Preditiva.ai

Inferência Estatística

Noções de Regressão Linear

Noções de Regressão Linear

Introdução

Preditiva.ai

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

Preditiva.ai

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?

Exemplo dos primeiros 15 funcionários

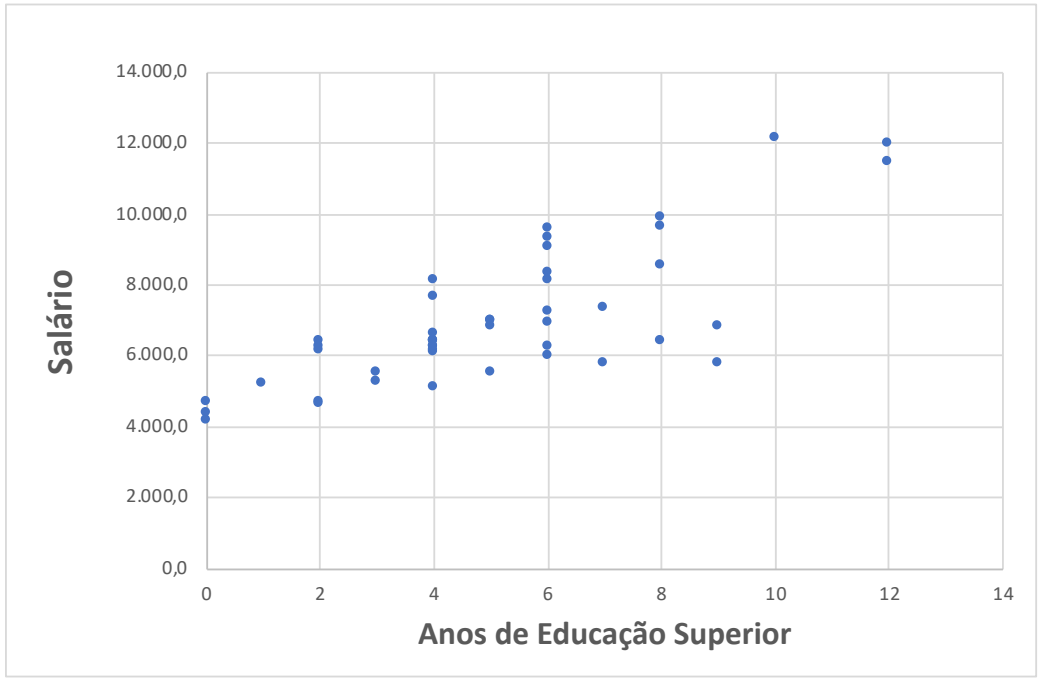
Núm. Funcionário	Salário	Anos de Educação Superior
1	5.517,4	3
2	6.399,9	4
3	6.206,7	6
4	6.060,6	4
5	6.122,7	2
6	6.955,0	5
7	7.643,0	4
8	6.210,2	2
9	5.761,0	9
10	8.086,9	6
11	6.375,4	4
12	9.568,8	6
13	9.316,0	6
14	6.822,4	9
15	6.570,9	4

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 núvem.

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

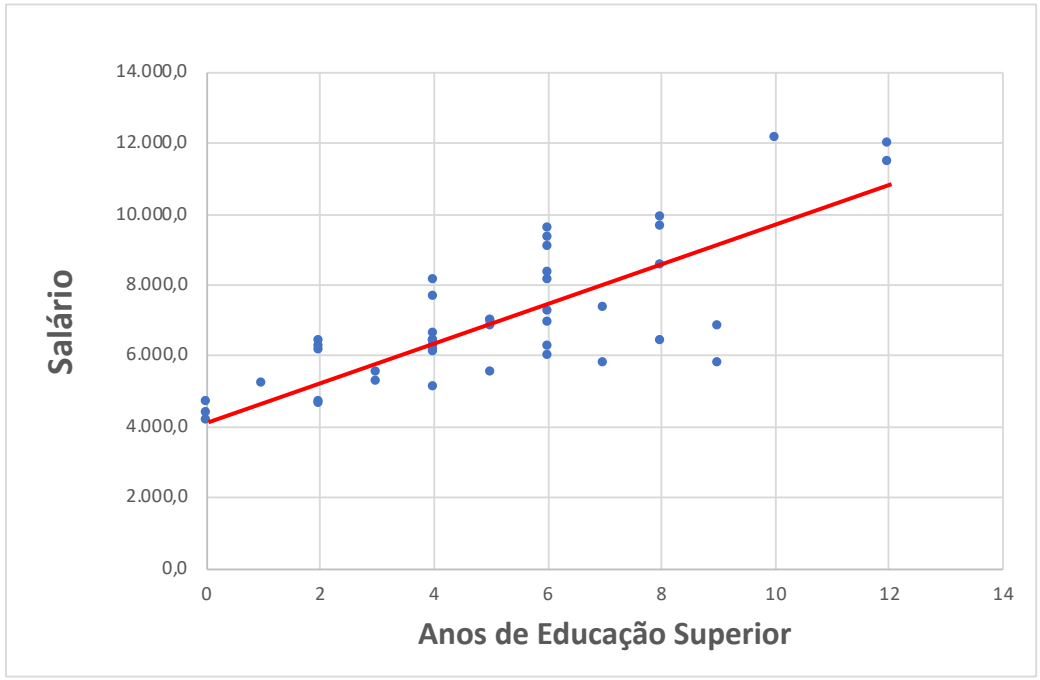
Noções de Regressão Linear

Exemplo – Uma variável explicativa

Preditiva.ai

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:

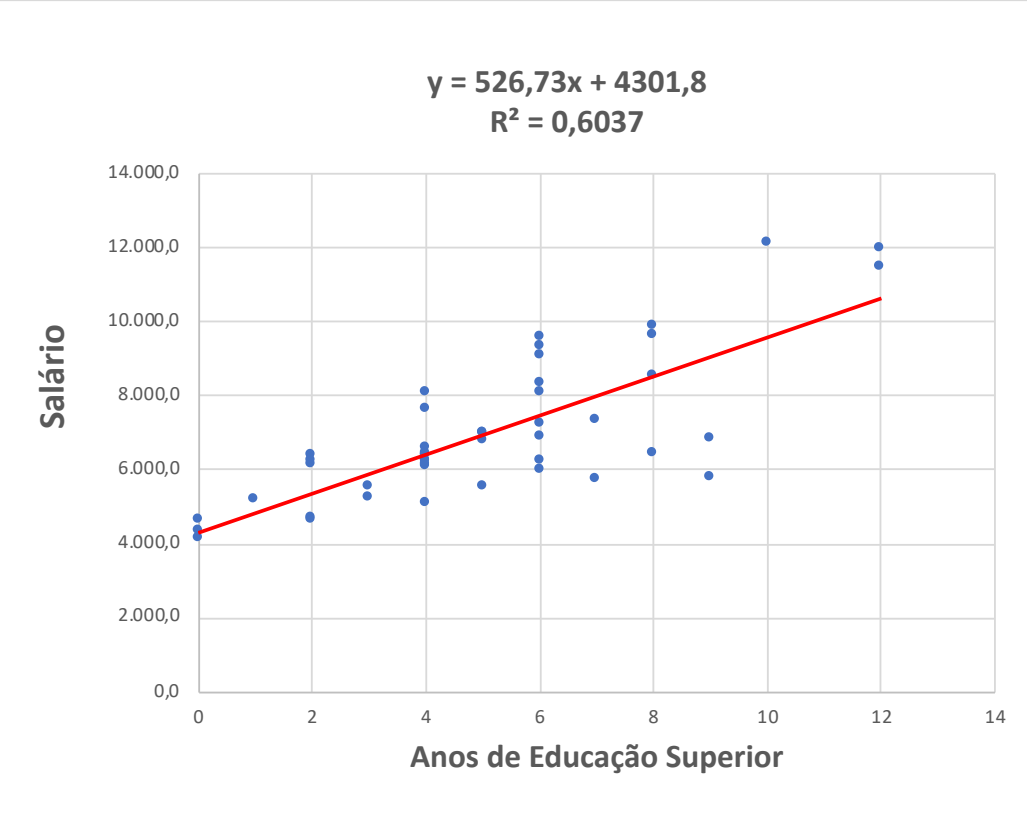
$$Salario = \underbrace{\beta_0}_{\text{(Intercepto)}} + \underbrace{\beta_1}_{\substack{\text{(Coeficiente Angular de Anos de Educação)}}} * Anos\ de\ Educacao$$

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

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

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**.

Noções de Regressão Linear

Exemplo

Preditiva.ai

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

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?

Exemplo dos primeiros 15 funcionários

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

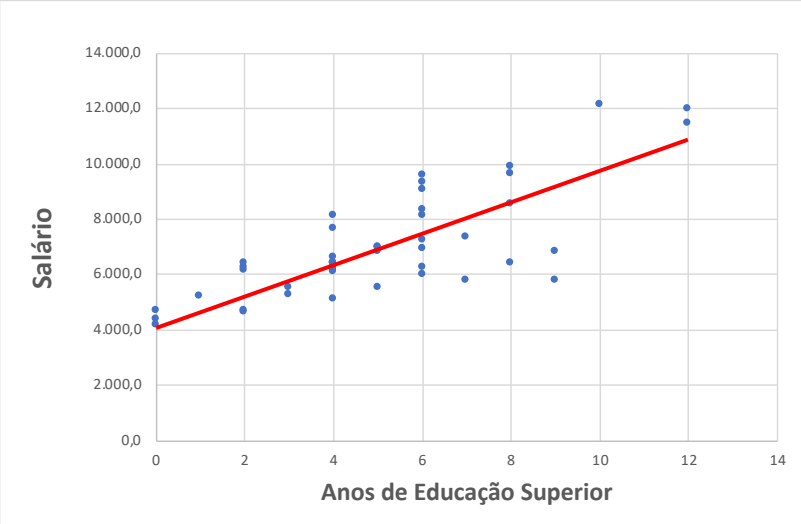
Noções de Regressão Linear

Exemplo – Mais de uma variável explicativa

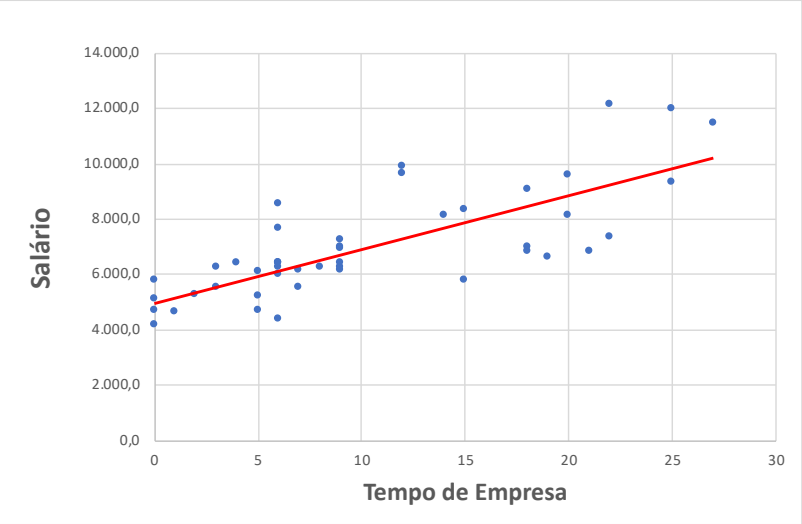
Preditiva.ai

Da mesma forma que fizemos para a varíavel “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

Preditiva.ai

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:

(Intercepto)

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

(Coeficiente de Anos de Educação)

(Coeficiente de 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

Preditiva.ai

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	0,00000000000027	
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

Preditiva.ai

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	0,00000000000027	
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

Preditiva.ai

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	0,00000000000027	
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

Estes são os coeficientes do modelo completo de regressão.

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

Noções de Regressão Linear

Resultado da regressão via Excel

Preditiva.ai

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	0,00000000000027	
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

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?

Preditiva.ai

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.

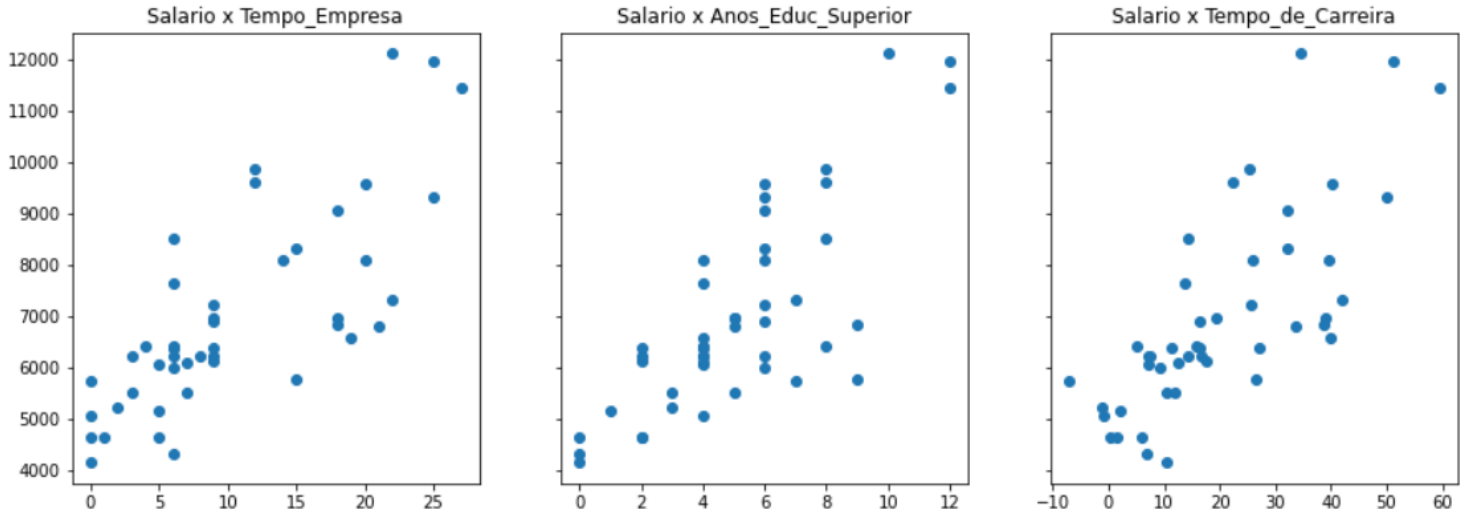
- ❑ Entre outros.

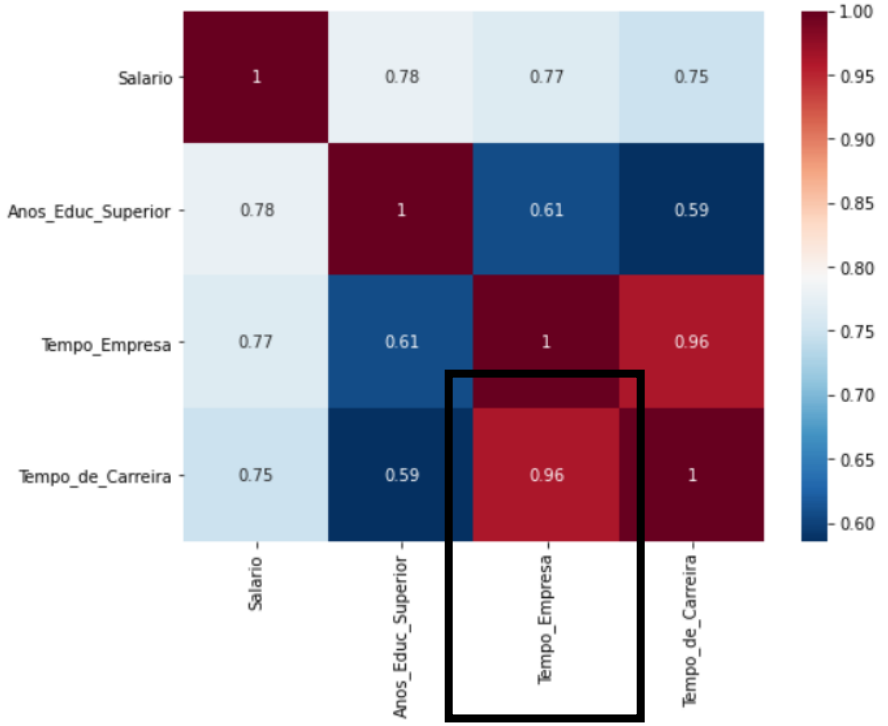
Regressão Linear

Problemas devido à Multicolinearidade

Preditiva.ai

Quando temos variáveis explicativas que são correlacionadas entre si, podemos prejudicar o **ajuste** do modelo (métrica e Performance não muda muito) e também sua **interpretabilidade** (a equação fica confusa e pouco intuitiva). Veja:





Todas as 3 variáveis explicativas são altamente correlacionadas com o target (Salário)

Tempo de Carreira e Tempo de Empresa **são correlacionadas entre si**, ou seja, existe **Multicolinearidade**.

Regressão Linear

Problemas devido à Multicolinearidade

Preditiva.ai

Em casos de multicolinearidade, quais as consequências para o modelo? Vejamos a seguir:

Modelo com baixa multicolinearidade

OLS Regression Results

Dep. Variable:	Salario	R-squared:	0.740
Model:	OLS	Adj. R-squared:	0.728
Method:	Least Squares	F-statistic:	61.17
Date:	Sun, 02 Aug 2020	Prob (F-statistic):	2.66e-13
Time:	19:13:48	Log-Likelihood:	-381.67
No. Observations:	46	AIC:	769.3
Df Residuals:	43	BIC:	774.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercepto	4056.0591	309.703	13.097	0.000	3431.483	4680.635
Anos_Educ_Superior	335.3868	66.372	5.053	0.000	201.535	469.238
Tempo_Empresa	117.6566	24.793	4.746	0.000	67.657	167.656

Modelo com alta multicolinearidade

OLS Regression Results

Dep. Variable:	Salario	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.723
Method:	Least Squares	F-statistic:	40.21
Date:	Sun, 09 Aug 2020	Prob (F-statistic):	2.06e-12
Time:	21:00:29	Log-Likelihood:	-381.51
No. Observations:	46	AIC:	771.0
Df Residuals:	42	BIC:	778.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercepto	4060.6182	312.397	12.998	0.000	3430.175	4691.061
Anos_Educ_Superior	335.3325	66.925	5.011	0.000	200.273	470.392
Tempo_Empresa	69.6669	92.205	0.756	0.454	-116.410	255.744
Tempo_de_Carreira	22.9483	42.440	0.541	0.592	-62.699	108.596

1

2

3

<u>Principais efeitos:</u>

1. O modelo mantém ou perde sua performance. R^2 não muda ou não aumenta.
2. Variáveis que antes importavam para o modelo (p-valor dentro da significância) perdem relevância (p-valor maior que significância).
3. É possível também que o sinal dos coeficientes fiquem trocados para compensar a multicolinearidade, o que leva a problemas da interpretabilidade do modelo para o negócio. Ex: Quanto maior o Tempo de Carreira, maior o salário. Faz sentido?

Regressão Linear
Problemas devido à Multicolinearidade

Preditiva.ai

Modelo com alta multicolinearidade

OLS Regression Results

Dep. Variable:	Salario	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.723
Method:	Least Squares	F-statistic:	40.21
Date:	Sun, 09 Aug 2020	Prob (F-statistic):	2.06e-12
Time:	21:00:29	Log-Likelihood:	-381.51
No. Observations:	46	AIC:	771.0
Df Residuals:	42	BIC:	778.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercepto	4060.6182	312.397	12.998	0.000	3430.175	4691.061
Anos_Educ_Superior	335.3325	66.925	5.011	0.000	200.273	470.392
Tempo_Empresa	69.6669	92.205	0.756	0.454	-116.410	255.744
Tempo_de_Carreira	22.9483	42.440	0.541	0.592	-62.699	108.596

O que fazer em situações como essa?

1. Remova a variável que estiver causando a multicolinearidade e ajuste/treine o modelo novamente. **Importante**: Faça isso variável a variável e não de uma única vez. Ex: Não remover Tempo de Empresa e Tempo de Carreira de uma única vez.

2. Para escolher entre essas variáveis, considere:

   - Qual a variável que mais tem relação causal com a variável resposta? Ex: As pessoas ganham mais pois têm mais tempo de empresa ou mais tempo de carreira?

   - Além disso, pondere pela dificuldade de ter acesso à variável no momento da implantação do modelo. Ex: Tempo de empresa é fácil medir pelo sistema de RH, porém Tempo de Carreira é necessário que o colaborador preencha um formulário.

Regressão Linear

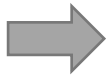
Lidando com variáveis Dummy

Preditiva.ai

Naturalmente, muitas variáveis do cotidiano não são numéricas. Como então trabalhamos com essas variáveis em um modelo de regressão linear? Através das variáveis Dummies. Vejamos:

Núm. Funcionário	Salario	Anos de Educação Superior	Tempo na Empresa	Possui Skill de Dados?
1	5.517,40	3	3	Não
2	6.399,90	4	6	Não
3	6.206,70	6	3	Não
4	6.060,60	4	5	Não
5	6.122,70	2	9	Não
6	6.955,00	5	9	Não
7	7.643,00	4	6	Sim
8	6.210,20	2	8	Não
9	5.761,00	9	15	Não
10	8.086,90	6	14	Sim
11	6.375,40	4	9	Não
12	9.568,80	6	20	Sim
13	9.316,00	6	25	Sim
14	6.822,40	9	18	Não
15	6.570,90	4	19	Não

Criando a Dummy



Núm. Funcionário	Salario	Anos de Educação Superior	Tempo na Empresa	Possui Skill de Dados?	Dummy
1	5.517,40	3	3	Não	0
2	6.399,90	4	6	Não	0
3	6.206,70	6	3	Não	0
4	6.060,60	4	5	Não	0
5	6.122,70	2	9	Não	0
6	6.955,00	5	9	Não	0
7	7.643,00	4	6	Sim	1
8	6.210,20	2	8	Não	0
9	5.761,00	9	15	Não	0
10	8.086,90	6	14	Sim	1
11	6.375,40	4	9	Não	0
12	9.568,80	6	20	Sim	1
13	9.316,00	6	25	Sim	1
14	6.822,40	9	18	Não	0
15	6.570,90	4	19	Não	0

Regressão Linear

Lidando com variáveis Dummy

Preditiva.ai

Mas o que fazemos quando temos muitas categorias em uma variável? Basta criar mais dummies, uma para cada categoria. Veja alguns exemplos:

	Dummies	
Escolaridade	**D_Medio**	**D_Graduacao**
Ensino Medio	1	0
Graduação	0	1
Pós Graduação	0	0

	Dummies		
Produto	**D_PC**	**D_Celular**	**D_TV**
PC	1	0	0
Celular	0	1	0
TV	0	0	1
Outros	0	0	0

A **quantidade de Dummies a serem criadas é sempre N-1**, sendo

N a quantidade de categorias da variável qualitativa.

Regressão Linear

Lidando com variáveis Dummy

Preditiva.ai

Hands on

Ajuste um modelo que usa uma variável Dummy e interprete seus coeficientes

Roteiro:

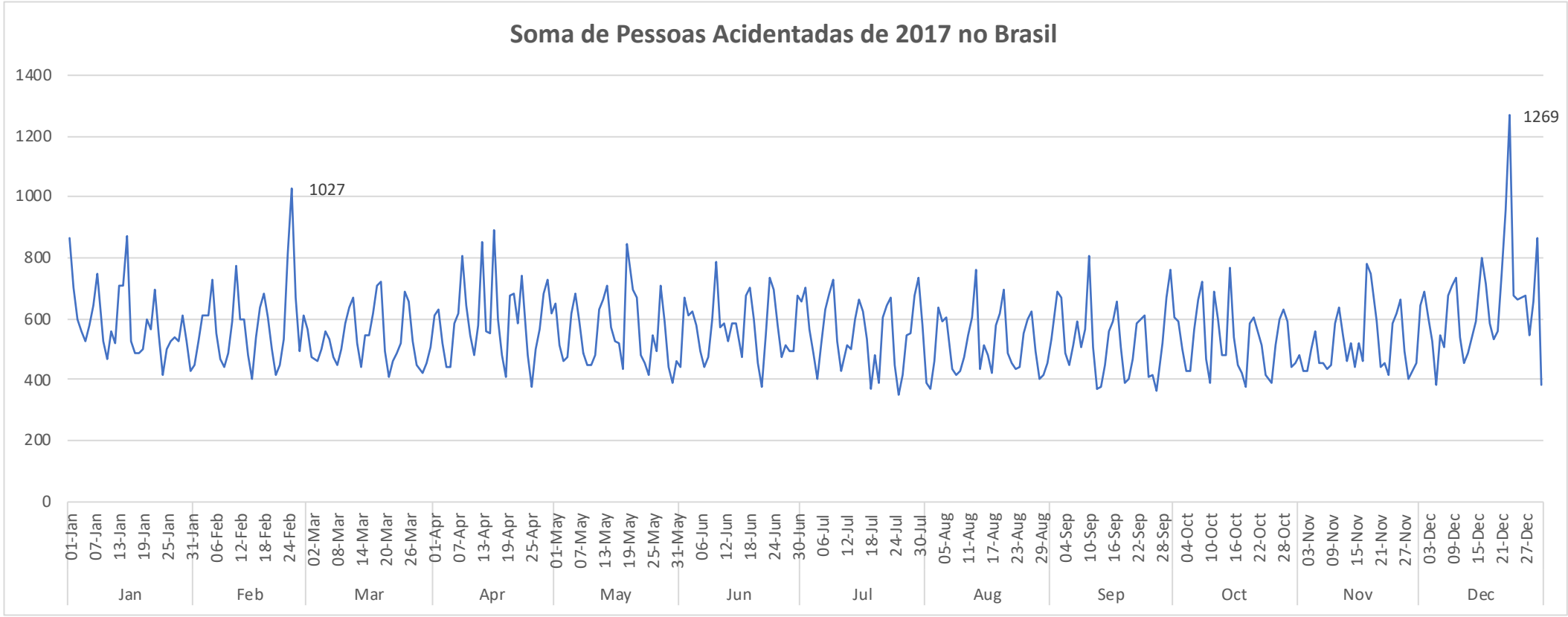
1. Importe a base da aba “Dados3” do arquivo “**base_regressao_salarios.xlsx**”.
2. Monte uma Matriz de Correlação entre as variáveis explicativas numéricas;
3. Ajuste uma regressão linear de Salários usado as variáveis disponíveis;
4. Construa o melhor modelo que conseguir;
5. Interprete o modelo.

Noções de Regressão Linear

Outras aplicações: Séries Temporais

Preditiva.ai

E quando usamos variáveis explicativas como sendo momentos anteriores no tempo da mesma variável? Neste caso, estamos usando a Regressão Linear para construir **Séries Temporais**. Veja um exemplo:



Fonte: https://portal.prf.gov.br/dados-abertos-acidentes

Preditiva.ai

Demonstração

(Séries Temporais no Excel)

Regressão Linear
Séries Temporais

Preditiva.ai

Hands on

Ajuste um modelo que usa uma variável Dummy e interprete seus coeficientes

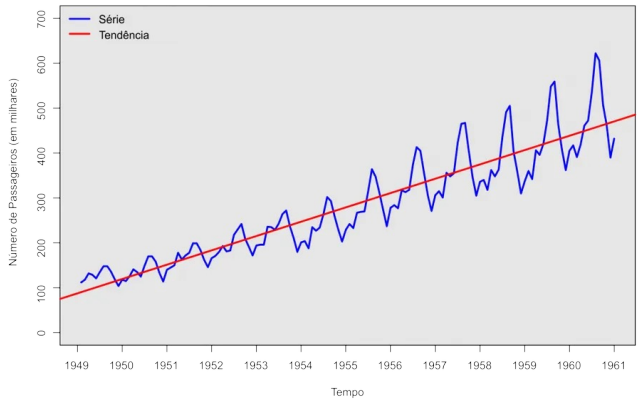
Roteiro:

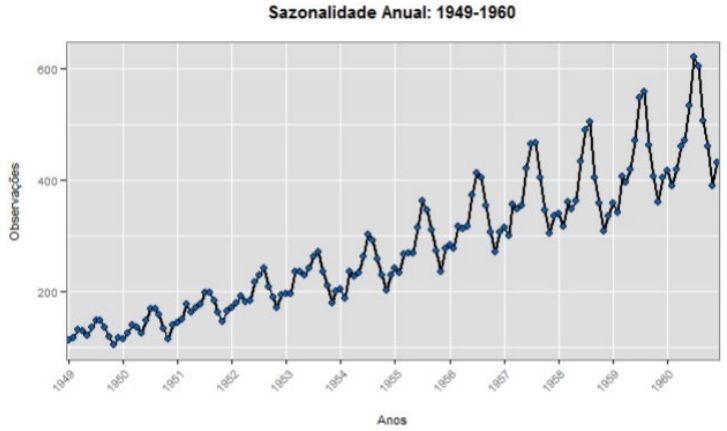
1. Importe "**Base_Acidentes_Datatran2017**".
2. Ajuste uma regressão linear para uma Série Temporal de 14 períodos de dias.
3. Interprete o modelo.
4. A previsão foi melhor que a de 3 períodos?

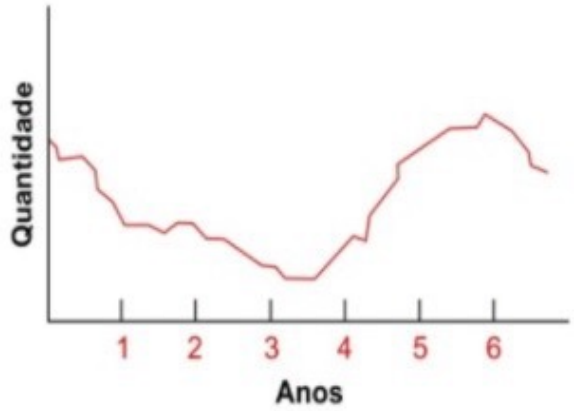
Noções de Regressão Linear

Outras aplicações: Séries Temporais

O assunto de Séries Temporais é extenso e envolve vários conceitos. Entre eles:







Tendência: Significa saber se uma determinada série está crescendo, diminuindo ou se está estável.

Sazonalidade: Variações periódicas, ou seja, fenômenos que se repetem a cada período idêntico de tempo.

Ciclos: É um aumento ou redução da frequência, mas sem intervalos fixos, o que difere da sazonalidade por não ter um intervalo frequente.

Outro conceito importante é a propriedade da **Estacionariedade**. Uma série estacionária é quando a média, variância dos períodos anteriores se mantém constantes durante o tempo. Os modelos de Séries Temporais funcionam muito melhor em series estacionárias.

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 linar. 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