



Data Science  
Academy

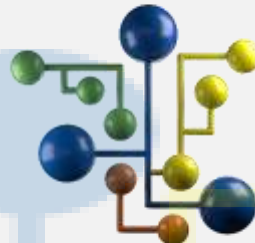
Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b

# Machine Learning



Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b



**Data Science  
Academy**

Seja muito bem-vindo(a)!



Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b



**Data Science  
Academy**

# Machine Learning - Regressão



# Regressão

- Modelos de Aprendizagem
- Fundamentos Teóricos
- Avaliação, Otimização, Regularização, Customizações
- Outros Conceitos Relacionados à Criação do Modelo
- Prática





Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b



**Data Science  
Academy**

O que é Regressão?



## Aprovação de Crédito de um Indivíduo

Atributo	Valor
Sexo	Masculino
Idade	34
Salário Mensal	R\$ 18.000,00
Anos no Emprego Atual	3
Anos de Residência	7
Saldo Bancário	R\$ 32.671,94

### Classificação

- Decisão de crédito (Sim/Não)

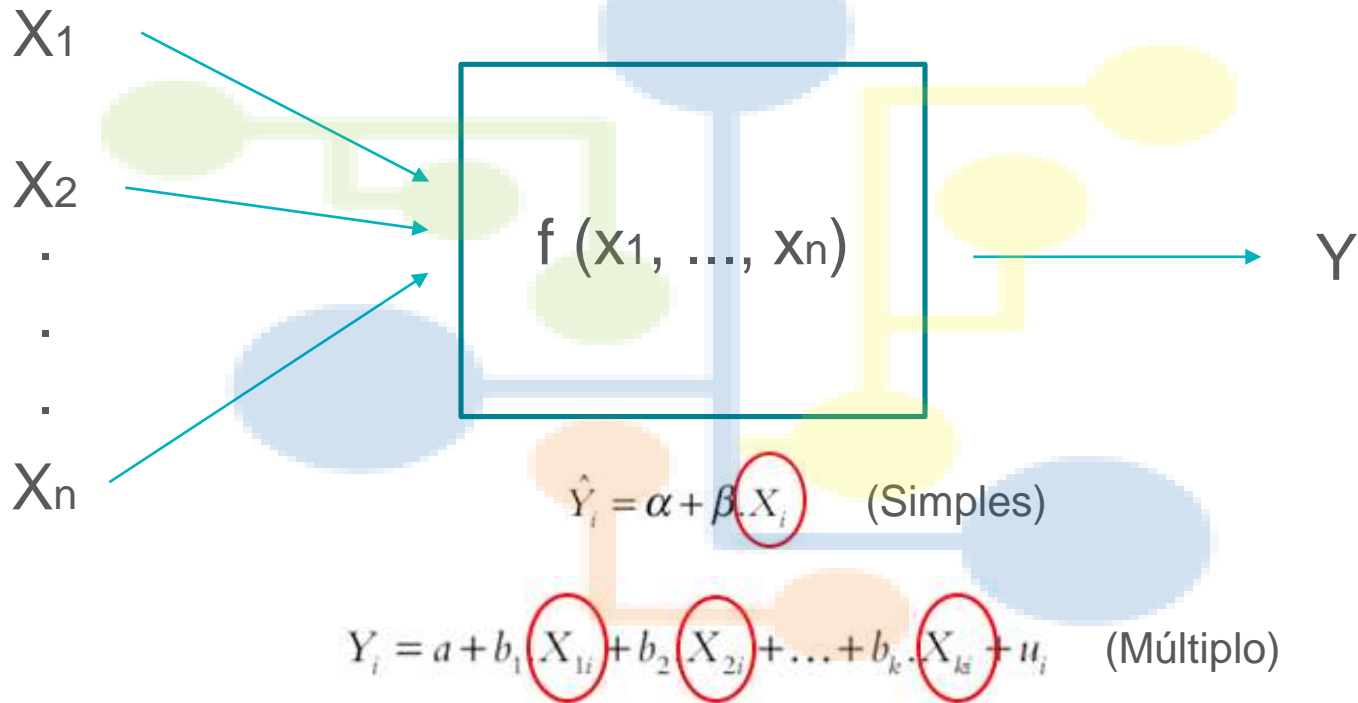
### Regressão

- Quantidade de crédito (dinheiro)





## Modelos de Regressão





Uma **variável independente x**, explica a variação em outra variável, que é chamada **variável dependente y**.  
Este relacionamento existe em apenas uma direção:

**variável independente (x) → variável dependente (y)**

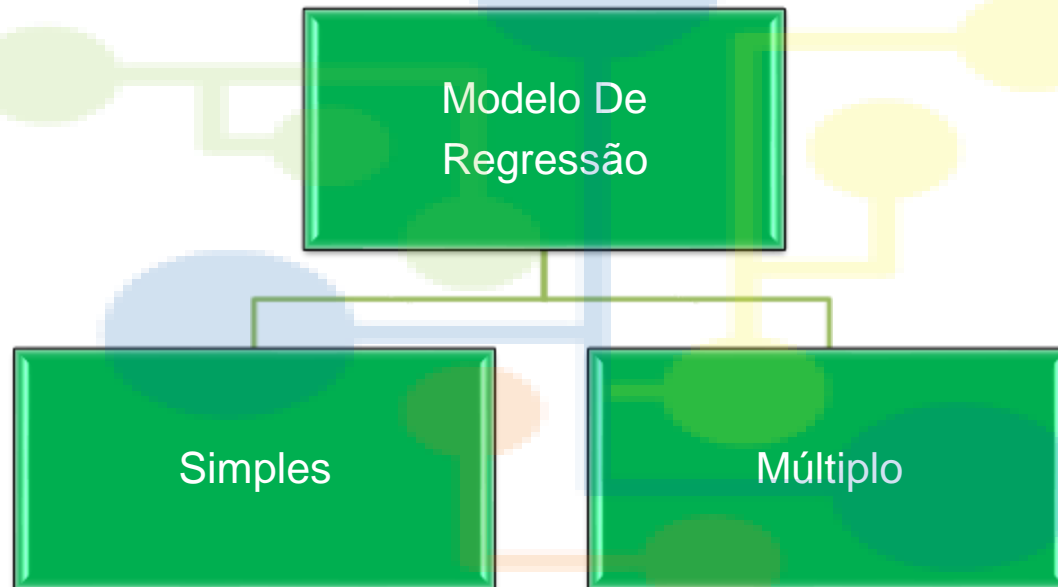




**Análise de regressão** é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas de tal forma que uma variável possa ser predita a partir de outra.



## Tipos de Modelos de Regressão Linear



**1 Variável Dependente Y**  
**1 Variável Independente X**

**1 Variável Dependente Y**  
**2 ou + Variáveis Independentes  $X, X_i$**



A análise de regressão compreende quatro tipos básicos de modelos:

Linear Simples

Linear Múltiplo

Não Linear  
Simples

Não Linear  
Múltiplo

A decorative background diagram consisting of a central blue circle with several lines radiating outwards to other circles of different colors (green, yellow, orange, and blue).

Regressão  
Linear Simples

Regressão  
Linear Múltipla

Regressão  
Logística



Qual o objetivo em se determinar a relação entre duas variáveis?



Prever a população futura de uma cidade simulando a tendência de crescimento da população no passado



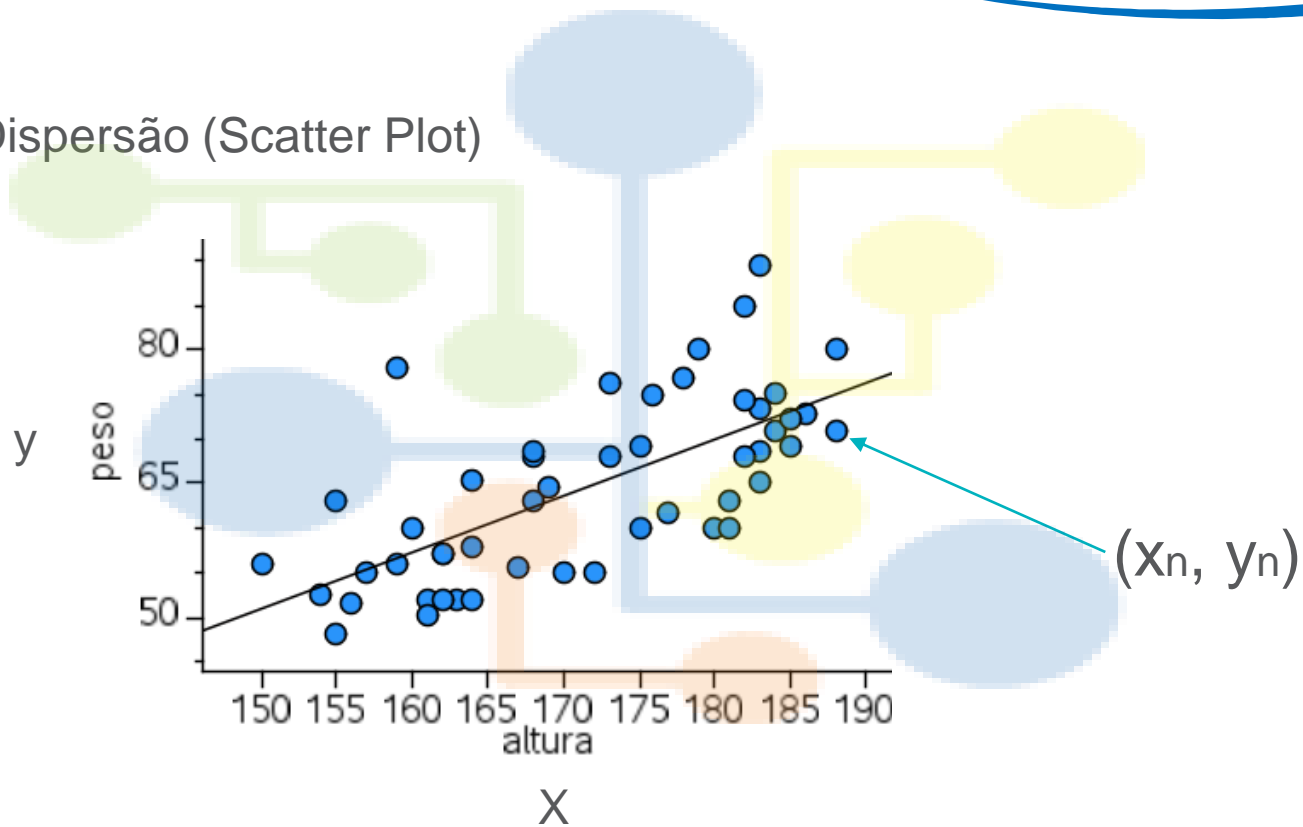
Qual o objetivo em se determinar a relação entre duas variáveis?



Produtividade (Y) de uma área agrícola é alterada quando se aplica certa quantidade (X) de fertilizante sobre a terra



## Diagrama de Dispersão (Scatter Plot)

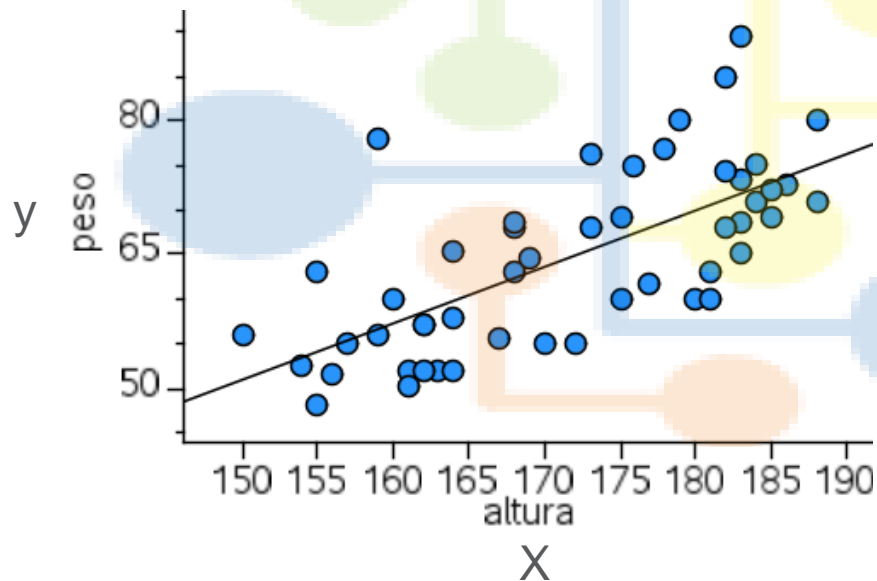






## Modelo de Regressão

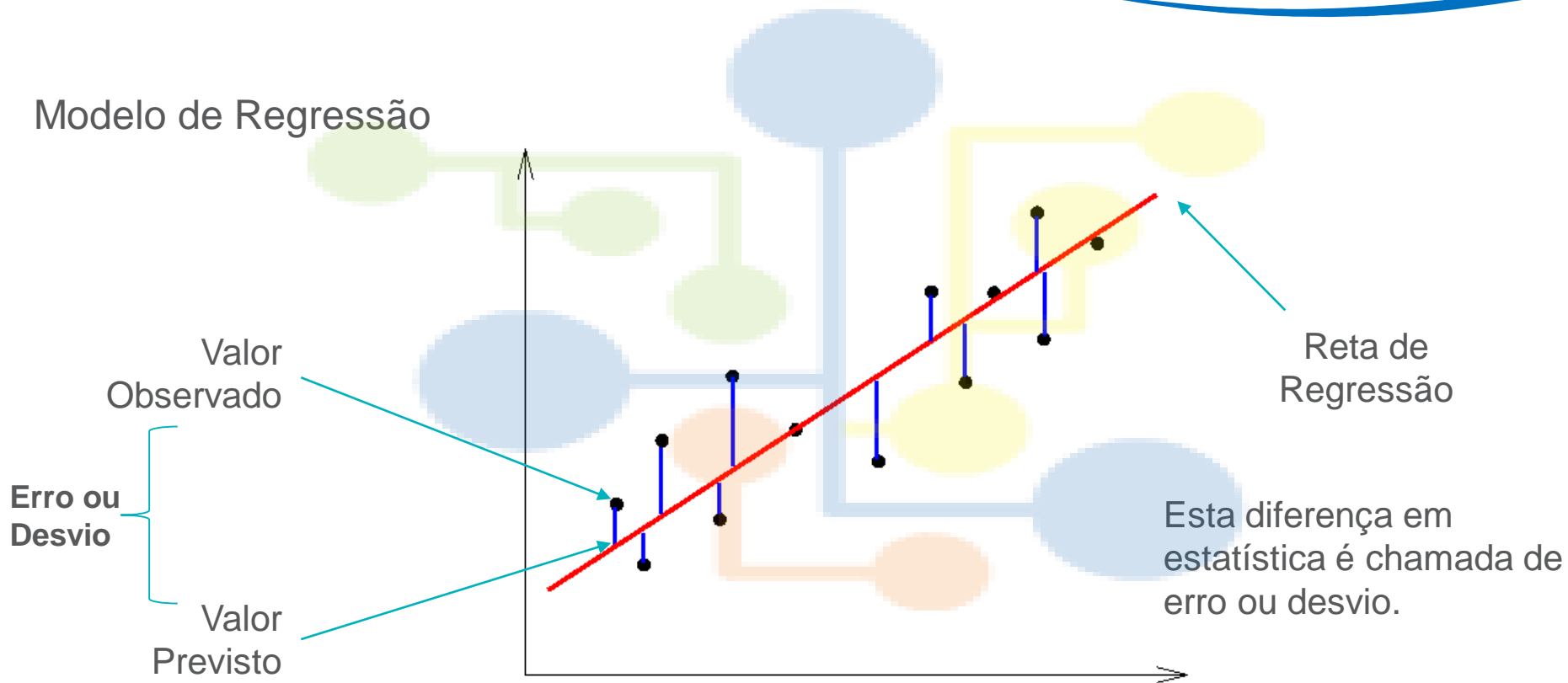
$$\hat{y} = a + bx$$



Reta de  
Regressão

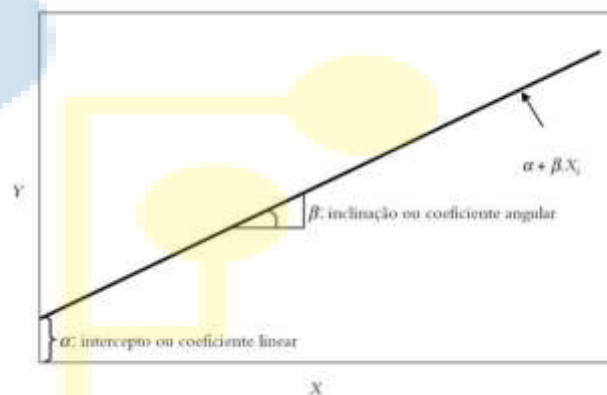


## Modelo de Regressão





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



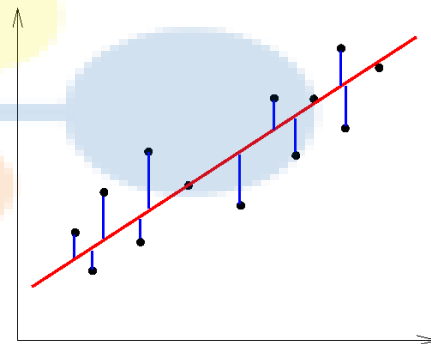
Onde:

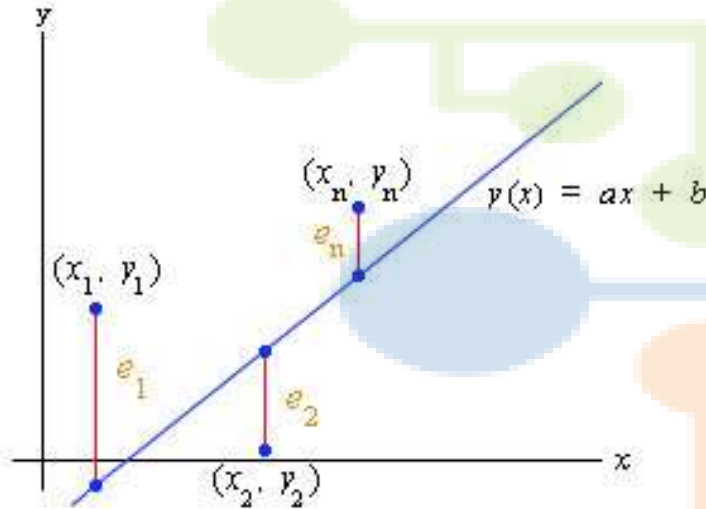
$\hat{y}$  = valor previsto de  $y$  dado um valor para  $x$

$x$  = variável independente

$a$  = ponto onde a linha intercepta o eixo  $y$

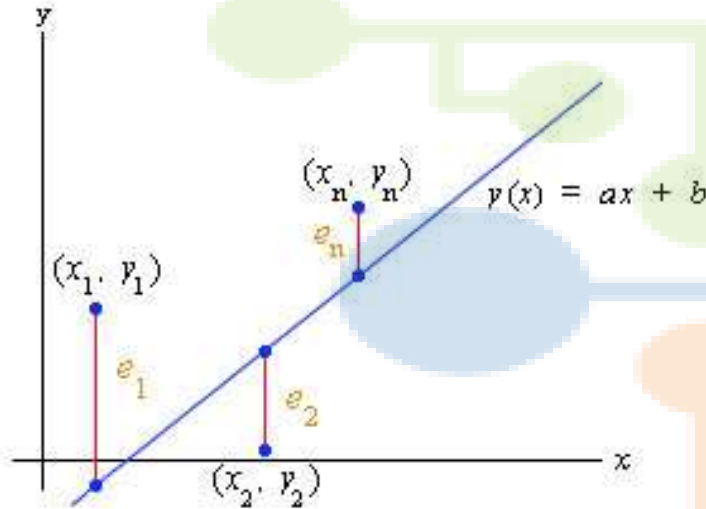
$b$  = inclinação da linha reta





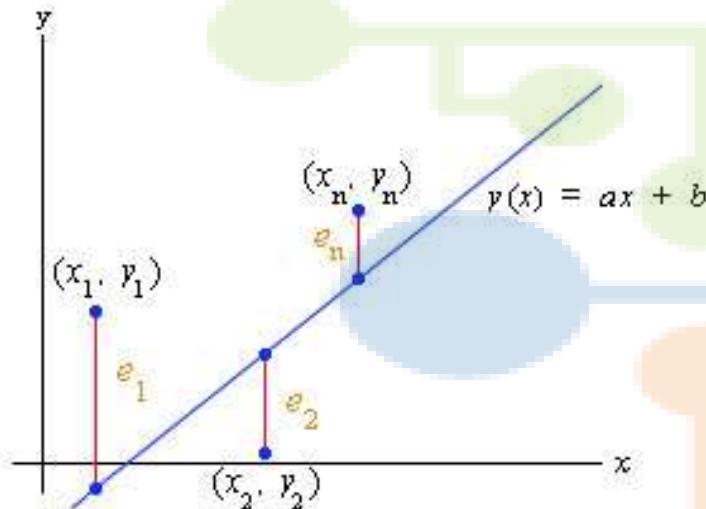
## Método dos Mínimos Quadrados

Esse método definirá uma reta que minimizará a soma das distâncias ao quadrado entre os pontos plotados ( $X, Y$ ) e a reta (que são os valores previstos de  $X', Y'$ ).



## Método dos Mínimos Quadrados

- Erro de Estimativa
- Coeficiente de Determinação

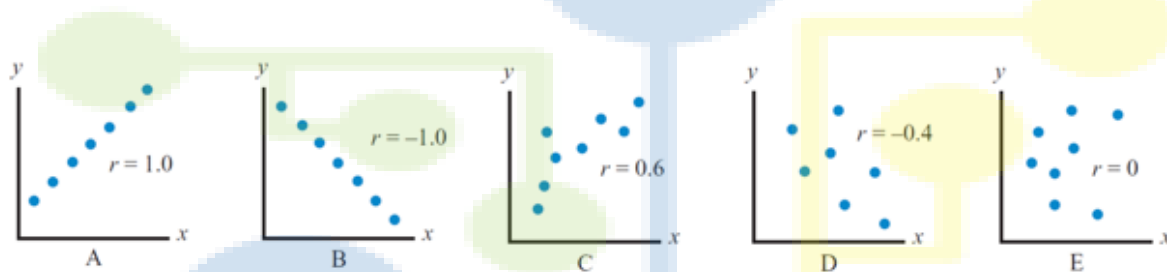


## Método dos Mínimos Quadrados

- Erro de Estimativa
- Coeficiente de Determinação



# Coeficiente de Correlação



**Gráfico A ( $r = 1.0$ ):** correlação positiva perfeita entre  $x$  e  $y$

**Gráfico B ( $r = -1.0$ ):** correlação negativa perfeita entre  $x$  e  $y$

**Gráfico C ( $r = 0.6$ ):** relação positiva moderada:  $y$  tende a aumentar se  $x$  aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

**Gráfico D ( $r = -0.4$ ):** relação negativa fraca: o coeficiente de correlação é próximo de zero ou negativo:  $y$  tende a diminuir se  $x$  aumenta

**Gráfico E ( $r = 0$ ):** Sem relação entre  $x$  e  $y$

Os valores de  $r$  variam entre **-1.0** (uma forte relação negativa) até **+1.0**, uma forte relação positiva.





## Coeficiente de Correlação

O coeficiente de determinação indica o quanto a reta de regressão explica o ajuste da reta, enquanto que o coeficiente de correlação deve ser usado como uma medida de força da relação entre as variáveis



- Soma Total dos Quadrados (STQ) – Mostra a variação em Y em torno da própria média.
- Soma dos Quadrados de Regressão (SQR) – Oferece a variação de Y considerando as variáveis X utilizadas no modelo.
- Soma dos Quadrados dos Resíduos (SQU) – Variação de Y que não é explicada pelo modelo elaborado.

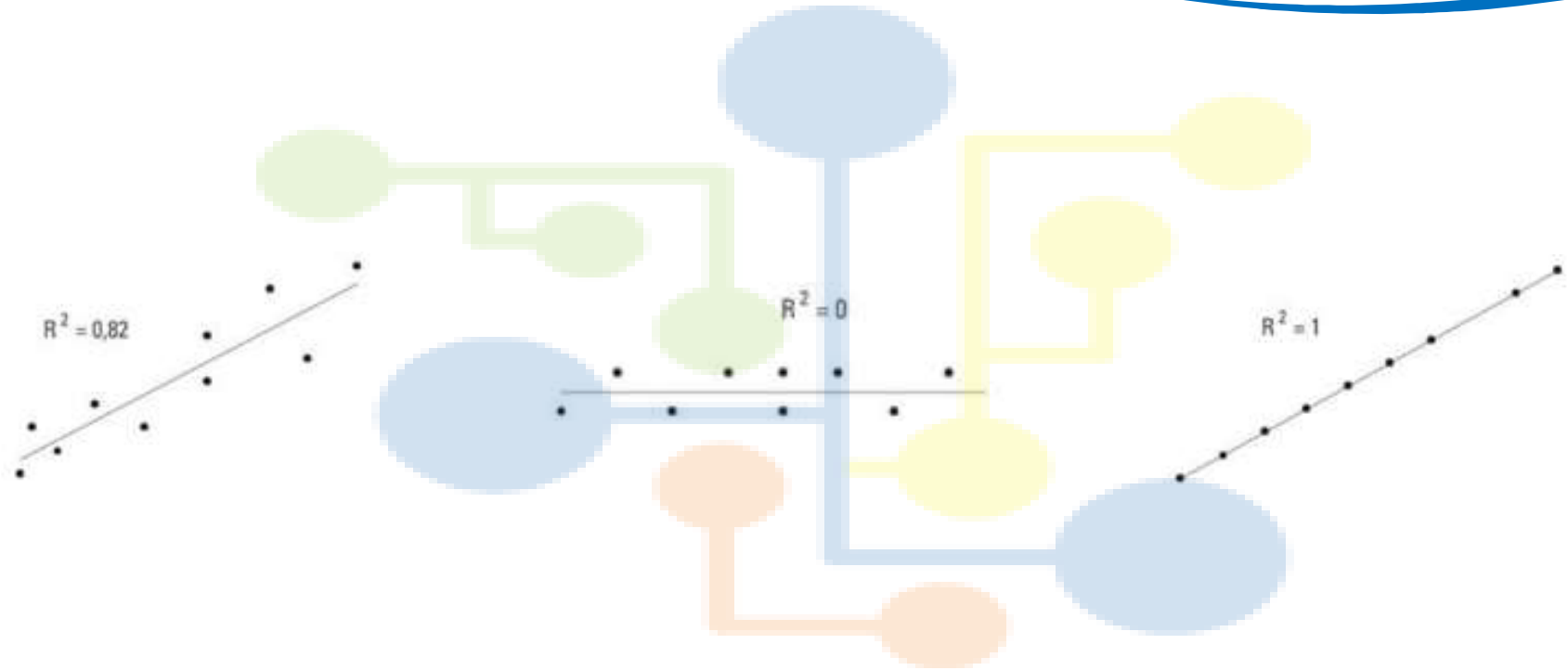
$$STQ = SQR + SQU$$



Nossa próxima etapa é compreender o poder explicativo do modelo de regressão

Coeficiente de Ajuste  $R^2$

$$R^2 = \frac{SQR}{SQR + SQU} = \frac{SQR}{SQT}$$





O coeficiente de ajuste  $R^2$  não diz aos analistas se uma determinada variável explicativa é estatisticamente significativa e se esta variável é a causa verdadeira da alteração de comportamento da variável dependente.



Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b



Data Science Academy

## Avaliando o Modelo de Regressão





## Típicos problemas que podem ser resolvidos com Regressão

- Quantos computadores serão vendidos no próximo mês?
- Quantas pessoas vão acessar nosso web site na próxima semana?
- Qual o salário de uma pessoa de acordo com a performance escolar?
- Qual o total de vendas relacionado ao número de seguidores em redes sociais?







Número de Funcionários Por Turno	Número de Seguidores nas Redes Sociais	Preço da Matéria-Prima (R\$)	Cotação do Dólar	Total de Vendas (R\$)
1400	54000	5000	3.44	1245900
1359	55000	5400	3.12	1302763
1402	55430	5300	3.50	1345119

Atributos ou Features  
(X)

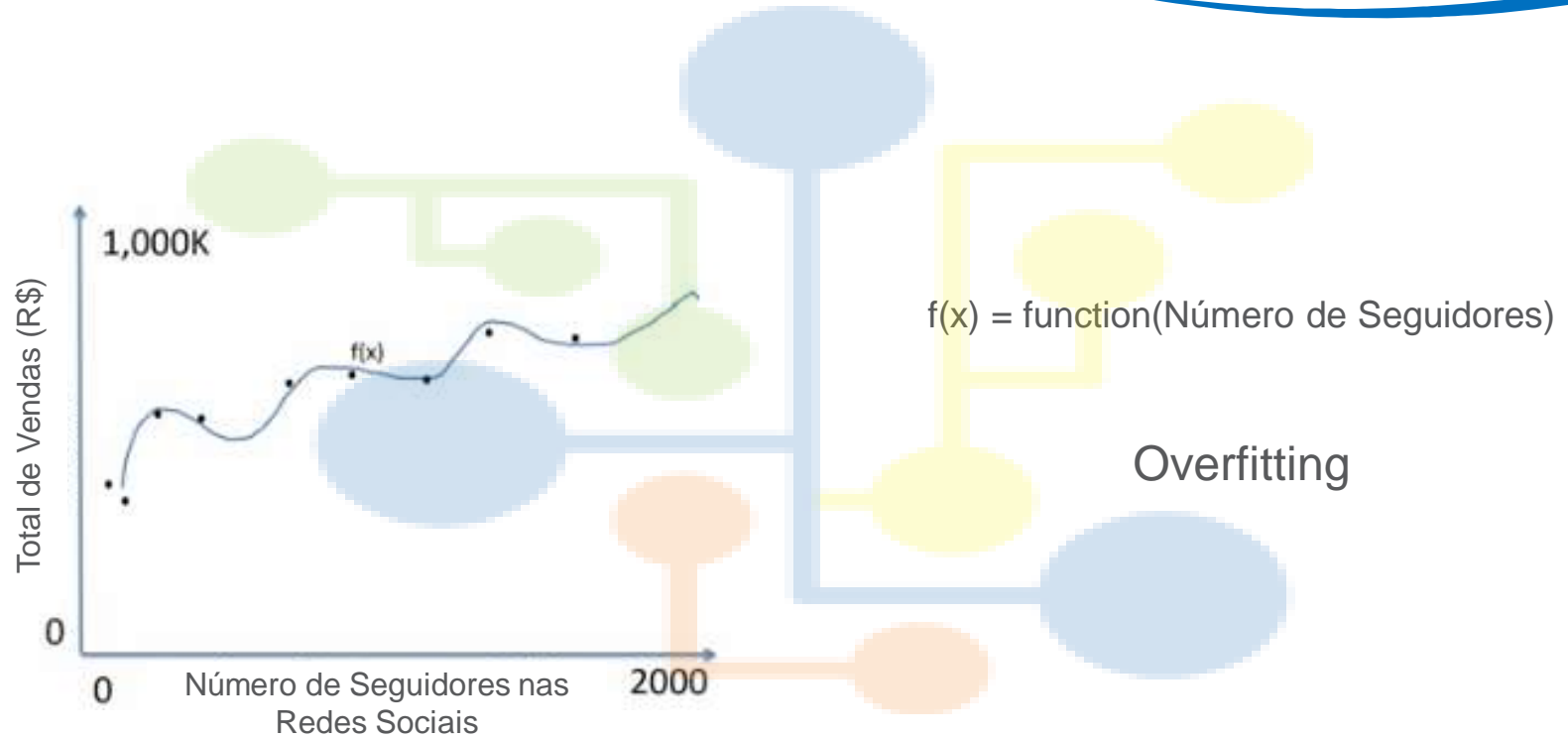
Variável Resposta  
(y)

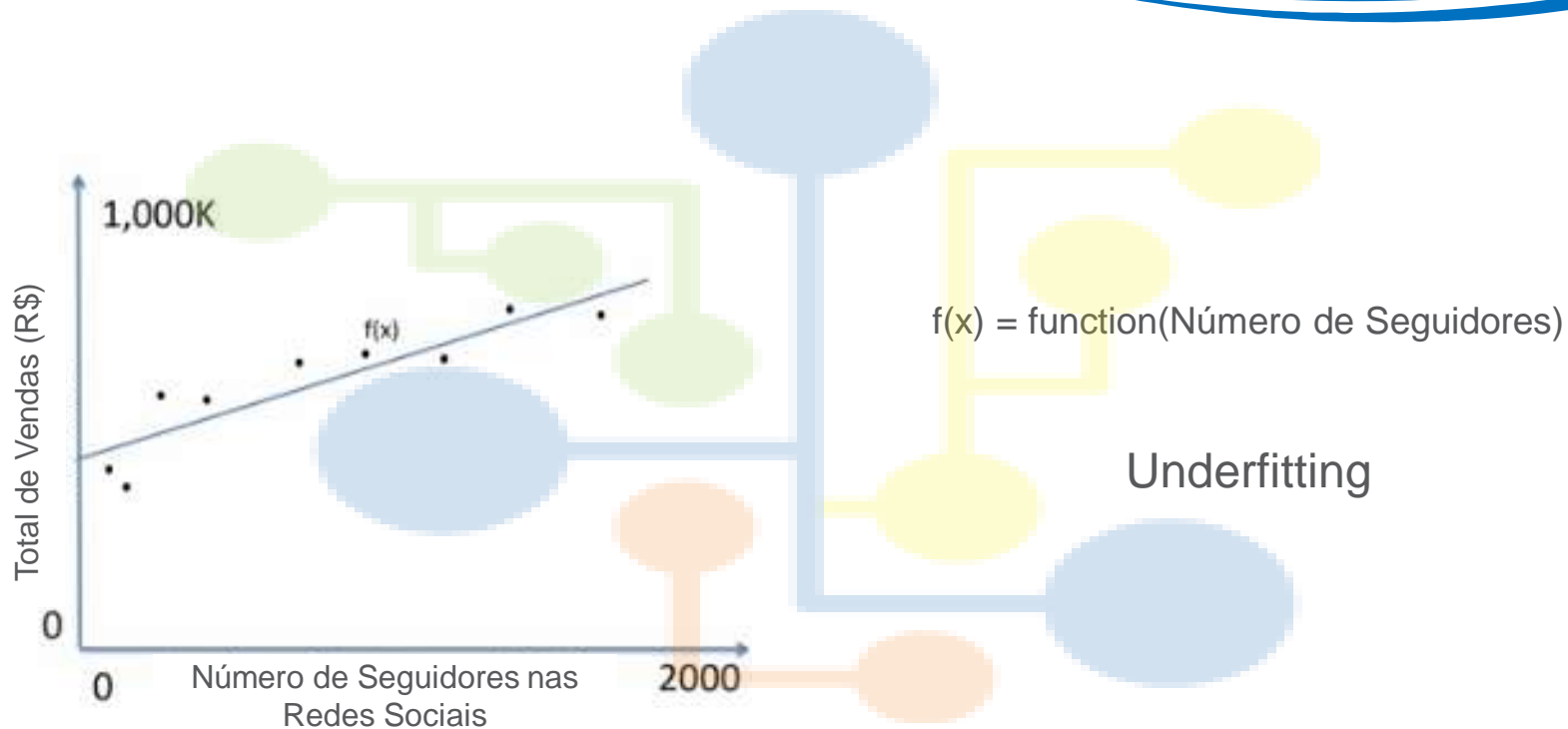


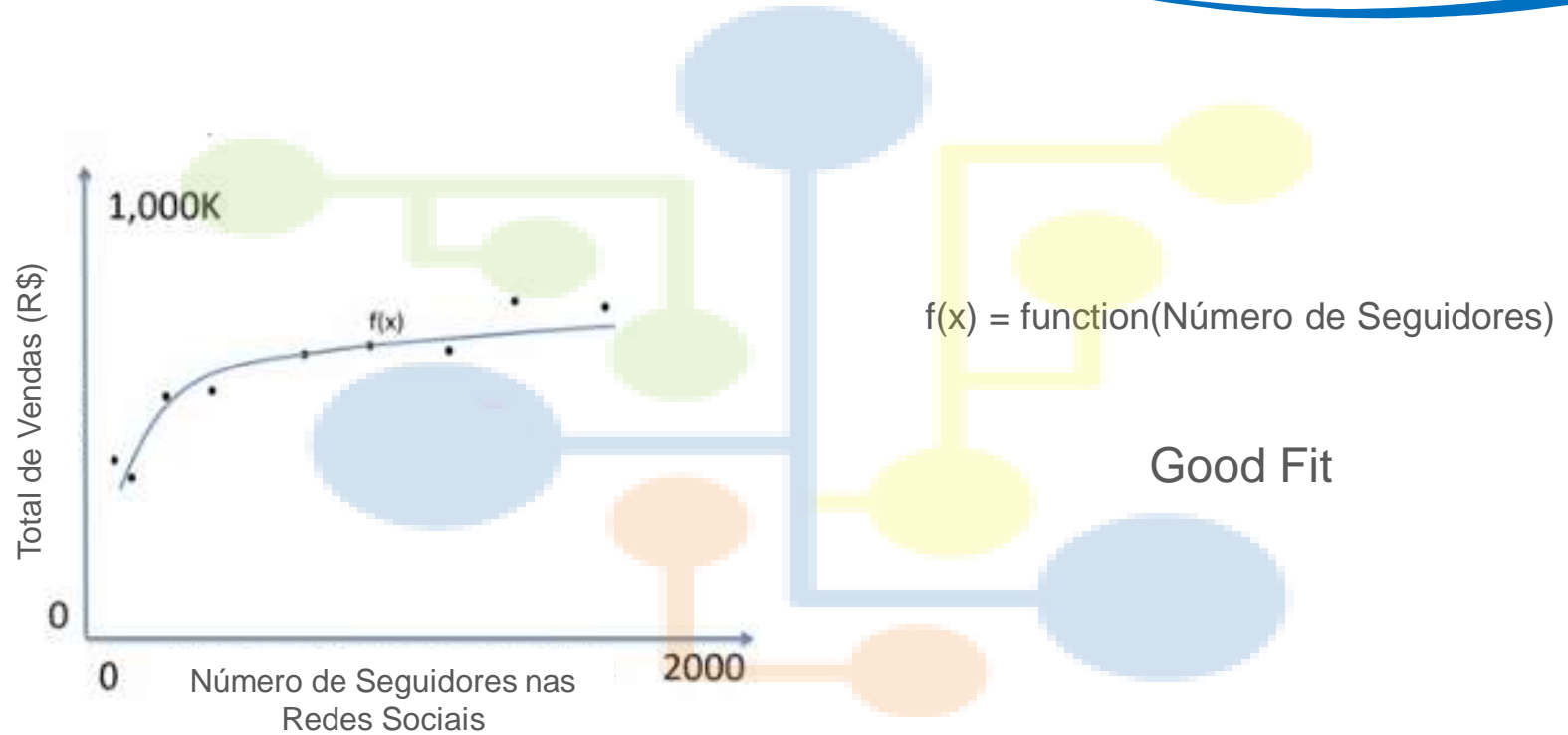
	Número de Seguidores nas Redes Sociais	Total de Vendas (R\$)
	54000	1245900
	55000	1302763
	55430	1345119

Atributo  
(X)

Variável Resposta  
(y)









	Número de Seguidores nas Redes Sociais		Total de Vendas (R\$)	Total de Vendas Previsto (R\$)
	54000		1245900	1278450
	55000		1302763	1302763
	55430		1345119	1320876

Atributos ou Features (X)

Variável Resposta (y)

Previsão  $f(x)$



$$y_i - f(x_i)$$

$$f(x_i) - y_i$$

$$|f(x_i) - y_i|$$

$$(y_i - f(x_i))^2$$

$$\text{Mean absolute error (MAE)} = \sum_{i=1}^n |f(x_i) - y_i|$$

$$\text{SSE/MSE} = \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^n (y_i - f(x_i))^2}$$

**Método dos Mínimos  
Quadrados  
(Least Square Error)**

Total de Vendas (R\$)	Total de Vendas Previsto (R\$)
1245900	1278450
1302763	1334789
1345119	1320876

Variável  
Resposta  
(y)

Previsão  
f(x)

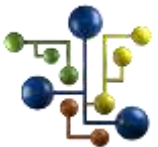
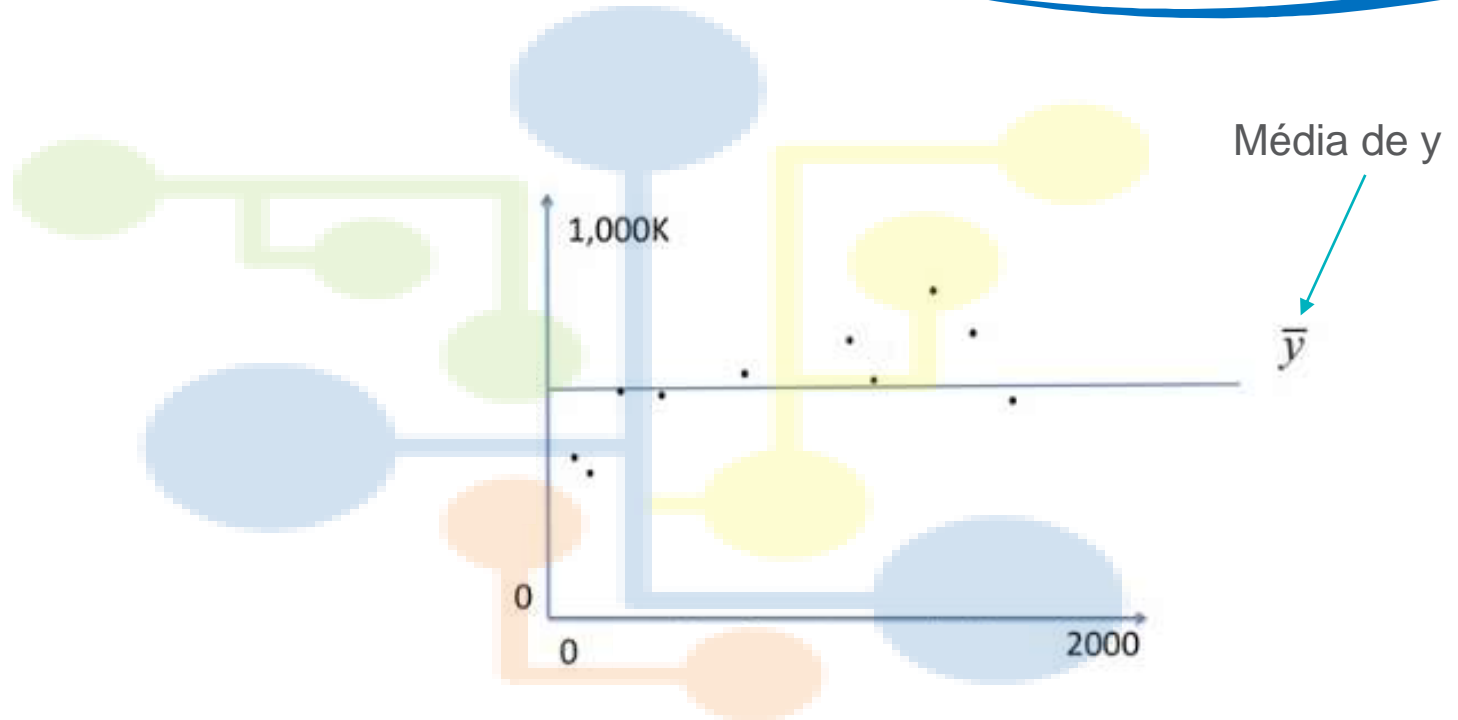




- **(SST – Sum Square Total)** Soma Total dos Quadrados (STQ) – Mostra a variação em Y em torno da própria média.
- **(SSR – Sum Square Regression)** Soma dos Quadrados de Regressão (SQR) – Oferece a variação de Y considerando as variáveis X utilizadas no modelo.
- **(SSE – Sum Square Error)** Soma dos Quadrados dos Resíduos (SQU) – Variação de Y que não é explicada pelo modelo elaborado.

$$SST = SSE + SSR$$





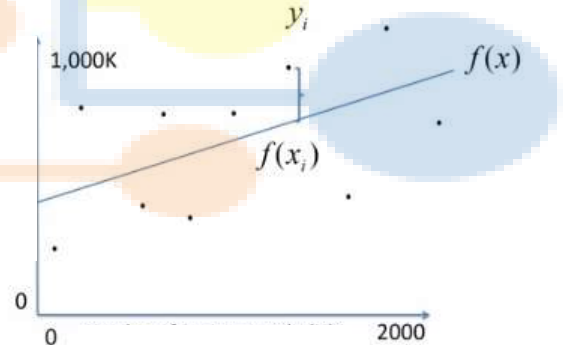
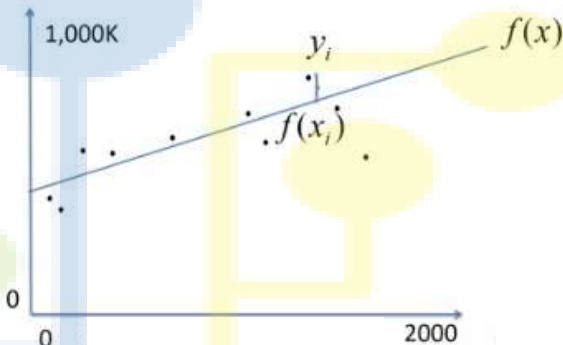


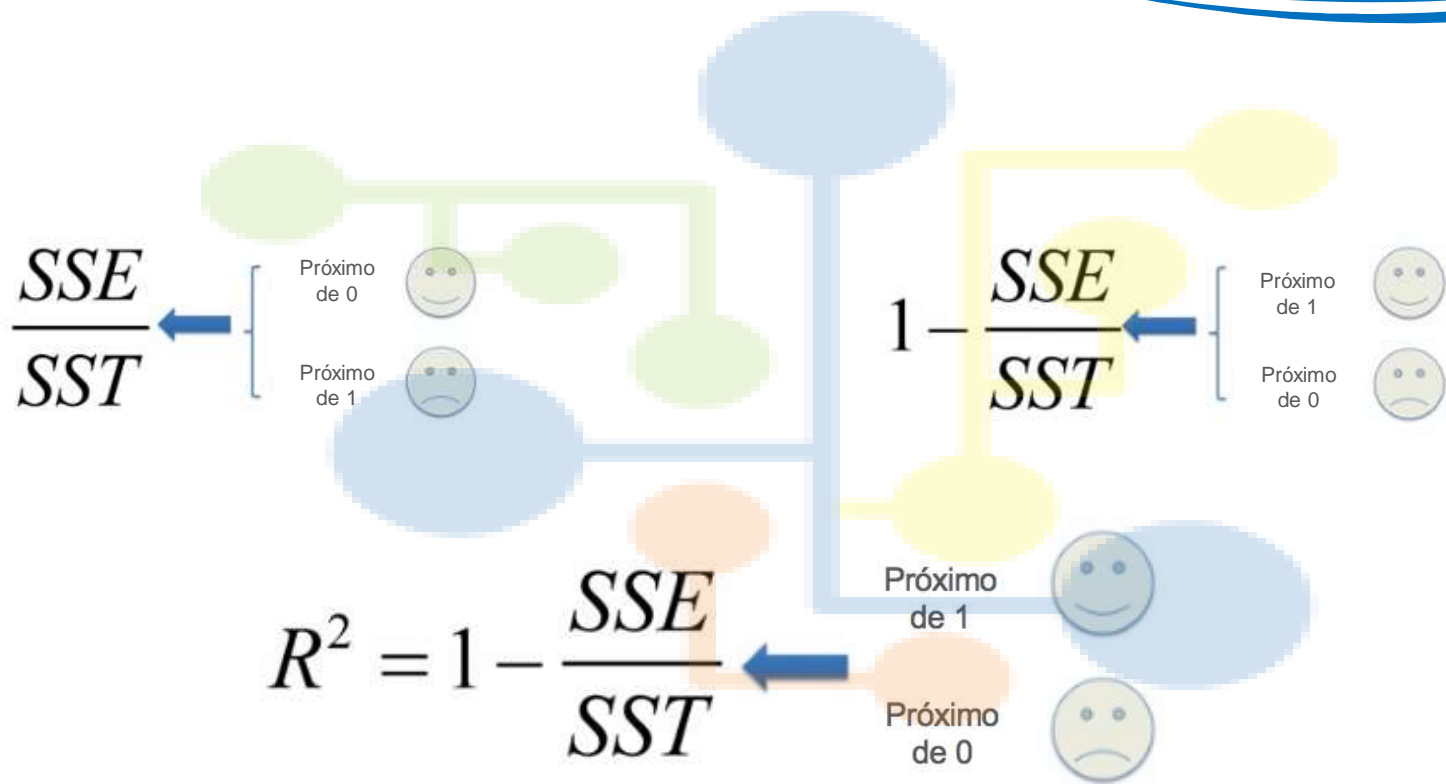
$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^n (f(x_i) - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (f(x_i) - y_i)^2$$

$$SST = SSE + SSR$$







$$SST = SSE + SSR$$

Se o SSR é alto e o SSE é baixo, o Modelo de Regressão explica bem a variação nas previsões

Se o SSR é baixo e o SSE é alto, o Modelo de Regressão não explica bem a variação nas previsões

- **SSR** = medida da variação que pode ser explicada
- **SSE** = medida da variação que não pode ser explicada
- **SST** = medida da variação total





Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b



Data Science Academy

Regressão Linear Simples

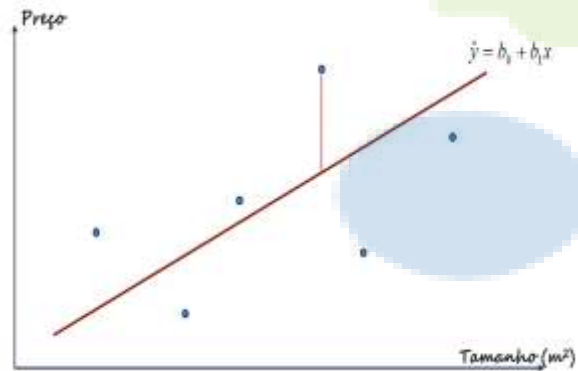
x

Regressão Linear Múltipla





## Regressão Linear Simples



Tamanho (m2)			Preço (R\$)
105			89.000
120			145.000
115			123.000



## Regressão Linear Múltipla

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

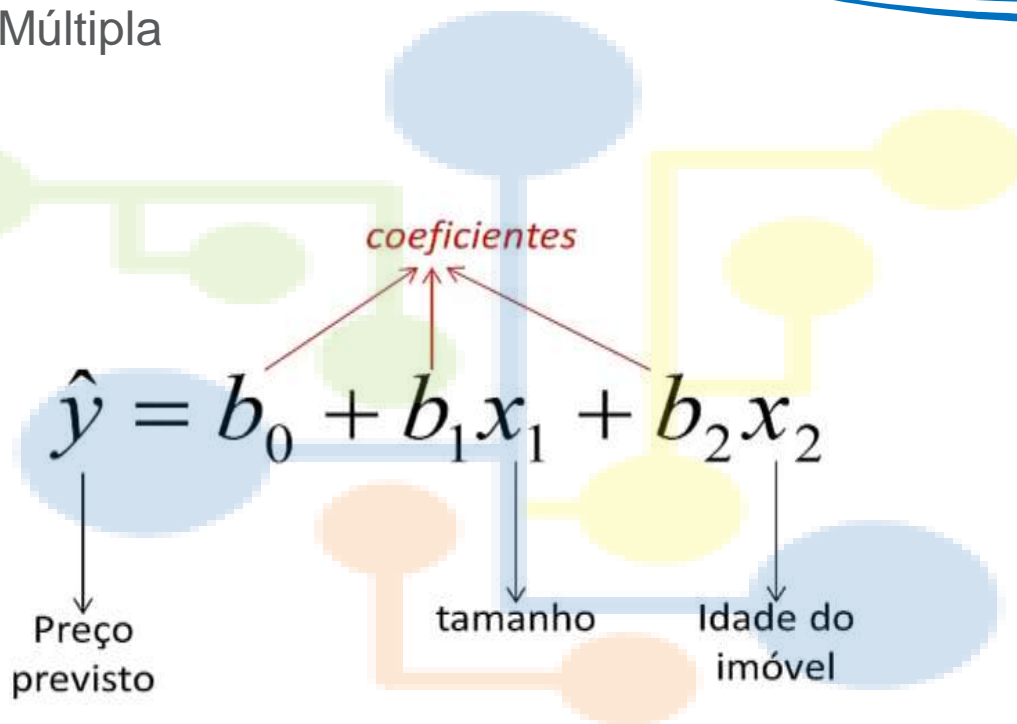
Tamanho (m2)	Idade do Prédio (Anos)	Número Vagas na Garagem	Número de Quartos	Preço (R\$)
105	15	2	2	89.000
120	4	3	3	145.000
115	8	2	3	123.000







## Regressão Linear Múltipla





# Interpretando Modelos de Regressão Linear Simples e Múltipla

- Teste F de Significância Global
- Testes de Significância Individuais
- Coeficientes  $R^2$  e  $R^2$  Ajustado
- Coeficientes





## Teste F de Significância Global

O modelo é útil para prever o preço?

Estatística de regressão	
R múltiplo	0,66
R-Quadrado	0,44
R-quadrado ajustado	0,41
Erro padrão	132352,0
Observações	40

### ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	2	5,135E+11	2,567E+11	1,466E+01	0,000
Resíduo	37	6,481E+11	1,752E+10		
Total	39	1,162E+12			

F de significação: teste F de significância global do modelo.  
"Há evidências de que pelo menos uma variável no modelo está relacionada com o preço?"  
Como **valor-p do teste F < 0,05**, há evidências estatísticas.

Valor-p do teste F

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	440107,0	182742,3	2,408	0,021	69836,0	810378,1
tamanho	6772,1	1555,7	4,353	0,000	3620,0	9924,2
idade do prédio	-19129,7	8372,9	-2,285	0,028	-36094,8	-2164,5



## Testes de Significância Individuais

Quais variáveis estão relacionadas com o preço?

### Estatística de regressão

R múltiplo	0,66
R-Quadrado	0,44
R-quadrado ajustado	0,41
Erro padrão	132352,0
Observações	40

### ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	2	5,135E+11	2,567E+11	1,466E+01	0,000
Resíduo	37	6,481E+11	1,752E+10		
Total	39	1,162E+12			

	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	440107,0	182742,3	2,408	0,021	69836,0	810378,1
tamanho	6772,1	1555,7	4,353	<b>0,000</b>	3620,0	9924,2
idade do prédio	-19129,7	8372,9	-2,285	<b>0,028</b>	-36094,8	-2164,5

Há evidências estatísticas de relação de tamanho e idade com preço, pois **valores-p<0,05**.



## Coeficientes R2 e R2 Ajustado

Qual percentual de variabilidade é explicado pelas variáveis usadas no modelo?

Estatística de regressão						
R múltiplo	0,66					
<b>R-Quadrado</b>	<b>0,44</b>					
R-quadrado ajustado	0,41					
Erro padrão	132352,0					
Observações	40					

**R<sup>2</sup>: 44% da variabilidade do preço é explicada pelo modelo com tamanho e idade.**

**R<sup>2</sup> ajustado: sempre é um pouco menor que o R<sup>2</sup>; deve ser utilizado para comparar modelos com diferentes quantidades de variáveis.**

ANOVA					
	gl	SQ	MQ	F	F de significação
Regressão	2	5,135E+11	2,567E+11	1,466E+01	0,000
Resíduo	37	6,481E+11	1,752E+10		
Total	39	1,162E+12			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	440107,0	182742,3	2,408	0,021	69836,0	810378,1
tamanho	6772,1	1555,7	4,353	0,000	3620,0	9924,2
idade do prédio	-19129,7	8372,9	-2,285	0,028	-36094,8	-2164,5



## Coeficientes

Valores que compõe a equação.

### Estatística de regressão

R múltiplo	0,66
R-Quadrado	0,44
R-quadrado ajustado	0,41
Erro padrão	132352,0
Observações	40

### ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	2	5,135E+11	2,567E+11	1,466E+01	0,000
Resíduo	37	6,481E+11	1,752E+10		
Total	39	1,162E+12			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	440107,0	182742,3	2,408	0,021	69836,0	810378,1
tamanho	6772,1	1555,7	4,353	0,000	3620,0	9924,2
idade do prédio	-19129,7	8372,9	-2,285	0,028	-36094,8	-2164,5

$$\hat{y} = 440107 + 6772,1 \cdot \text{tamanho} - 19129,7 \cdot \text{idade}$$



## Regras Gerais

Modelo é útil para prever o preço, se o valor-p do teste F é menor que 0,05.

O  $R^2$  indica quanto da variabilidade de  $y$  é explicado pelas variáveis preditoras. Pode ser necessário incluir mais variáveis no modelo para aumentar este coeficiente.

Há evidências de que uma variável está relacionada com o valor previsto, se o valor-p for menor que 0,05.

O objetivo da regressão é encontrar os coeficientes que permitem construir a equação de regressão e fazer as previsões.





Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b

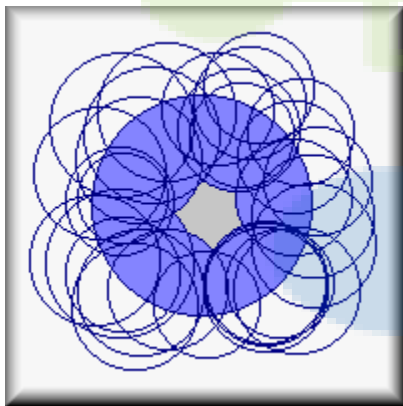


Data Science Academy

Regularização







Isso significa que muitas variáveis seriam ajustadas e o modelo ficaria super estimado, com uma variância infinita, sendo inviável o método dos mínimos quadrados





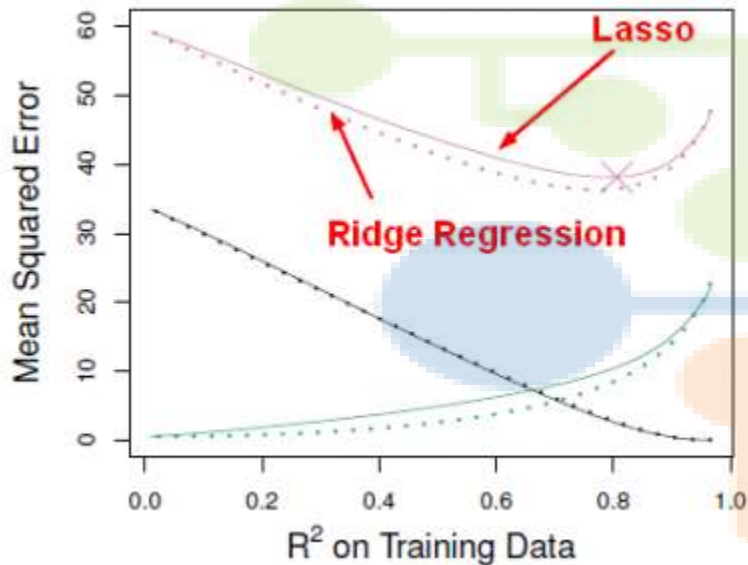
Temos basicamente 3 métodos que nos auxiliam quando o número de variáveis é maior que o número de observações:

Seleção de um subconjunto de coeficientes

Reduzir a dimensão

Reduzir o valor dos coeficientes  
(Regularização)





Uma regressão com diversos coeficientes regressores torna o modelo como um todo muito mais complexo e pode tirar características de interpretabilidade





## Shrinkage Methods (Métodos de Encolhimento)

Ridge Regression

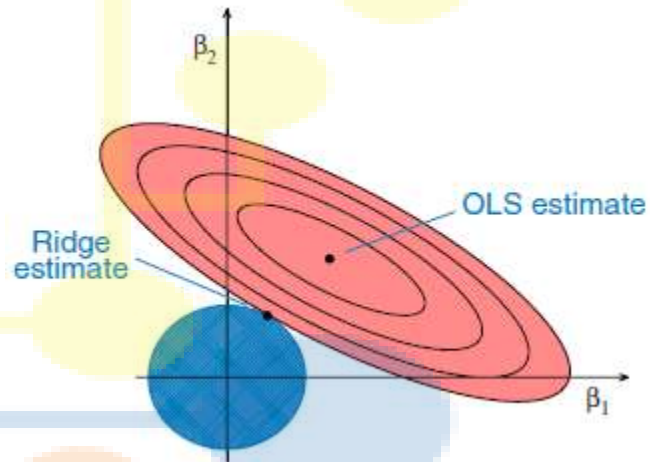
**LASSO Regression**  
(Least Absolute Shrinkage and Selection Operator)





## Ridge Regression

A Ridge Regression é um método de regularização do modelo que tem como principal objetivo suavizar atributos que sejam relacionados uns aos outros e que aumentam o ruído no modelo (multicolinearidade).

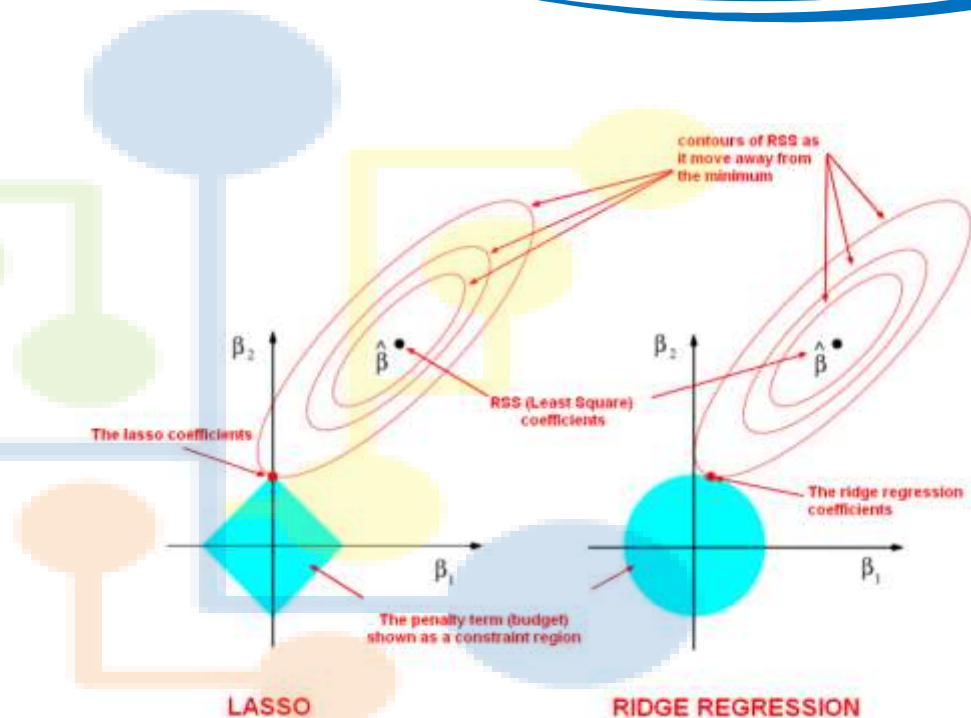


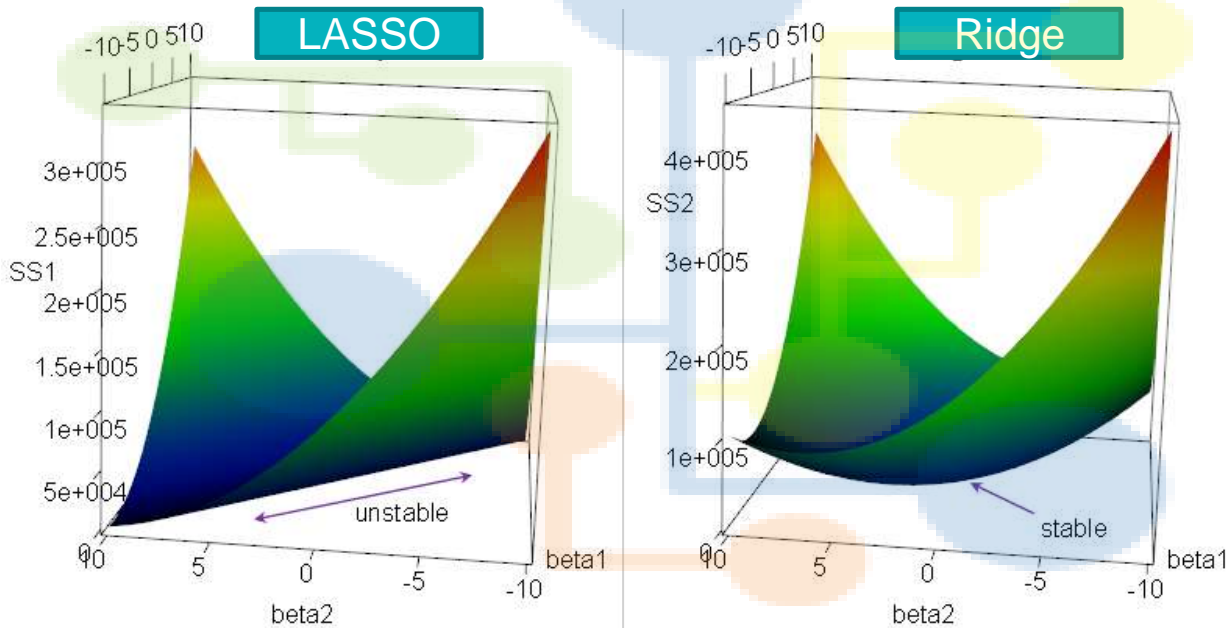


## LASSO Regression

(Least Absolute Shrinkage and Selection Operator)

O LASSO tem o mesmo mecanismo de penalização dos coeficientes com um alto grau de correlação entre si, mas que usa o mecanismo de penalizar os coeficientes de acordo com o seu valor absoluto.







Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b

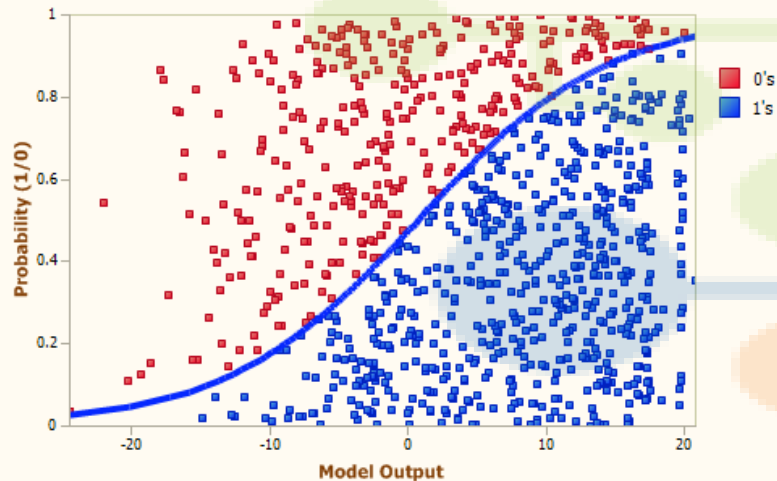


Data Science Academy

Regressão Logística

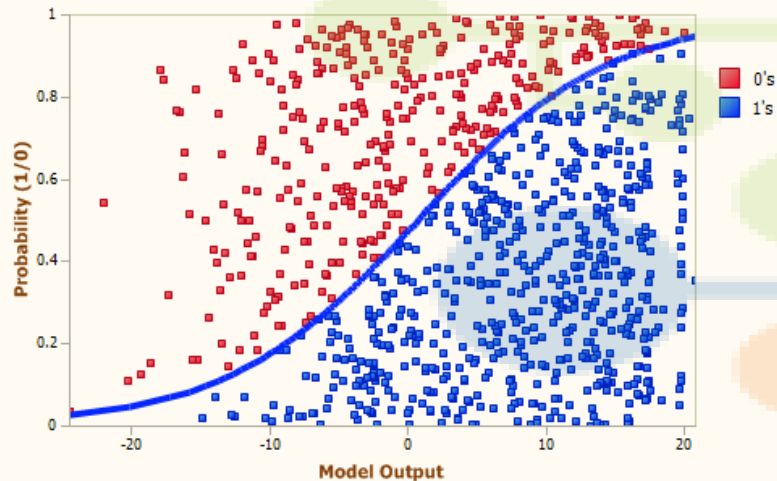






A regressão logística é uma técnica estatística que tem como objetivo modelar, a partir de um conjunto de observações, a relação “logística” entre uma variável resposta e uma série de variáveis explicativas numéricas (contínuas, discretas) e/ou categóricas.





A regressão logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como **modelo logístico**, **modelo logit**, e **classificador de máxima entropia**.





**Na Regressão Logística, a variável resposta é binária**

1 → acontecimento de interesse (sucesso)

0 → acontecimento complementar (insucesso)





$$g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right)$$

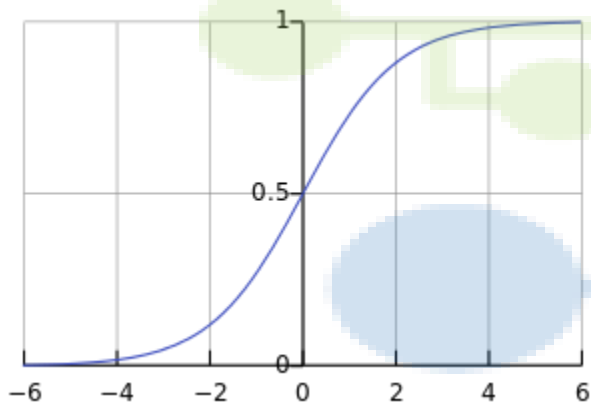
$$g(x) = \ln \left( \frac{\frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}{1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}} \right) = \ln \left( \frac{\frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}{\frac{1}{1 + e^{\beta_0 + \beta_1 x}}} \right)$$

**Transformação logit**

$$g(x) = \ln(e^{\beta_0 + \beta_1 x}) = \beta_0 + \beta_1 x$$

Logaritmo





Regressão Logística é útil para modelar a probabilidade de um evento ocorrer como função de outros fatores. É um modelo linear generalizado que usa como função de ligação a função logit.





A regressão logística é utilizada em áreas tais como:





- Em medicina, permite por exemplo determinar os fatores que caracterizam um grupo de indivíduos doentes em relação a indivíduos saudáveis.
- Na área de seguros, permite encontrar frações de clientes que sejam sensíveis a determinada política securitária em relação a um dado risco particular.
- Em instituições financeiras, pode detectar os grupos de risco para a subscrição de um crédito.
- Em econometria, permite explicar uma variável discreta, como por exemplo as intenções de voto em atos eleitorais.





Data Science  
Academy

Data Science Academy marxv49@gmail.com 5e686b2be32fc3447a0e403b



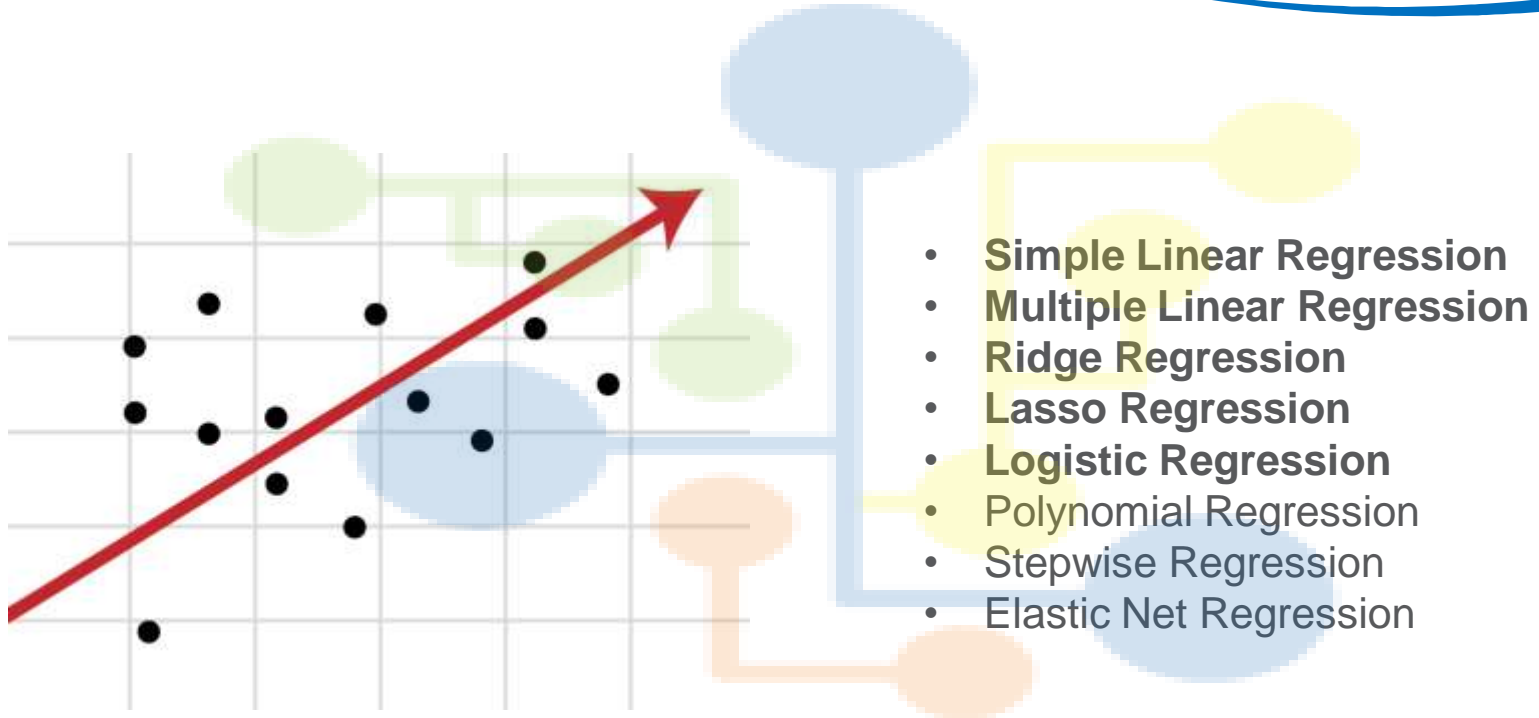
Data Science Academy

# Regressão

## Vantagens e Desvantagens











## Previsão do Futuro





## Apoio à Tomada de Decisão



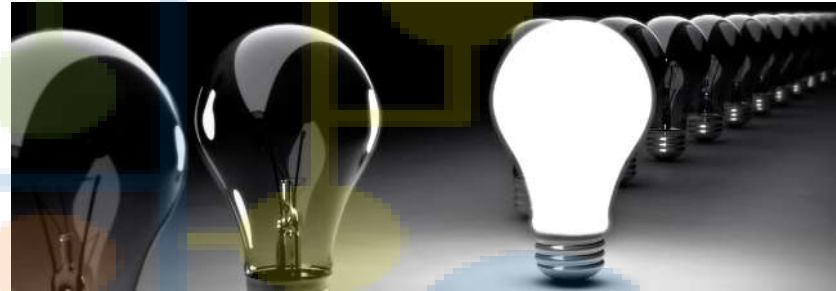


## Correção de Erros





Novos Insights





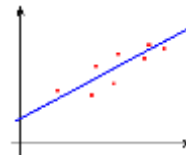
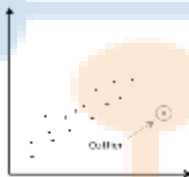
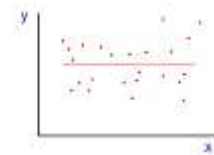
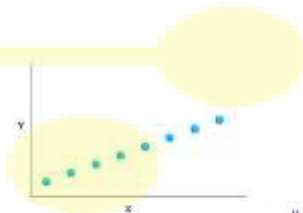
## Importantes Desvantagens:





## Importantes Desvantagens:

- Apenas consideram relacionamento linear
- Toma como base a média da variável dependente
- Sensível a Outliers
- Regressão Linear assume que os dados são independentes







Data Science  
Academy

Data Science Academy [marxv49@gmail.com](mailto:marxv49@gmail.com) 5e686b2be32fc3447a0e403b



Continue Trilhando uma Excelente Jornada de Aprendizagem!

**Muito Obrigado!**