CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Regressão

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com

Regressão Linear

- **Objetivo**
  - Estabelecer uma função matemática que descreva a relação entre uma variável contínua (dependente) e uma ou mais variáveis explicativas ou independentes
- O modelo de **regressão linear simples** se define por uma relação linear entre a variável dependente e uma variável independente
- Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se de **regressão linear múltipla**

Exemplos

- Altura dos pais e altura dos filhos
- Tempo de prática de esportes e ritmo cardíaco
- Tempo de estudo e nota na prova
- Taxa de desemprego e taxa de criminalidade
- Expectativa de vida e taxa de analfabetismo
- Renda familiar e gasto com cartão de crédito
- Gastos com publicidade e preço do produto
- Tamanho e preço de um imóvel
- Número de agrotóxicos liberados e casos de cancer

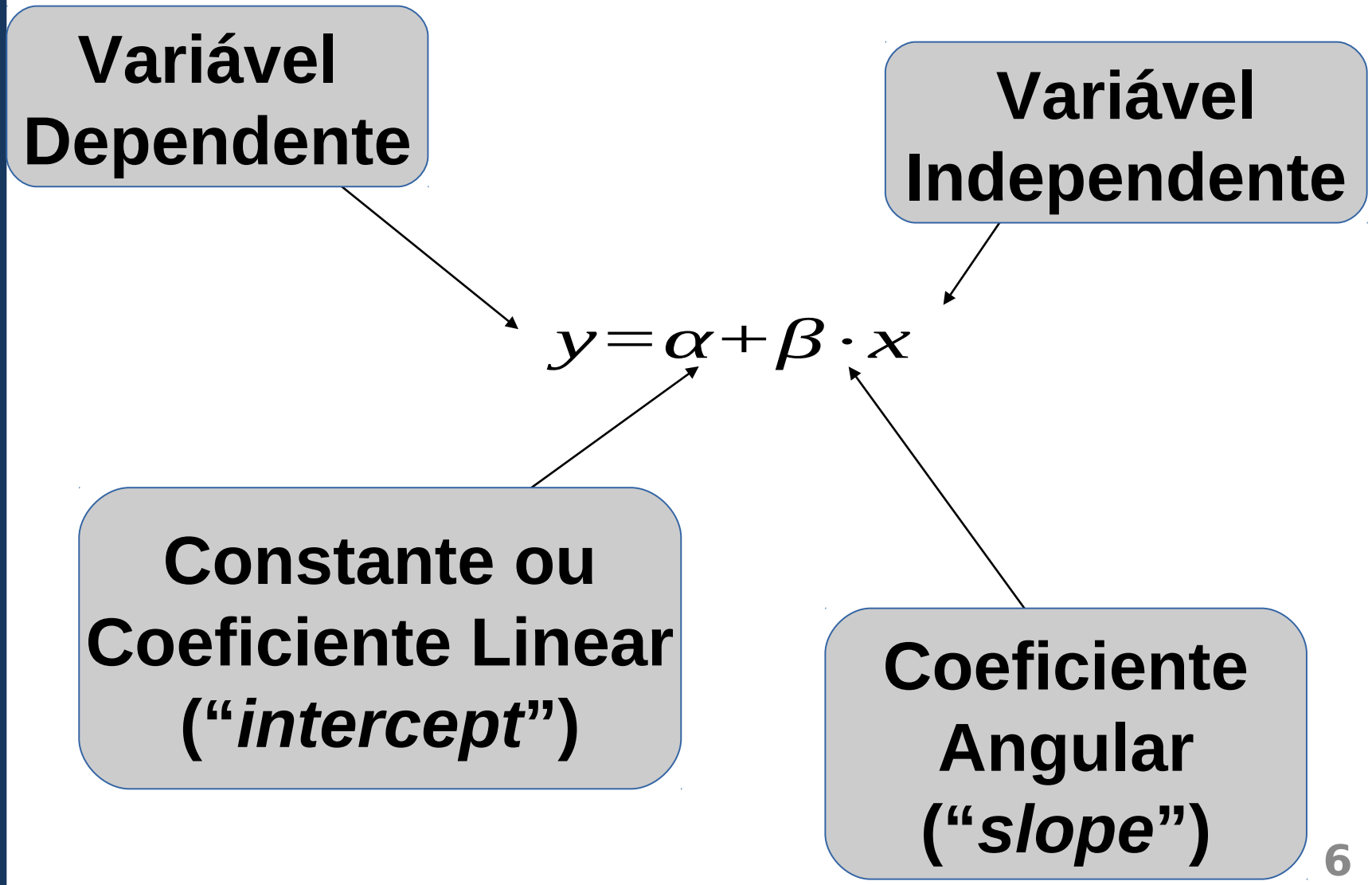
A relação entre as variáveis

- A presença ou ausência de relação linear pode ser investigada sob dois pontos de vista
  - Explicitando a forma dessa relação: **regressão**
  - Quantificando a força dessa relação: **correlação**
- A relação pode entre as variáveis pode ser
  - **Direta** (ou positiva) quando os valores de Y aumentam em decorrência do aumento dos valores de X
  - **Inversa** (ou negativa) quando os valores de Y variam inversamente em relação aos de X

Diagrama de dispersão

- Os dados para a análise de regressão e correlação simples são da forma
  - (x_1,y_1), (x_2,y_2),..., (x_i,y_i),..., (x_n,y_n)
- Com os dados constrói-se um diagrama de dispersão, que deve exibir a tendência
- Este diagrama permite decidir empiricamente
- Se **existe uma relação linear** entre as variáveis X e Y
- Se a **relação linear é direta ou inversa** entre as variáveis, conforme o modo como os pontos se dispersam ao redor da equação da reta obtida através valores dos pontos

Regressão Linear Simples



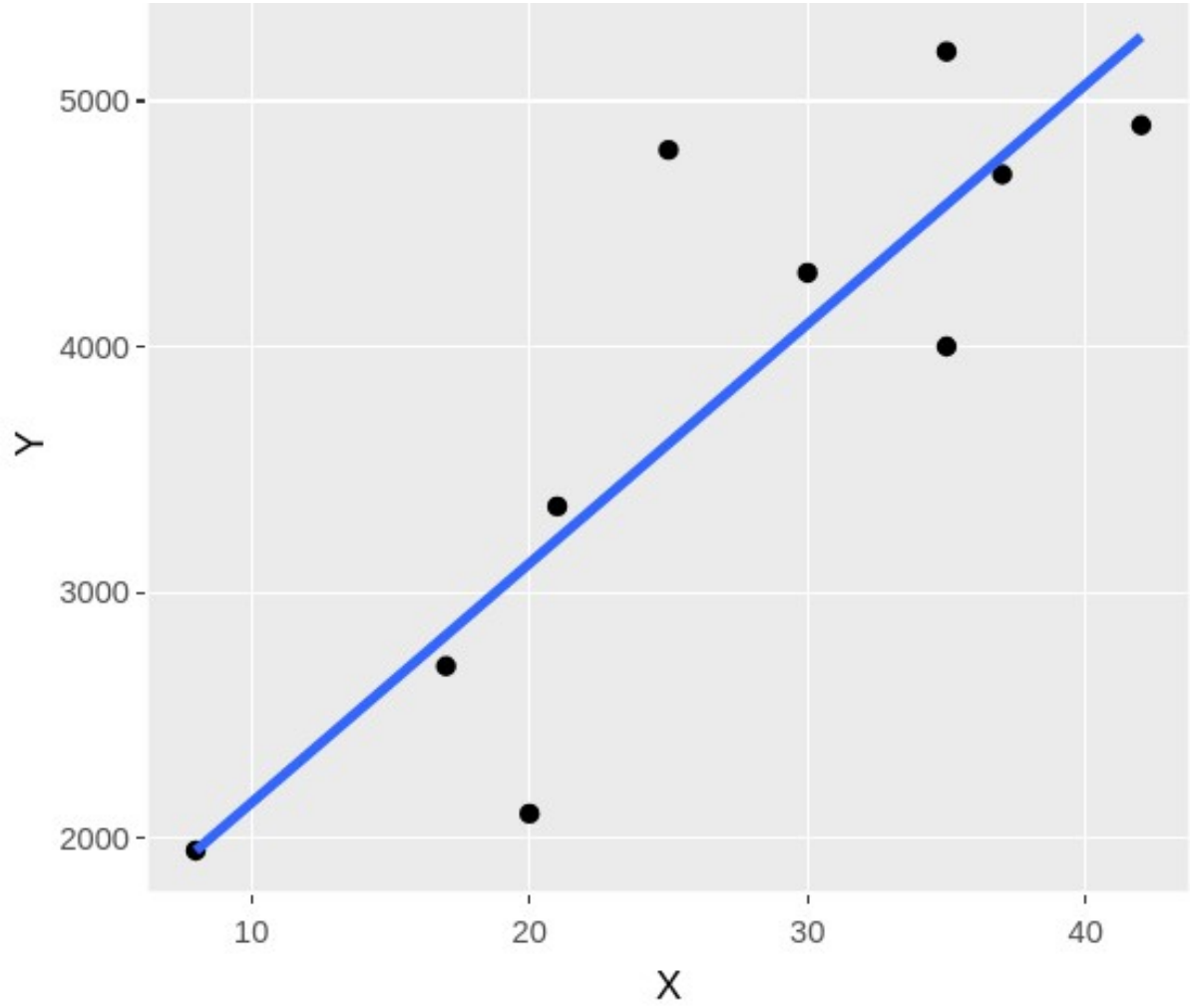
Relação Direta

#	X	Y
1	30	4300
2	21	3350
3	35	5200
4	42	4900
5	37	4700
6	20	2100
7	8	1950
8	17	2700
9	35	4000
10	25	4800

X = Idade

e

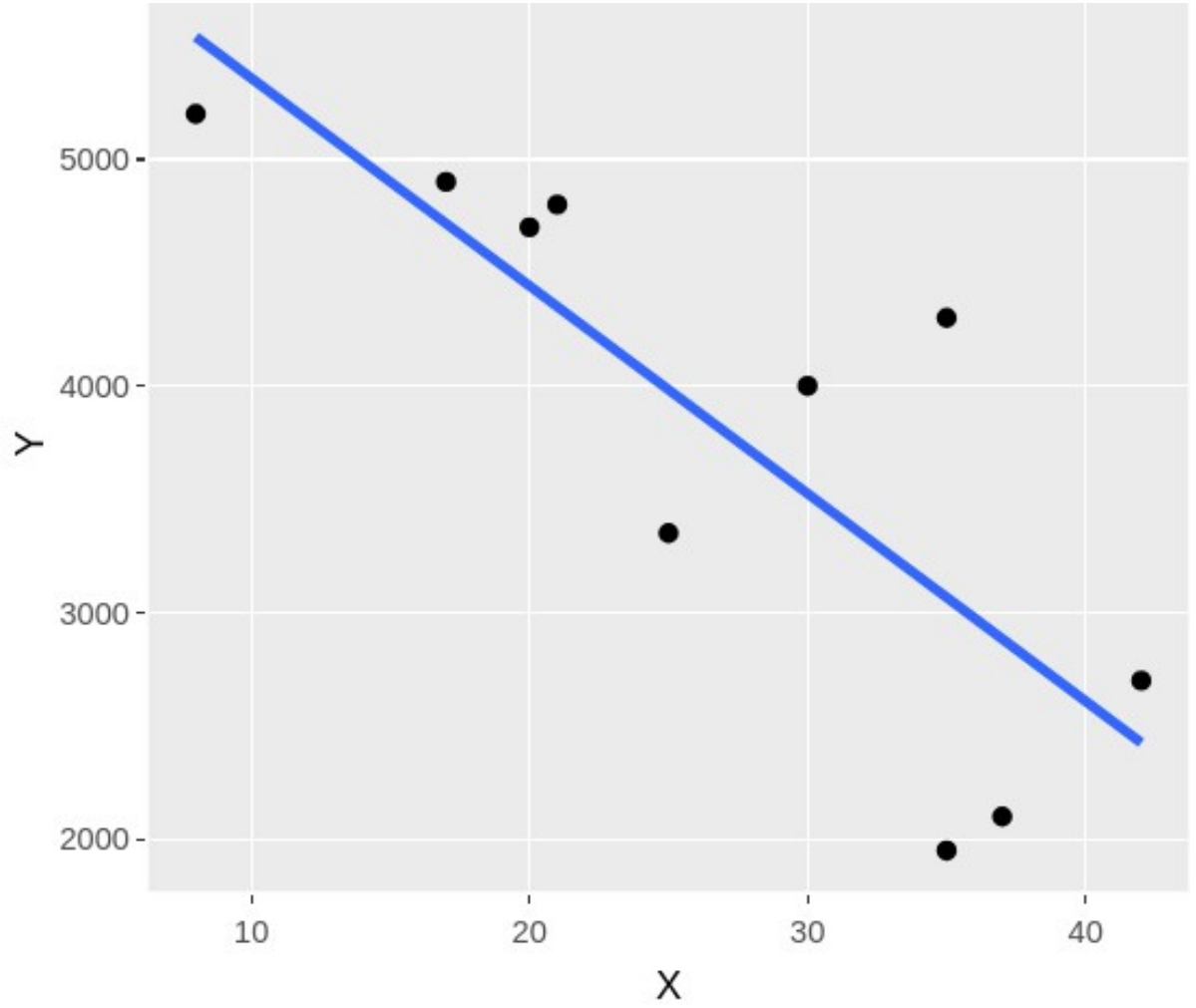
Y = Renda mensal



Relação Inversa

#	X	Y
1	35	4300
2	25	3350
3	8	5200
4	17	4900
5	20	4700
6	37	2100
7	35	1950
8	42	2700
9	30	4000
10	21	4300

X = Distância do Centro e Y = Renda



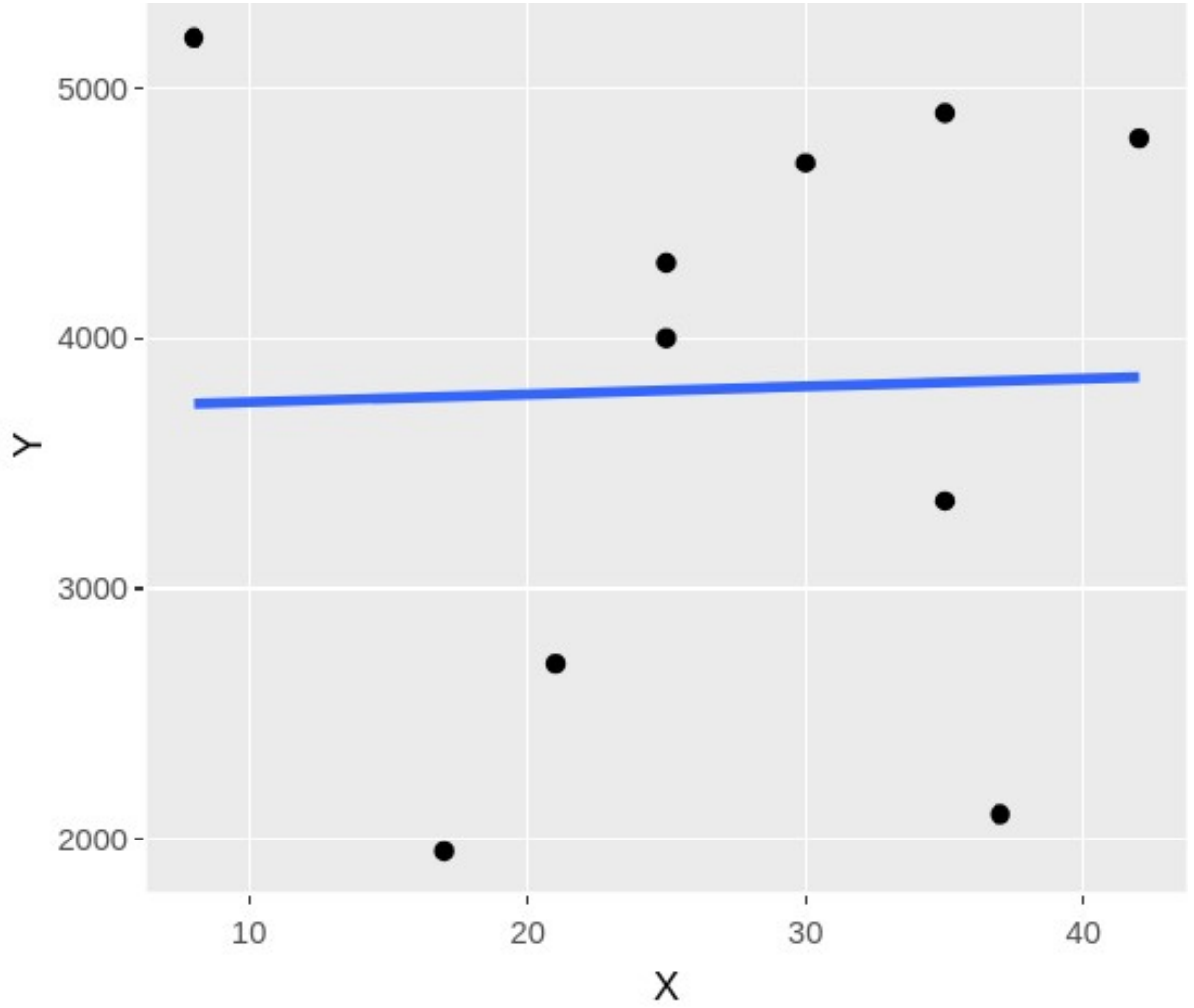
Sem Relação

#	X	Y
1	35	4300
2	25	3350
3	8	5200
4	17	4900
5	20	4700
6	37	2100
7	35	1950
8	42	2700
9	30	4000
10	21	4800

X = Idade

e

Y = Distância do Centro



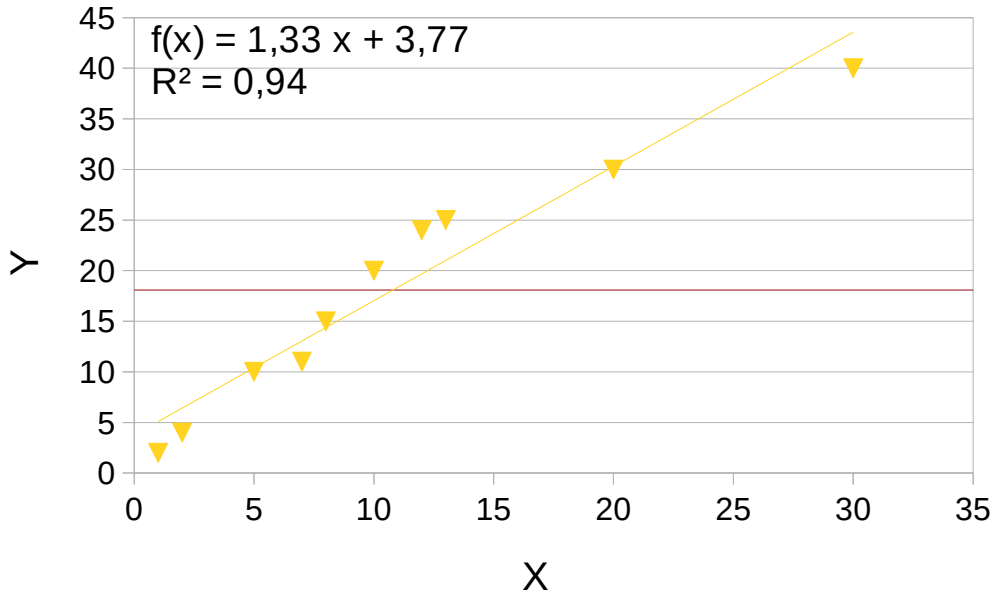
Erro ou Desvio

- Haverá sempre alguma diferença entre o valor observado Y e o valor estimado Y'. Essa diferença em estatística é chamada de erro ou desvio:
  - erro = Y – Y'
- O erro indica que
  - As variações de Y não são perfeitamente explicadas pelas variações de X ou
  - Existem outras variáveis das quais Y depende ou
  - Os valores de X e Y são obtidos de uma amostra particular que não é representativa da realidade

Exemplo - Regressão

- A regressão significa que os pontos plotados no gráfico são regredidos, isto é, são definidos ou modelados por uma reta que corresponde à menor distância entre cada ponto plotado e a reta

Modelo de Regressão Linear



$$y = \alpha + \beta \cdot x$$

Fórmulas para encontrar o coeficiente linear α e angular β

$$\beta = \frac{\frac{\sum \left(x_i - \bar{x}\right) \cdot \left(y_i - \bar{y}\right)}{n}}{\frac{\sum \left(x_i - \bar{x}\right)^2}{n}}$$

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

Medidas de Avaliação

Table 5.8 Performance Measures for Numeric Prediction

Mean-squared error	$\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}$
Root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}}$
Mean absolute error	$\frac{\lvert p_1-a_1\rvert+\cdots+\lvert p_n-a_n\rvert}{n}$
Relative squared error	$\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}$
	(in this formula and the following two, $\bar{a}$ is the mean value over the training data)
Root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}}$
Relative absolute error	$\frac{\lvert p_1-a_1\rvert+\cdots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\cdots+\lvert a_n-\bar{a}\rvert}$
Correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i(p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_P=\frac{\sum_i(p_i-\bar{p})^2}{n-1}$, $S_A=\frac{\sum_i(a_i-\bar{a})^2}{n-1}$ (here, $\bar{a}$ is the mean value over the test data)

Qual medida utilizar?

Table 5.9 Performance Measures for Four Numeric Prediction Models

	A	B	C	D
Root mean-squared error	67.8	91.7	63.3	57.4
Mean absolute error	41.3	38.5	33.4	29.2
Root relative squared error	42.2%	57.2%	39.4%	35.8%
Relative absolute error	43.1%	40.1%	34.8%	30.4%
Correlation coefficient	0.88	0.88	0.89	0.91

- Depende do contexto
- Coeficiente de correlação mais uma ou duas

CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Regressão

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com