TV de 40 ?

Previsão

- Qual a Previsão de Vendas Referente a Gastos com Propaganda em Jornal de 65 e Gastos com TV de 40 ?

$$Y = 8{,}384 + 1{,}082 \cdot Gjornal + 0{,}84 \cdot Gtv$$

$$Y = 112{,}314$$

Procedimento Stepwise

Se o conjunto de variáveis explicativas for razoável, o número de modelos a serem analisados torna enfadonho e pouco produtivo.

O **algoritmo Stepwise** consiste no ajuste de diversos modelos, partindo do modelo bem elementar, com apenas uma variável explicativa, e sucessivamente, acrescentar ou retirar as demais variáveis.

Procedimento Stepwise.

- ✓ A primeira variável a entrar no modelo é a mais correlacionada com a variável resposta.
- ✓ A segunda variável a entrar é aquela tem a maior estatística F modificada.
- ✓ No novo modelo, é feito um teste se alguma variável deve ser retirada antes da inclusão da próxima variável.
- ✓ E assim sucessivamente, até que esse aumento se torne irrelevante.

Procedimento Stepwise

Neste tipo de procedimento, a sequência de variáveis no modelo atualizado pode ser:

1. X_1;
2. X_1X_2;
3. $X_1X_2X_3$;
4. X_2X_3;
5. $X_2X_3X_4$;
6. $X_2X_3X_4X_5$;
7. $X_2X_4X_5$.

Procedimento Stepwise

As variações do algoritmo consistem em:

- ✓ **Seleção para frente**, semelhante a anterior sem retirar variáveis;
- ✓ **Eliminação para trás**, onde se parte do modelo completo ou com todas as variáveis, e sucessivamente vai se retirando variáveis que menos contribuem para o modelo.

Variável Dummy

- A análise de Regressão pressupõe que as variáveis estejam em escala intervalar. Entretanto em algumas situações o pesquisador pode querer incluir variáveis qualitativas na análise.
- A forma de fazer isso é utilizar uma variável Dummy que é transformar a variável qualitativa em uma variável binária que indica presença ou ausência de alguma característica.

- Por exemplo, um banco está fazendo uma análise de sua carteira de clientes e deseja colocar no modelo se o cliente possui ou não cartão de crédito. A Variável seria:

X = 0 , se o cliente não possui cartão de crédito
1, se o cliente possui cartão de crédito

Análise de Resíduos

- Suposições
  - Distribuição Normal
  - Variância Constante
  - Ausência de Correlação
  - Ausência de Outliers

Tamanho da Amostra

- Número Mínimo de Observações por Variável Independente: 5

- Número Recomendado: 15 à 20 Observações por Variável Independente

Exercícios

1) Um modelo foi construído para prever o Lucro anual de uma empresa do setor financeiro (em R$ Milhões) utilizando a taxa de Juros média praticada no período (em %) e o valor total dos empréstimos concedidos no ano (em R$ Bilhões). O resultado do modelo de regressão foi:

RESUMO DOS RESULTADOS

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	-14,98	7,8026	2,164	0,026
Tx_Juros	-2,88	0,9099	-3,21	0,002
Vl_Emprestado	25,06	10,9361	2,47	0,037

Com base nesses resultados responda:

a) Avalie os resultados do modelo
b) Espera-se que em 2019 os Juros médios sejam de 19% e que a empresa empreste R$ 1,80 Bilhões. Se essa expectativa se confirmar, qual deverá ser o lucro da empresa.

Exercícios

2) O gerente de uma loja de variedades precisa prever o tempo que leva para se fazer o checkout de um cliente. Ele decidiu usar as seguintes variáveis independentes: a quantidade de produtos que o cliente compra e a quantia comprada em moeda. Os resultados são mostrados a seguir:

Correlação

	Tempo de check out (Minutos)	*Quantia gasta (em R$)*	*Quantidade de itens comprados*
Tempo de check out (Minutos)	1		
Quantia gasta (em R$)	0,95948566	1	
Quantidade de itens comprados	0,876109252	0,92302654	1

RESUMO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,959804771
R-Quadrado	0,921225199
R-quadrado ajustado	0,910721892
Erro padrão	0,85751084
Observações	18

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	2	128,9879052	64,49395258	87,70811	5,2833E-09
Resíduo	15	11,02987262	0,735324841		
Total	17	140,0177778			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
Interseção	0,421662251	0,586384916	0,71908782	0,48314	-0,828188381	1,671512883	-0,828188381	1,671512883
Quantia gasta (em R$)	0,087146866	0,016111001	5,409152734	7,24E-05	0,052807059	0,121486673	0,052807059	0,121486673
Quantidade de itens compra	-0,038627658	0,113111096	-0,34150193	0,737463	-0,279718399	0,202463084	-0,279718399	0,202463084

Avalie os resultados.

O que você faria como próximo passo? Por que?