



Preditiva.ai

# Inferência Estatística

## Noções de Regressão Linear

# Noções de Regressão Linear

## Introdução



A **Regressão Linear** é um dos métodos mais famosos para se interpretar a relação de uma variável de seu interesse (**variável resposta**) com outra variável (**variável explicativa**).

Enquanto com a correlação linear de Pearson conseguimos quantificar o nível de associação entre as duas variáveis, com a regressão conseguimos inclusive “**modelar**” esse grau de associação construindo uma equação que explica a mudança da variável resposta em relação às mudanças da variável explicativa.

Vejamos um exemplo a seguir.

# Noções de Regressão Linear

## Exemplo

Uma empresa está interessada em verificar se a **remuneração** dos seus colaboradores está de acordo com as políticas vigentes. Para isso coletou uma amostra de **46 colaboradores** conforme a seguir.

A pergunta de negócio é:

O salário muda de acordo com os anos de educação superior? Se sim, é possível construir um modelo que estime o salário de acordo com esses anos?



Preditiva.ai

Exemplo dos primeiros 15 funcionários

Núm. Funcionário	Salário	Anos de Educação Superior
1	5.517,4	3
2	6.399,9	4
3	6.206,7	6
4	6.060,6	4
5	6.122,7	2
6	6.955,0	5
7	7.643,0	4
8	6.210,2	2
9	5.761,0	9
10	8.086,9	6
11	6.375,4	4
12	9.568,8	6
13	9.316,0	6
14	6.822,4	9
15	6.570,9	4

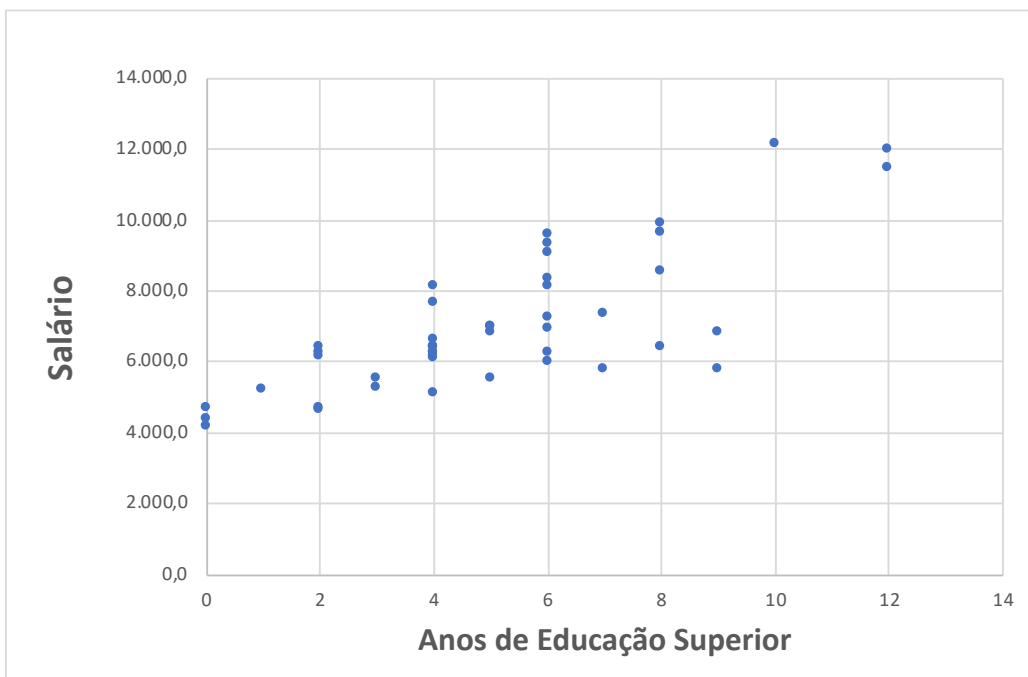
# Noções de Regressão Linear

## Exemplo



Como toda análise bidimensional de duas variáveis quantitativas, devemos **construir um gráfico de dispersão**.

Vejamos a seguir.



O gráfico de dispersão nos informa que as duas variáveis são **correlacionadas positivamente**, pois aparentemente quanto maior os anos de educação, maior é o salário nesta empresa.

Pelo formato da “nuvem de pontos”, também é possível sugerir que uma reta pode ser traçada no meio da nuvem.

Portanto, as variáveis têm uma associação aproximadamente linear e podemos quantificar essa associação com o **coeficiente de correlação de Pearson** ( Correlação = 0,78 ).

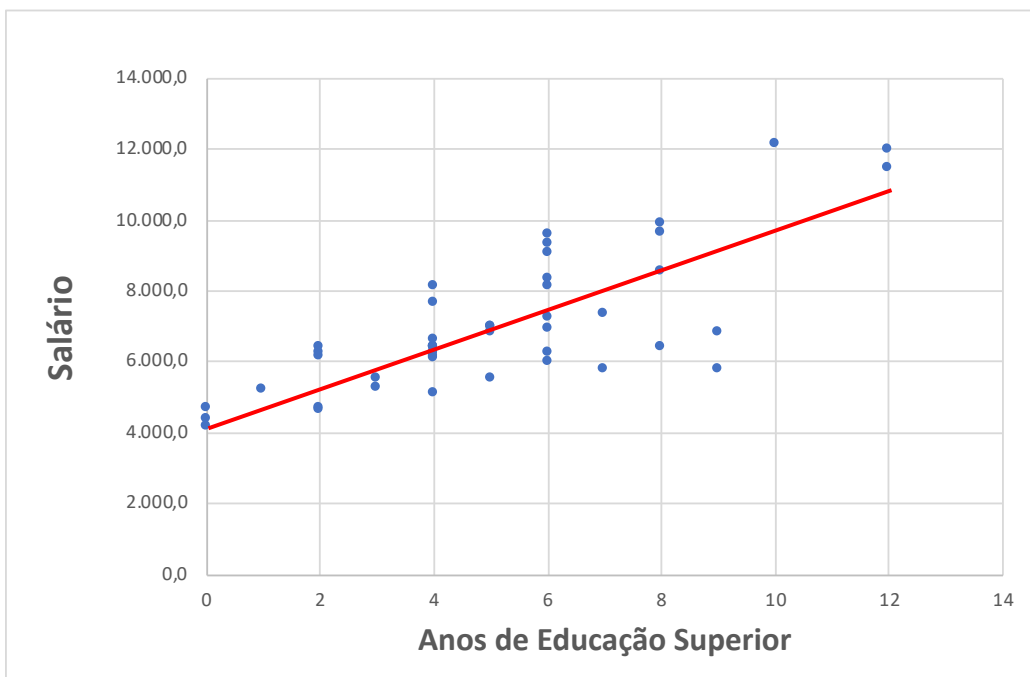
# Noções de Regressão Linear

## Exemplo – Uma variável explicativa



Como toda análise bidimensional de duas variáveis quantitativas, devemos **construir um gráfico de dispersão**.

Vejamos a seguir.



Das relações de geometria, podemos supor que a **reta vermelha** ao lado segue a seguinte relação:

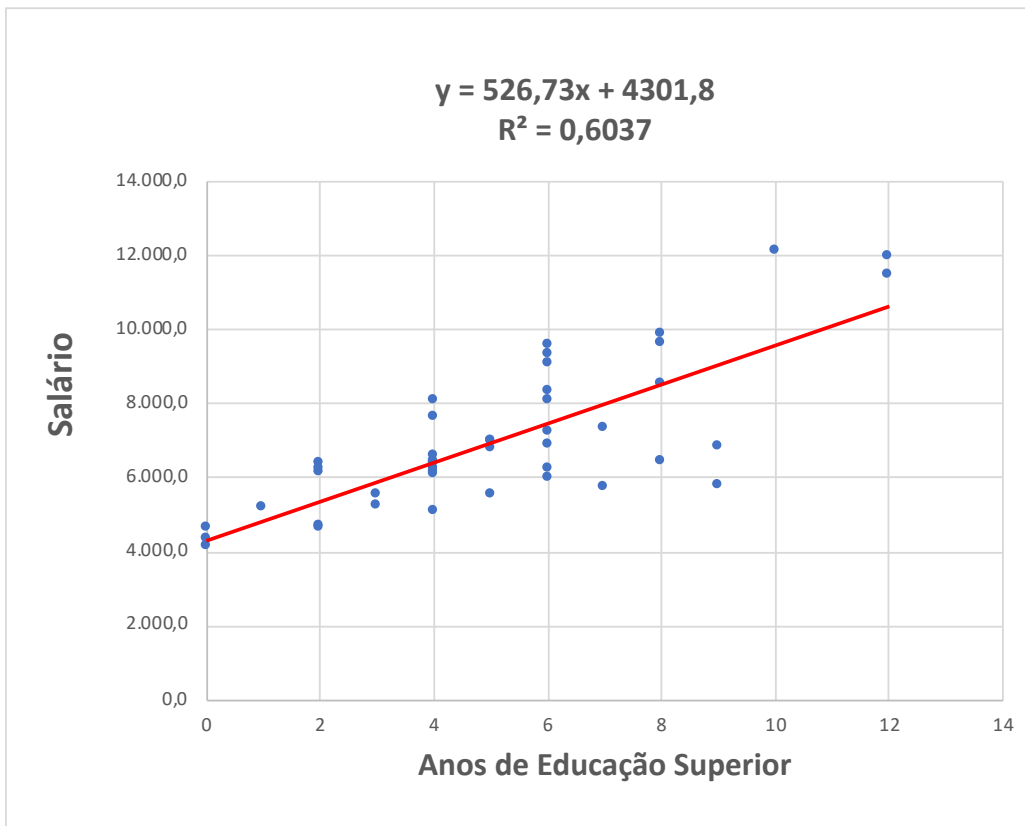
$$\text{Salário} = \underset{\substack{\text{(Intercepto)} \\ \text{(Coeficiente} \\ \text{Angular de} \\ \text{Anos de Educação)}}}{\beta_0 + \beta_1 * \text{Anos de Educacao}}$$

Resta agora saber quais são os valores de Beta 0 e Beta 1 que produzem a reta que passa no meio da nuvem de pontos.

Podemos encontrar esses valores no Excel utilizando a função **“Solver”** conforme mostramos a seguir.

# Noções de Regressão Linear

## Exemplo – Uma variável explicativa



Portanto, chegamos ao modelo desejado. O modelo que estima o salário pelos anos de educação é:

$$\text{Salario} = 4.301,8 + 526,7 * \text{Anos de Educacao}$$

Com posse deste modelo, podemos ter várias interpretações.

Seguem algumas:

1. O salário médio estimado de um funcionário com 0 anos de educação superior é de R\$ 4.301,8 reais, ou seja, o próprio valor do **intercepto da reta**.
2. O salário médio estimado de um funcionário aumenta em R\$ 526,7 reais a cada ano de educação superior, ou seja, o próprio valor do **coeficiente angular da reta**.

Além disso, o **R quadrado** calculado (0,6037) é uma medida da qualidade da explicação dos salários introduzida pela variável “Anos de Educação”. Sua interpretação é: **A variável “Anos de Educação” explica 60,37% da variabilidade dos Salários na empresa.**

Ao mostrar esse estudo em uma reunião de RH, um dos gerentes fez a seguinte pergunta:

**“Se os anos de **Educação Superior** dos funcionários explicam praticamente 60% de seus salários, quais outros fatores (variáveis) explicam o restante?”**

Essa é uma pergunta interessante. Se tivermos acesso a mais variáveis, é possível contruir um modelo que contemple todas essas **múltiplas variáveis**? A resposta é positiva e vamos mostrar a seguir.

# Noções de Regressão Linear

## Exemplo – Mais de uma variável explicativa



Preditiva.ai

Exemplo dos primeiros 15 funcionários

A mesma empresa tentando melhorar o modelo, conseguiu tabular outra informação a respeito da amostra de 46 funcionários: **o tempo (em anos) de empresa**. Veja a seguir:

A pergunta de negócio é:

**O salário muda de acordo com os anos de educação superior e Tempo de Empresa? Se sim, é possível construir um modelo que estime o salário de acordo com essas duas variáveis?**

Núm. Funcionário	Salario	Anos de Educação Superior	Tempo na Empresa
1	5.517,4	3	3
2	6.399,9	4	6
3	6.206,7	6	3
4	6.060,6	4	5
5	6.122,7	2	9
6	6.955,0	5	9
7	7.643,0	4	6
8	6.210,2	2	8
9	5.761,0	9	15
10	8.086,9	6	14
11	6.375,4	4	9
12	9.568,8	6	20
13	9.316,0	6	25
14	6.822,4	9	18
15	6.570,9	4	19



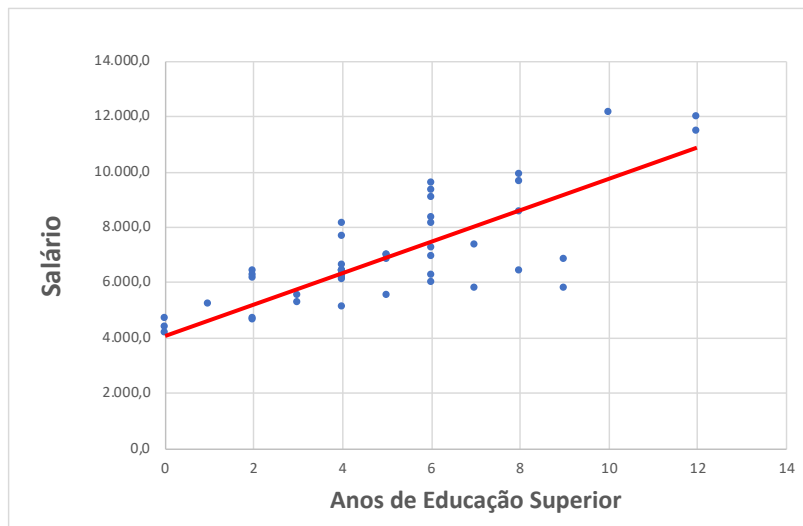
# Noções de Regressão Linear

## Exemplo – Mais de uma variável explicativa

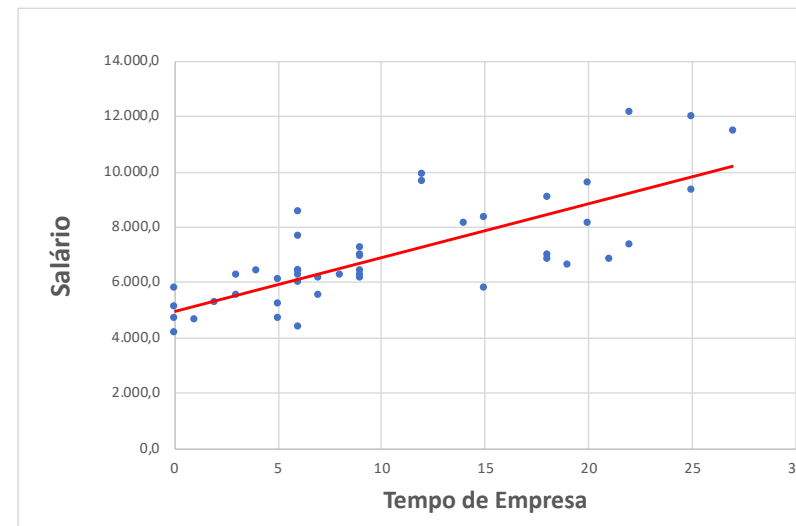


Da mesma forma que fizemos para a variável “Anos de Educação”, também devemos **construir um gráfico de dispersão** para a nova variável. Vejamos a seguir.

**Variável Anos de Educação (Correlação = 0,78)**



**Variável Tempo de Empresa (Correlação = 0,76)**



Como ambas as variáveis explicativas são aproximadamente lineares com a variável resposta (Salário), podemos construir um modelo que pode estimar o salário de um funcionário de acordo com seus **anos de educação e tempo de empresa**.

# Noções de Regressão Linear

## Exemplo – Mais de uma variável explicativa



No entanto, este modelo não segue mais a interpretação de “reta”, pois temos duas variáveis explicativas. Felizmente, mesmo assim podemos seguir o mesmo método para encontrar o modelo de interesse. Vejamos o seu novo formato:

$$\text{Salario} = \overset{\text{(Intercepto)}}{\beta_0} + \underset{\substack{\text{(Coeficiente} \\ \text{de} \\ \text{Anos de Educação)}}}{\beta_1} * \text{Anos de Educacao} + \underset{\substack{\text{(Coeficiente} \\ \text{de} \\ \text{Tempo de Empresa)}}}{\beta_2} * \text{Tempo de Empresa}$$

Podemos usar o mesmo processo anterior usando a função “Solver” do Excel. Porém, existe uma forma mais rápida utilizando o recurso “[Análise de Dados \ Regressão](#)” da aba Dados. Vejamos no exemplo a seguir.

# Noções de Regressão Linear

## Resultado da regressão via Excel



Vejamos abaixo como interpretar os resultados da regressão linear utilizando o recurso “Análise de Dados”:

	A	B	C	D	E	F	G
1	RESUMO DOS RESULTADOS						
2							
3	Estatística de regressão						
4	R múltiplo	0,8602	<div>Este é o R quadrado da Regressão. O nome “R quadrado que é feita quanto temos mais de Podemos concluir que o <b>acréscimo de uma explicação dos salários de 6</b></div>				
5	R-Quadrado	0,7399					
6	R-quadrado ajustado	0,7278					
7	Erro padrão	1.004,20					
8	Observações	46					
9							
10	ANOVA						
11		gl	SQ	MQ	F	F de significação	
12	Regressão	2	123.368.161	61.684.081	61,169	0,000000000000027	
13	Resíduo	43	43.362.038	1.008.419			
14	Total	45	166.730.199				
15							
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
17	Interseção	4.056,06	309,70	13,10	0,000000	3431,48	4680,63
18	Anos de Educação Superior	335,39	66,37	5,05	0,000009	201,53	469,24
19	Tempo na Empresa	117,66	24,79	4,75	0,000023	67,66	167,66

Este é o R quadrado da Regressão. O nome “ajustado”, vem de uma correção do R quadrado que é feita quanto temos mais de uma variável explicativa no modelo. Podemos concluir que o **acréscimo de uma nova variável aumentou o grau de explicação dos salários de 60,37% para 72,78%.**

# Noções de Regressão Linear

## Resultado da regressão via Excel



Vejamos abaixo como interpretar os resultados da regressão linear utilizando o recurso “Análise de Dados”:

	A	B	C	D	E	F	G
1	RESUMO DOS RESULTADOS						
2							
3	Estatística de regressão						
4	R múltiplo	0,8602	Esta é a “Soma dos erros ao quadrado” da regressão método de regressão utilizando o				
5	R-Quadrado	0,7399					
6	R-quadrado ajustado	0,7278					
7	Erro padrão	1.004,20					
8	Observações	46					
9							
10	ANOVA						
11		gl	SQ	MQ	F	F de significação	
12	Regressão	2	123.368.161	61.684.081	61,169	0,000000000000027	
13	Resíduo	43	43.362.038	1.008.419			
14	Total	45	166.730.199				
15							
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
17	Interseção	4.056,06	309,70	13,10	0,000000	3431,48	4680,63
18	Anos de Educação Superior	335,39	66,37	5,05	0,000009	201,53	469,24
19	Tempo na Empresa	117,66	24,79	4,75	0,000023	67,66	167,66

Esta é a “Soma dos erros ao quadrado” da regressão (já calculamos este valor usando o método de regressão utilizando o “Solver” do Excel).

# Noções de Regressão Linear

## Resultado da regressão via Excel



Vejamos abaixo como interpretar os resultados da regressão linear utilizando o recurso “Análise de Dados”:

	A	B	C	D	E	F	G
1	RESUMO DOS RESULTADOS						
2							
3	Estatística de regressão						
4	R múltiplo	0,8602					
5	R-Quadrado	0,7399					
6	R-quadrado ajustado	0,7278					
7	Erro padrão	1.004,20					
8	Observações	46					
9							
10	ANOVA						
11		gl	SQ	MQ	F	F de significação	
12	Regressão	2	123.368.161	61.684.081	61,169	0,000000000000027	
13	Resíduo	43	43.362.038	1.008.419			
14	Total	45	166.730.199				
15							
16		Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
17	Interseção	4.056,06	309,70	13,10	0,000000	3431,48	4680,63
18	Anos de Educação Superior	335,39	66,37	5,05	0,000009	201,53	469,24
19	Tempo na Empresa	117,66	24,79	4,75	0,000023	67,66	167,66

Estes são os coeficientes do modelo completo de regressão.

$$\text{Salario} = \beta_0 + \beta_1 * \text{Anos de Educacao} + \beta_2 * \text{Tempo de Empresa}$$

# Noções de Regressão Linear

## Resultado da regressão via Excel



Vejamos abaixo como interpretar os resultados da regressão linear utilizando o recurso “Análise de Dados”:

	A	B	C	D	E	F	G
1	RESUMO DOS RESULTADOS						
2							
3	Estatística de regressão						
4	R múltiplo	0,8602					
5	R-Quadrado	0,7399					
6	R-quadrado ajustado	0,7278					
7	Erro padrão	1.004,20					
8	Observações	46					
9							
10	ANOVA						
11		gl	SQ	MQ	F	F de significação	
12	Regressão	2	123.368.161	61.684.081	61,169	0,000000000000027	
13	Resíduo	43	43.362.038	1.008.419			
14	Total	45	166.730.199				
15							
16		Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
17	Interseção	4.056,06	309,70	13,10	0,000000	3431,48	4680,63
18	Anos de Educação Superior	335,39	66,37	5,05	0,000009	201,53	469,24
19	Tempo na Empresa	117,66	24,79	4,75	0,000023	67,66	167,66

Estes são os valores P dos Testes de Hipótese realizados para cada coeficiente. A hipótese nula do teste é se o coeficiente é igual a 0 (zero). Como podemos concluir, rejeita-se a hipótese nula de igualdade a 0 (zero) ao nível de significância de 5% e, portanto, **o salário na população é explicado pelos Anos de Educação Superior e Tempo na Empresa.**

# Noções de Regressão Linear

## O que mais é necessário saber?



Esta aula serviu como uma introdução ao riquíssimo tema de modelos de regressão. Existem vários outros fatores e cuidados que devemos tomar, como por exemplo:

- ☐ Quando trabalhar com regressões em amostras (processo inferencial), é preciso respeitar as suposições da regressão. Ex: Inexistência de Multicolinearidade Perfeita, Homocedasticidade (Variância dos Resíduos constante), Resíduos seguindo uma distribuição Normal, independência entre as observações, entre outros.
- ☐ Utilização de variáveis Dummy quando tivermos variáveis qualitativas no modelo. Ex: Homem ou Mulher; Estado Civil etc.
- ☐ Verificação da Causalidade das variáveis explicativas.
- ☐ Entre outros.

# Revisão e próximos passos

---

O método para se interpretar a relação de uma variável de seu interesse (variável resposta) com outra variável (variável explicativa) é chamado de Regressão Linear. Existem basicamente dois tipos de regressão linear. São eles:

- **Regressão Linear Simples**: Quando temos apenas **uma** variável explicativa no modelo.
- **Regressão Linear Múltipla**: Quando temos **mais de uma** variável explicativa no modelo.

O **R quadrado** é a principal medida da qualidade dos modelos estimados. Essa medida informa o grau de variabilidade da variável resposta explicada pelas variáveis explicativas.

Nos vemos nas próximas aulas!







Preditiva.ai