my first regression project

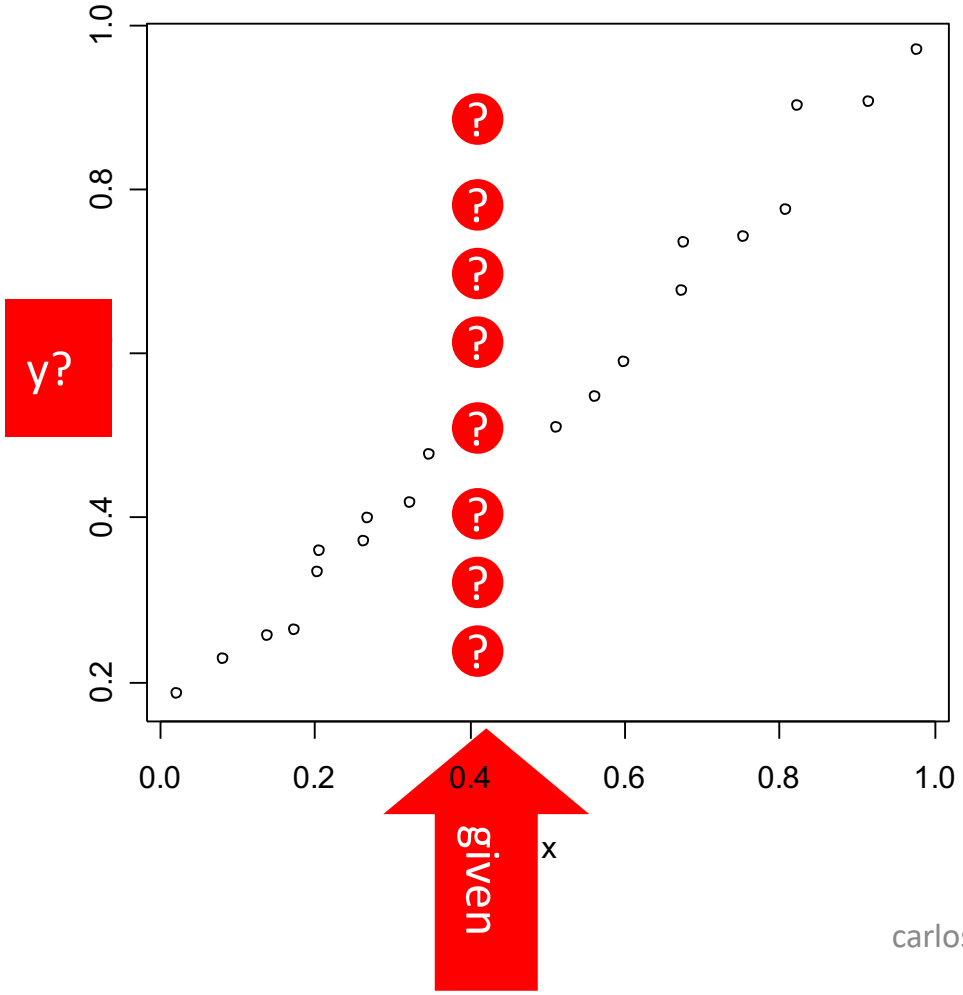
Carlos Soares

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

http://www.flickr.com/photos/upv
cfrontdoor/4642103837/

predictive: regression to estimate customer value

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



- y is customer value
- x is family income
  - and other characteristics
  - ... only 1 here for simplicity

plan & goals

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- regression
  - introduction
  - my first regression (in RM...)
- linear regression
- evaluation of Regressors
- other algorithm

- regression concepts
  - interpretation of the linear model
  - evaluation measures of regression models
- understand the need to use different sets of data for modelling and for evaluation
- know how to evaluate the results of a classification model
  - conceptually and in RapidMiner

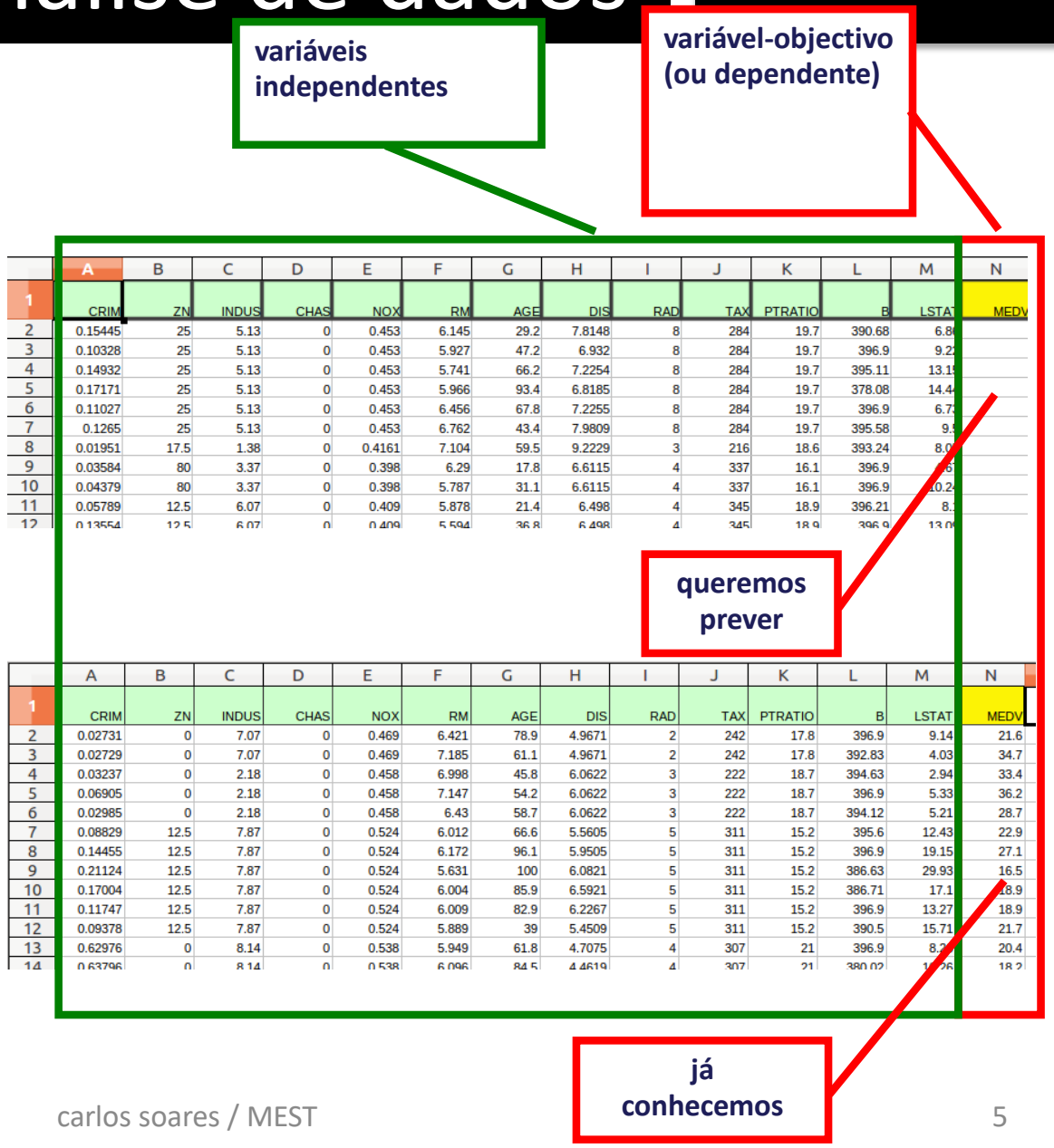
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

REGRESSÃO

previsão de valores numéricos: abordagem de análise de dados

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- novas observações para as quais queremos fazer a previsão
  - ex. nova zona da cidade
- observações conhecidas
  - ex. bairros onde empresa já está implantada



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
2	0.15445	25	5.13	0	0.453	6.145	29.2	7.8148	8	284	19.7	390.68	6.8	
3	0.10328	25	5.13	0	0.453	5.927	47.2	6.932	8	284	19.7	396.9	9.2	
4	0.14932	25	5.13	0	0.453	5.741	66.2	7.2254	8	284	19.7	395.11	13.1	
5	0.17171	25	5.13	0	0.453	5.966	93.4	6.8185	8	284	19.7	378.08	14.4	
6	0.11027	25	5.13	0	0.453	6.456	67.8	7.2255	8	284	19.7	396.9	6.7	
7	0.1265	25	5.13	0	0.453	6.762	43.4	7.9809	8	284	19.7	395.58	9.5	
8	0.01951	17.5	1.38	0	0.4161	7.104	59.5	9.2229	3	216	18.6	393.24	8.0	
9	0.03584	80	3.37	0	0.398	6.29	17.8	6.6115	4	337	16.1	396.9	4.6	
10	0.04379	80	3.37	0	0.398	5.787	31.1	6.6115	4	337	16.1	396.9	10.24	
11	0.05789	12.5	6.07	0	0.409	5.878	21.4	6.498	4	345	18.9	396.21	8.1	
12	0.13554	12.5	6.07	0	0.409	5.594	36.8	6.498	4	345	18.9	396.9	13.0	

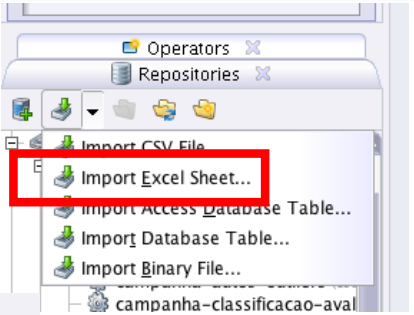
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
11	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
12	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
13	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.2	20.4
14	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2

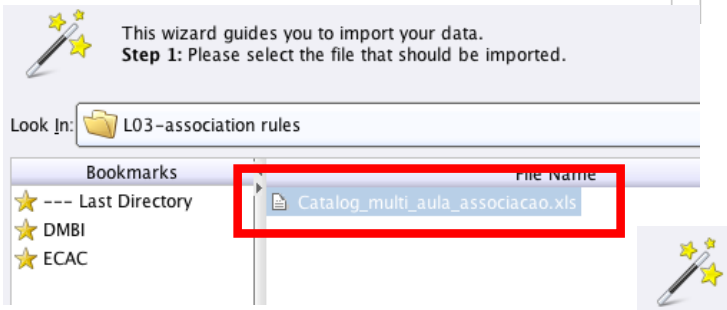
my first regression (in RM): load data

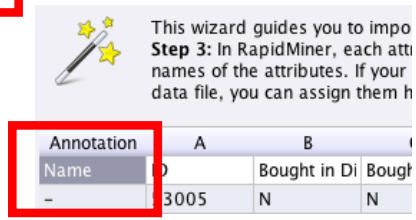
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

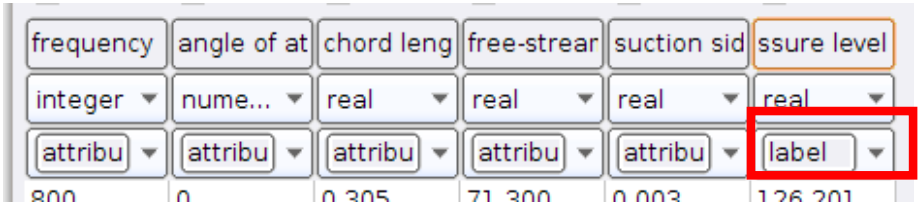
pleases note: the figures are for another file!

- data file: regression data.xlsx
- load data into repositor
  1. choose file
  2. choose worksheet
     - housing
  3. indicate row with column names (if any)
     - first row
  4. indicate column with ids of rows (if any)
     - none
  5. indicate column with target variable (if any)
     - last one
  6. give name
     - boston-housing
  7. finish





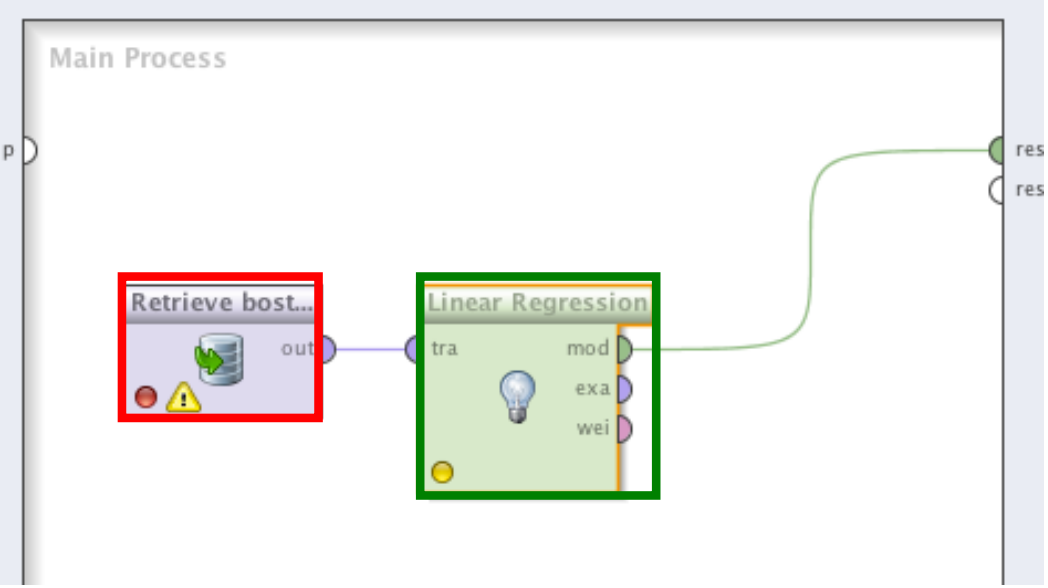




construir processo de regressão no rapid miner

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

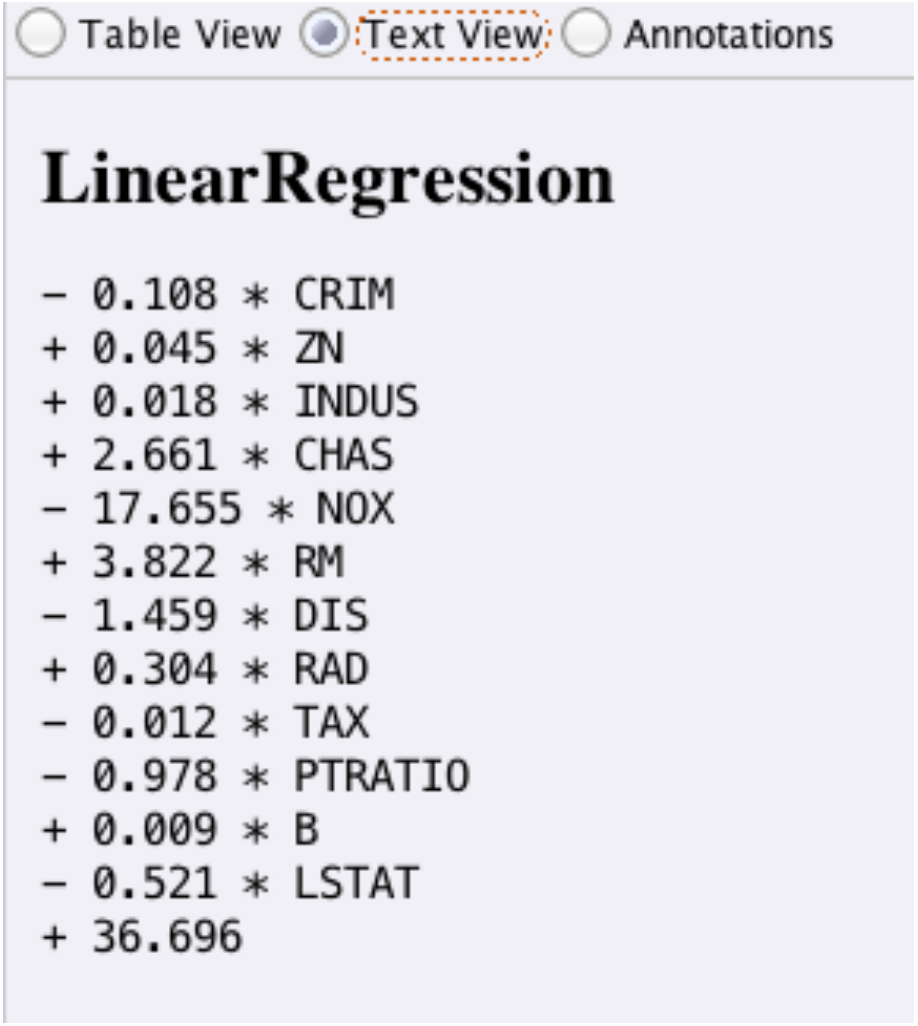
- carregar dados do repositório
- aplicar algoritmo de regressão linear
  - ex. operators ➔ modeling ➔ classification and regression ➔ function fitting ➔ linear regression
- executar processo



analyze linear regression model

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- assumes variables are not correlated
  - influence of each variable explained separately
  - coefficients are not influenced by changes in the set of independent variables
- variation depends on magnitude of correlation
  - sign might change!
- ... but empirical results indicate robustness



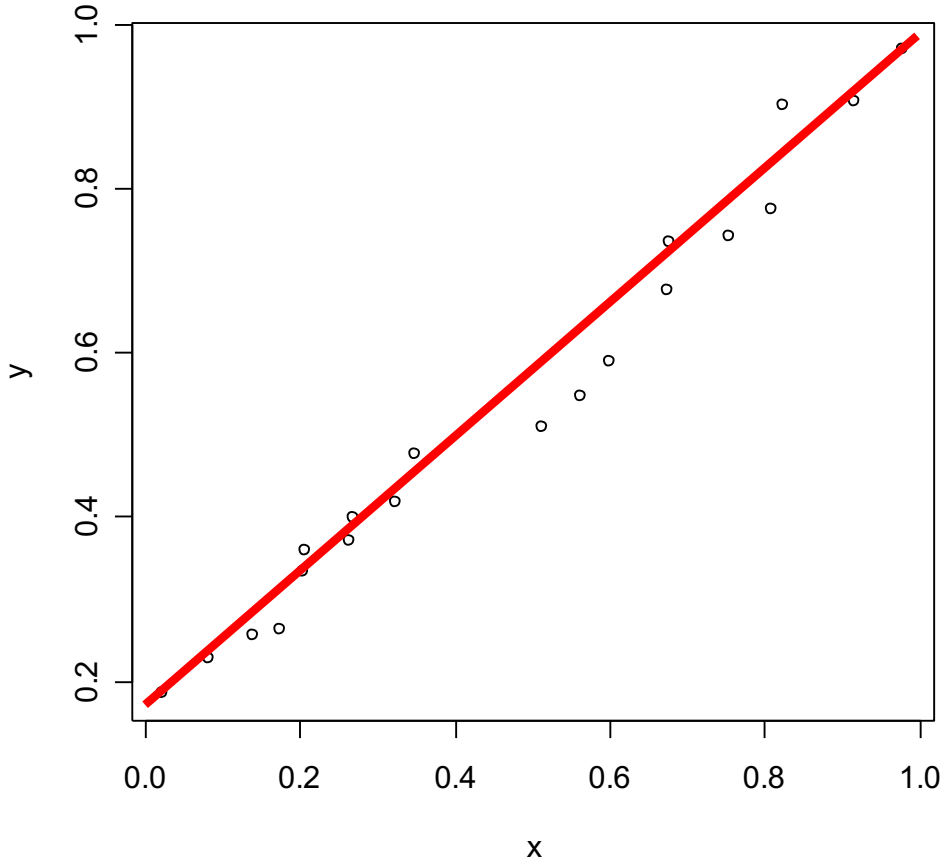
linear regression

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- simple case: 2 variables
  x and *y*

- equation of the line

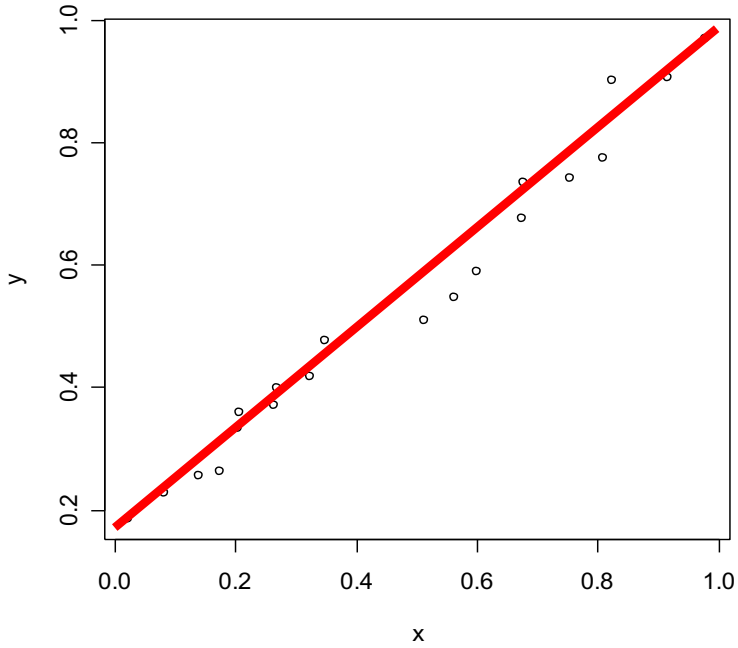
$$y = f(x) = b_0 + b_1 x$$



interpretion of coefficients

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

$$y = b_0 + b_1 x$$



- b_0: intersection of the line with the *yy* axis
  - frequently hard to interpret
- b_1: slope of the line
  - variation of the value of *y* given an increase of 1 unit of *x*

make prediction for new examples

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

given a neighborhood with the following characteristics

CRIM	0,04294
ZN	28
INDUS	15,04
CHAS	0
NOX	0,464
RM	6,249
AGE	77,3
DIS	3,615
RAD	4
TAX	270
PTRATIO	18,2
B	396,9
LSTAT	10,59

… what is the predicted value?

Table View | Text View | Annotations

LinearRegression

```
- 0.108 * CRIM
+ 0.045 * ZN
+ 0.018 * INDUS
+ 2.661 * CHAS
- 17.655 * NOX
+ 3.822 * RM
- 1.459 * DIS
+ 0.304 * RAD
- 0.012 * TAX
- 0.978 * PTRATIO
+ 0.009 * B
- 0.521 * LSTAT
+ 36.696
```

estimating parameters: statistics meets optimization meets algorithms

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

$$y = b_0 + b_1 x$$

$$\widehat{b_1} = \frac{S_{XY}}{S_{XX}}$$

where $\widehat{b_1}$ is an estimate of b

$$S_{XY} = \sum_{i=1}^{n} [(X_i - \overline{X}).(Y_i - \overline{Y})]$$

$$S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- $\widehat{b_1}$ should be statistically significantly different from zero
  - if not, there is no meaningful dependency between *Y* and X
  - this should be tested

$$\widehat{b_0} = \overline{Y} - \hat{\beta}.\overline{X}$$

where $\widehat{b_0}$ is an estimate of b_0

- $\widehat{b_0}$ may or may not be statistically significantly different from zero
  - If not there is no evidence that *Y≠0* when *X=0.*
  - … which could make sense
    - e.g. value of a customer with 0 income
  - … or not…
    - e.g. minimum sales of a product without shelf space

Simple linear regression: assumptions

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Linear relationship between *x* and *y*
  - also additive
- Errors
    - i.e. unexplained variation in *y*
  - ... are independently and identically distributed
  - ... homoscedasticity
    - constant variance
  - ... normally distributed

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

AVALIAÇÃO DE MODELOS DE REGRESSÃO

regressão: resumo (até agora)

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

pleases note: the figures are for another file!

exemplos com valor conhecido da variável-objetivo

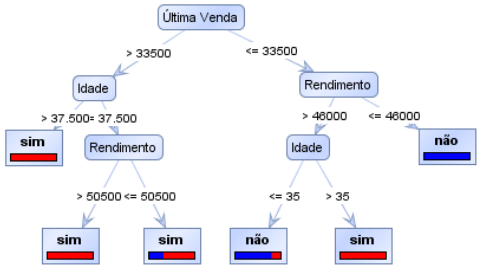
Comprou	Idade	Rendimento	Ag.fam	Vendas anteriores	Última Venda
não	37	49000	2	1	42000
sim	43	68000	3	0	0
sim	42	61000	4	0	0
sim	26	52000	2	0	0
sim	40	64000	1	1	21000
sim	38	52000	1	0	0
sim	45	43000	4	1	47000
sim	35	45000	2	1	34000
não	39	43000	2	0	0

(novos) exemplos com valor desconhecido da variável-objetivo

	A	B	C	D	
1	Comprou	Idade	Rendimento	Ag.fam	Vendas
2		41	50000	2	
3		39	68000	2	
4		58	61000	4	
5		26	25000	3	
6		21	50000	1	
7		38	43000	2	
8		44	43000	4	
9		27	47000	2	
10		70	23000	2	

1

2



3

previsões para os (novos) exemplos com valor desconhecido da variável-objetivo

row no.	Comprou	prediction(...	confidence(...	confidence(...	Ida
1	?	sim	0	1	41
2	?	sim	0	1	39
3	?	sim	0	1	58
4	?	não	1	0	26
5	?	não	0.818	0.182	21
6	?	não	1	0	38
7	?	sim	0	1	44
8	?	não	0.818	0.182	27
9	?	não	1	0	70

usariam as previsões feitas por este modelo?

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- regression
- linear regression
- evaluation of regressors
  - measures
  - methodology

prediction and evaluation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

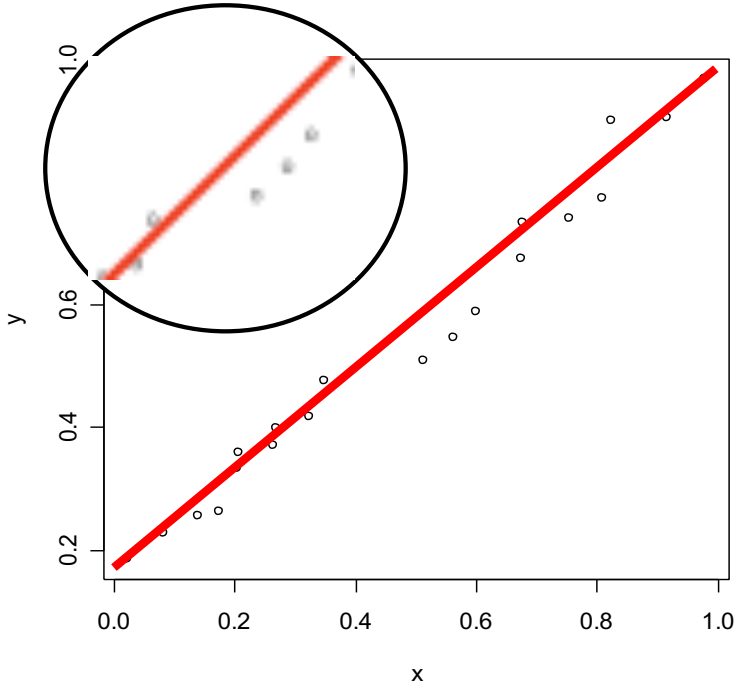
- given the value of *x*
- model estimates the value of *y*

$$\hat{y} = b_0 + b_1 x$$

- but the estimate is not perfect!

$$\hat{y} - y$$

- error:
  - y : true value
  - ŷ : value estimated by the model

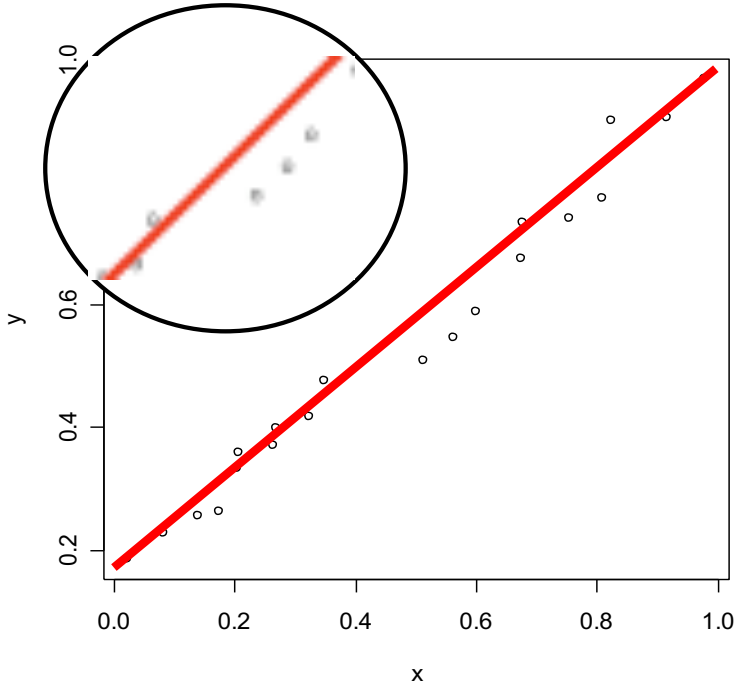


common evaluation measures

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- average error
  - do not use!

$$1/m \sum_i \hat{y}_i - y_i$$



- mean absolute deviation

$$1/m \sum_i |\hat{y}_i - y_i|$$

- mean squared error

$$1/m \sum_i (\hat{y}_i - y_i)^2$$

common evaluation measures: MAD vs MSE

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

$$\frac{1}{m}\sum_i |\hat{y}_i - y_i| \qquad \frac{1}{m}\sum_i (\hat{y}_i - y_i)^2$$

	i=1	… 2	… 3	…	… i=10	MAD	MSE
y_i	100	100	100	…	100		
$\hat{y}_i^{M_1}$	99	99	99	…	99		
$\lvert\hat{y}_i^{M_1} - y_i\rvert$	1	1	1	…	1	$\frac{1\times10}{10} = 1$	
$(\hat{y}_i^{M_1} - y_i)^2$	1	1	1	…	1		$\frac{1\times10}{10} = 1$
$\hat{y}_i^{M_2}$	90	100	100	…	100		
$\lvert\hat{y}_i^{M_2} - y_i\rvert$	10	0	0	0	0	$\frac{10 + 0\times9}{10} = 1$	
$(\hat{y}_i^{M_2} - y_i)^2$	100	0	0	0	0		$\frac{100 + 0\times9}{10} = 10$

analysis of evaluation measures

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- mean absolute deviation

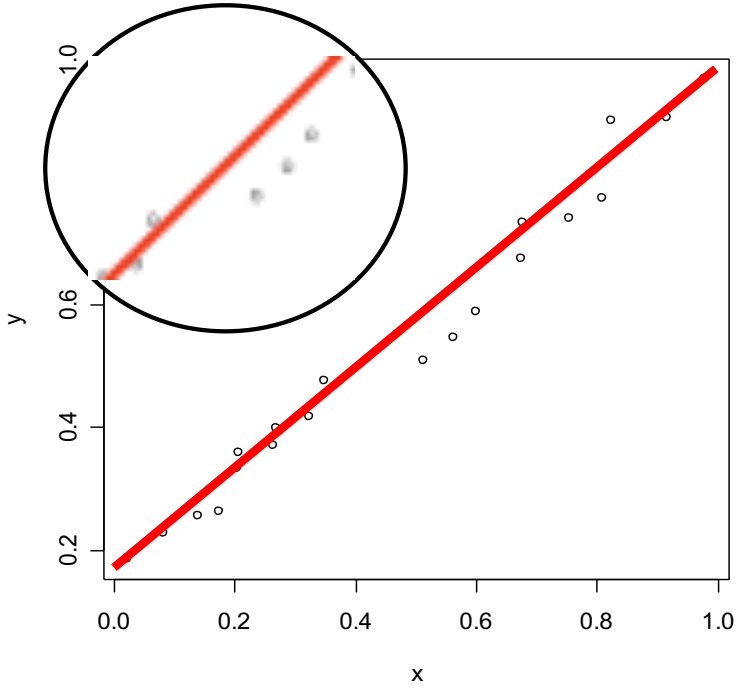
$$\frac{1}{m}\sum_i |\hat{y}_i - y_i|$$

  - estimates “typical” error

- mean squared error

$$\frac{1}{m}\sum_i (\hat{y}_i - y_i)^2$$

  - gives more importance to larger errors
  - ... possible dominated by a few errors



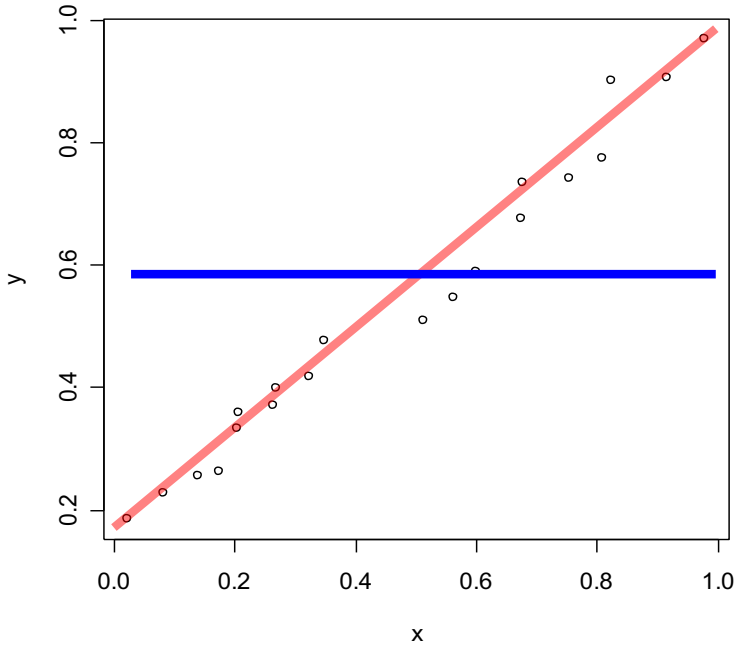
- values depend on the scale of the target variable
  - good or bad?
    - business perspective
    - but is it really the relationship between *x* and *y*?

baseline

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- if nothing is known about the new case
- ... what is the best prediction we can make?
  - random vs average
- baseline

$$\hat{y}_i = \bar{y}$$



- regression is only useful if the error is less than the baseline error
  - eg MSE

$$\frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

0 if prediction model is perfect

]0,1[if it is useful

1 if it is the same as the baseline

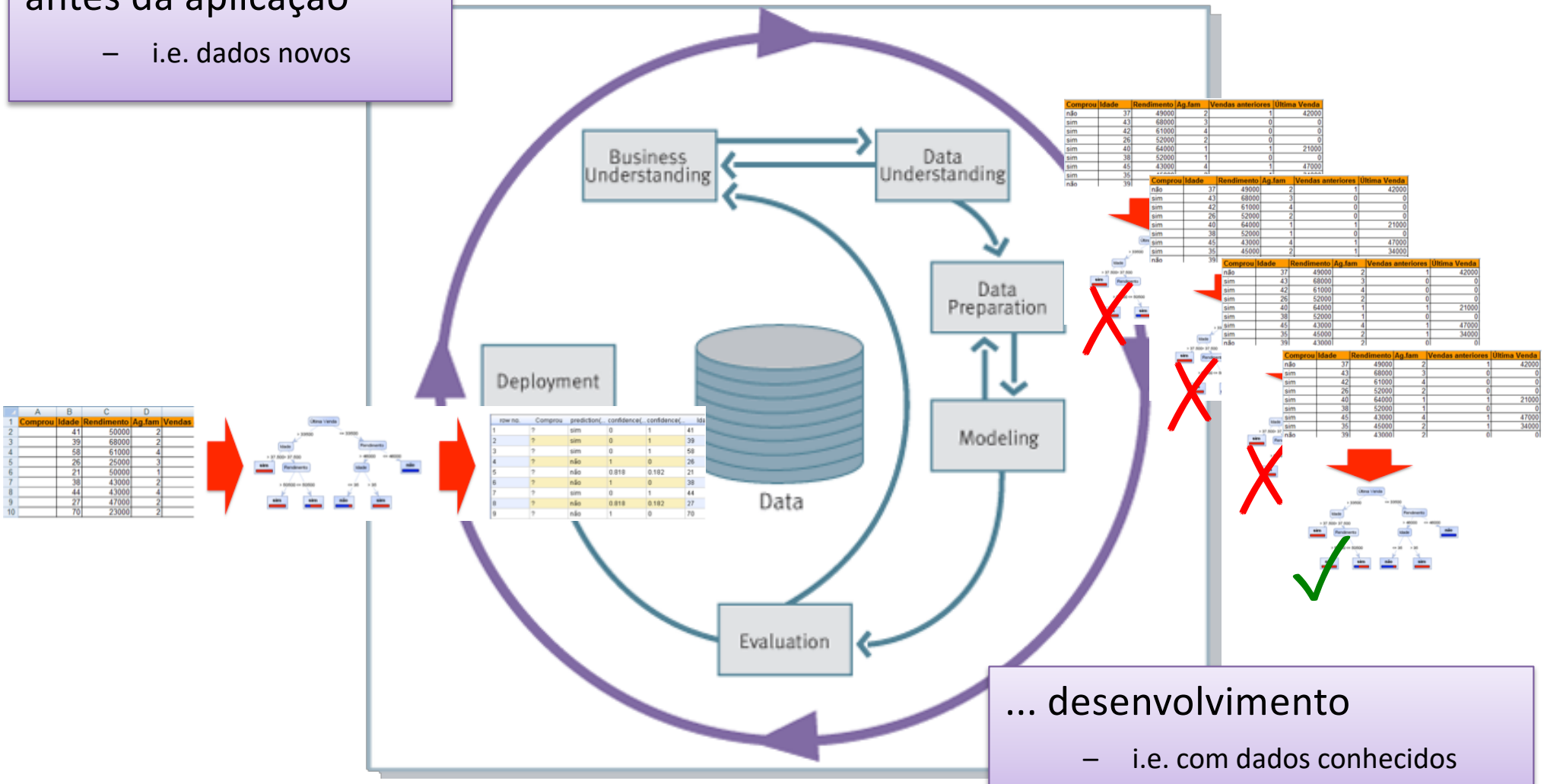
>1 if it is worse than the baseline

desenvolvimento de modelos

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

antes da aplicação

- i.e. dados novos

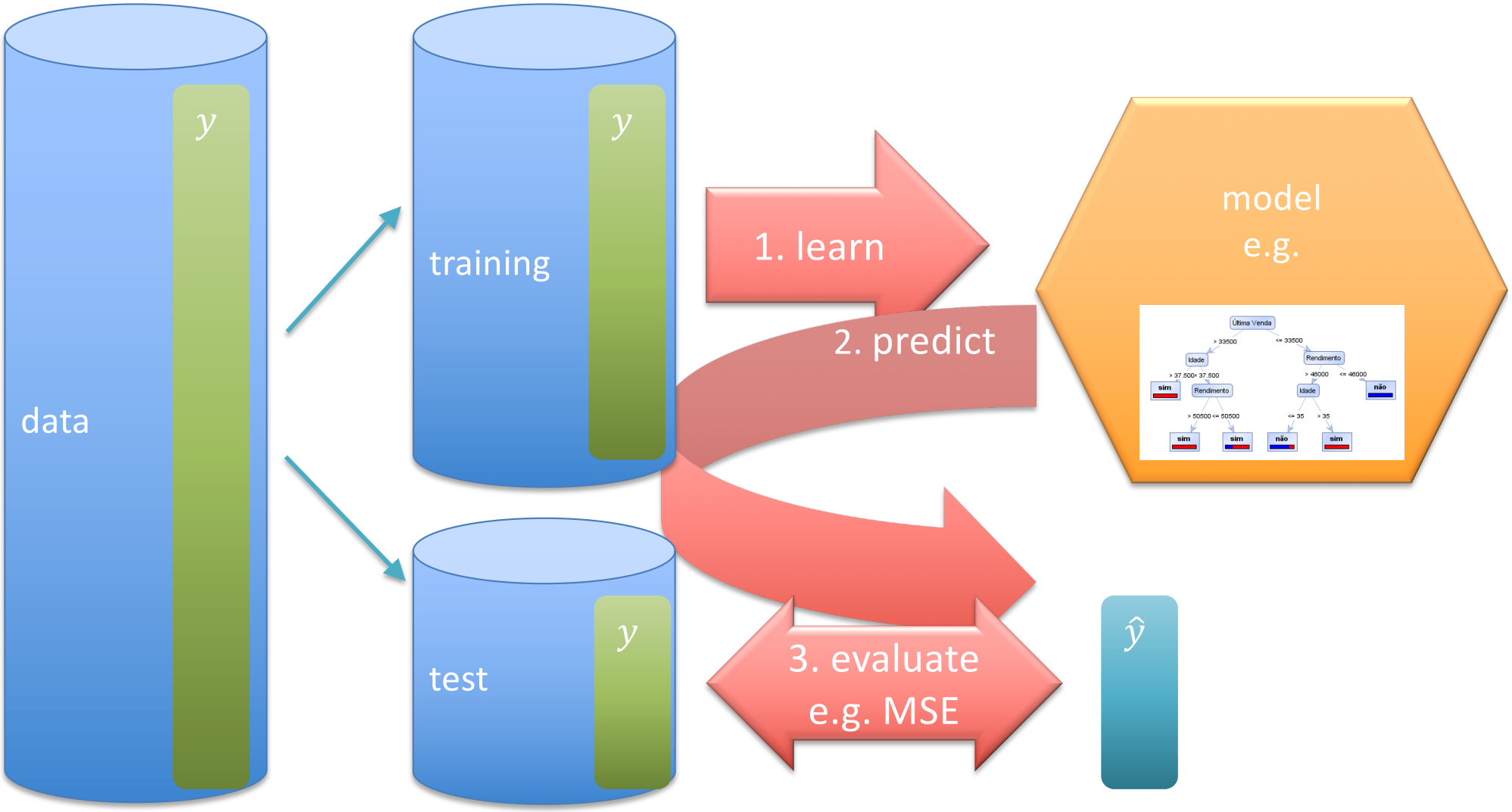


... desenvolvimento

- i.e. com dados conhecidos

evaluation methodology: do not forget!

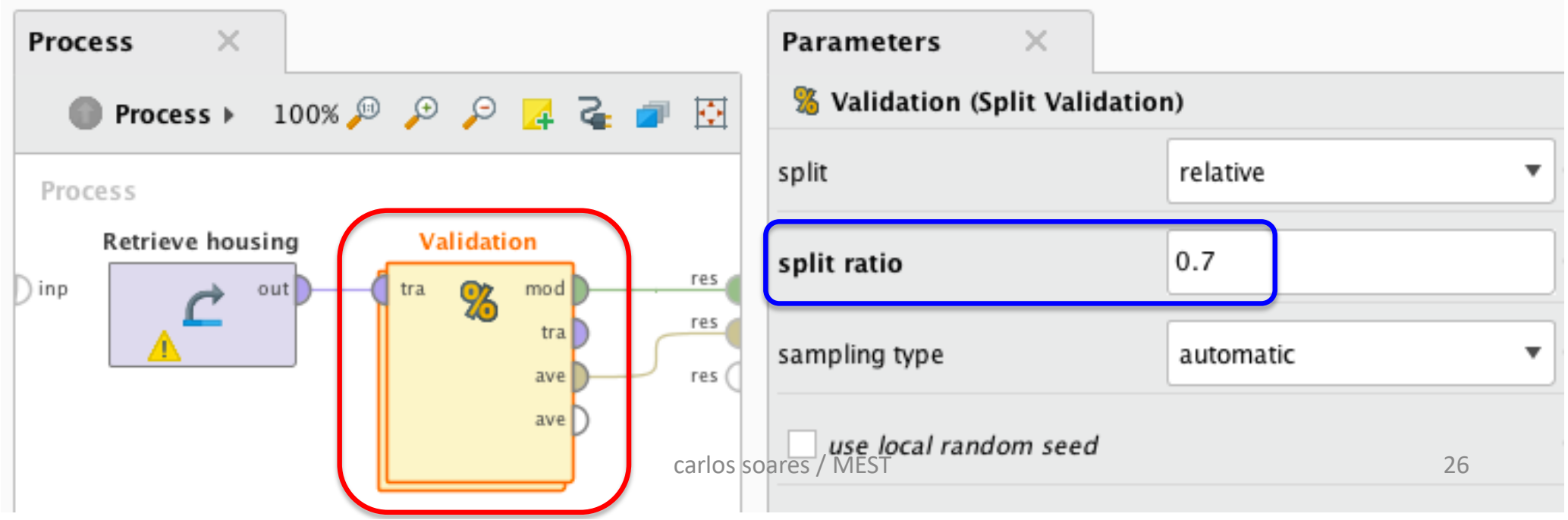
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



avaliar modelo de regressão em rapid miner (1/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

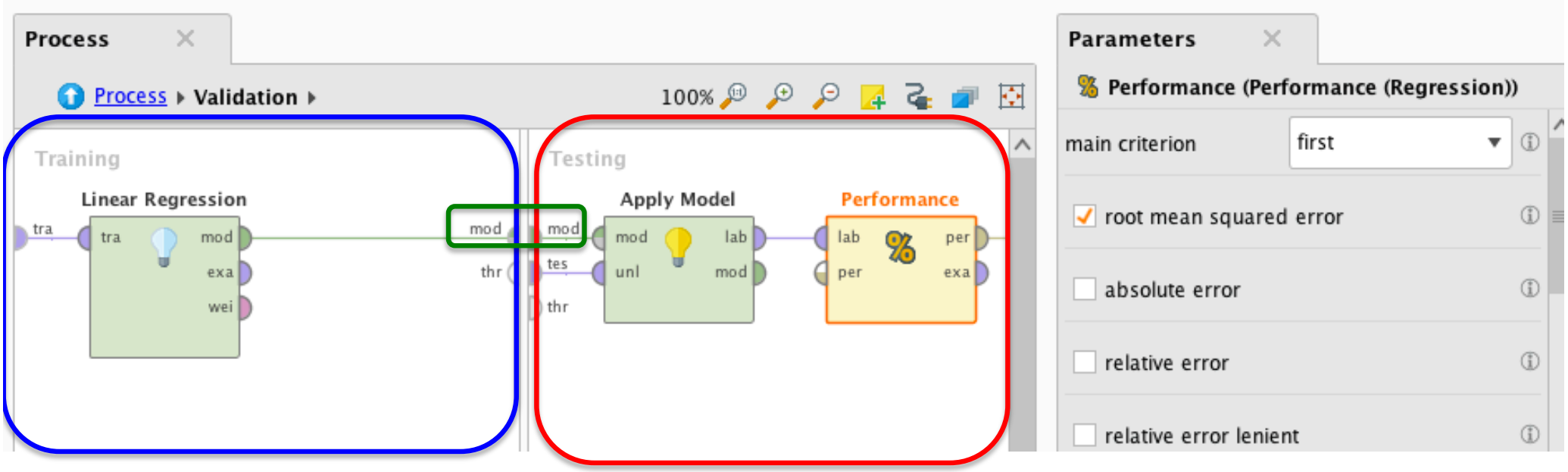
- operador split validation
- sub-processo
  - operador que contém outros operadores
  - duplo clique para entrar
- distribuir aleatoriamente os dados por conjunto de treino e conjunto de teste
  - proporção
    - 70% dos casos para treino
    - 30% dos casos para teste



estimate predictive error with rapid miner (2/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

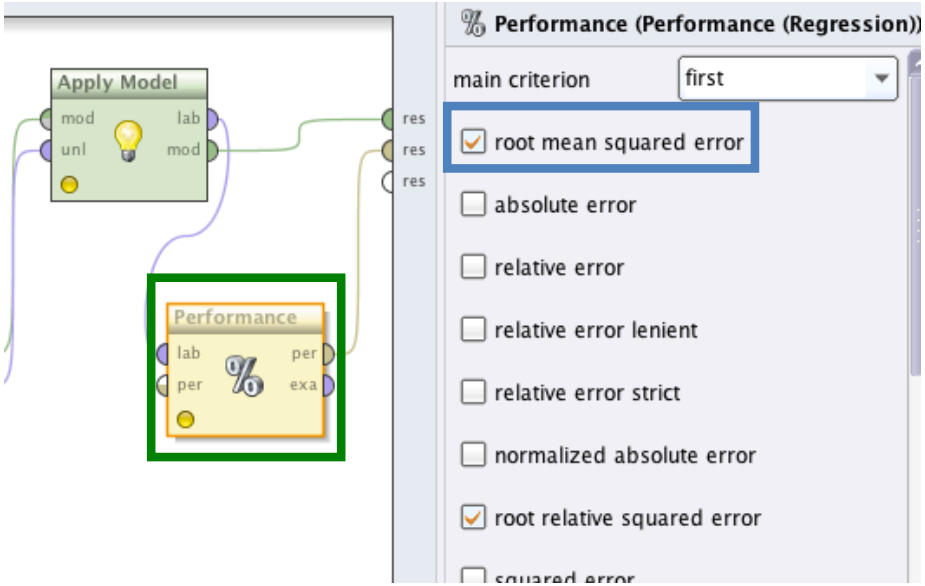
- split validation
  - different operations applied to training and test data
  - model obtained with train data is applied to test data



avaliar modelo de regressão em rapid miner (3/3)

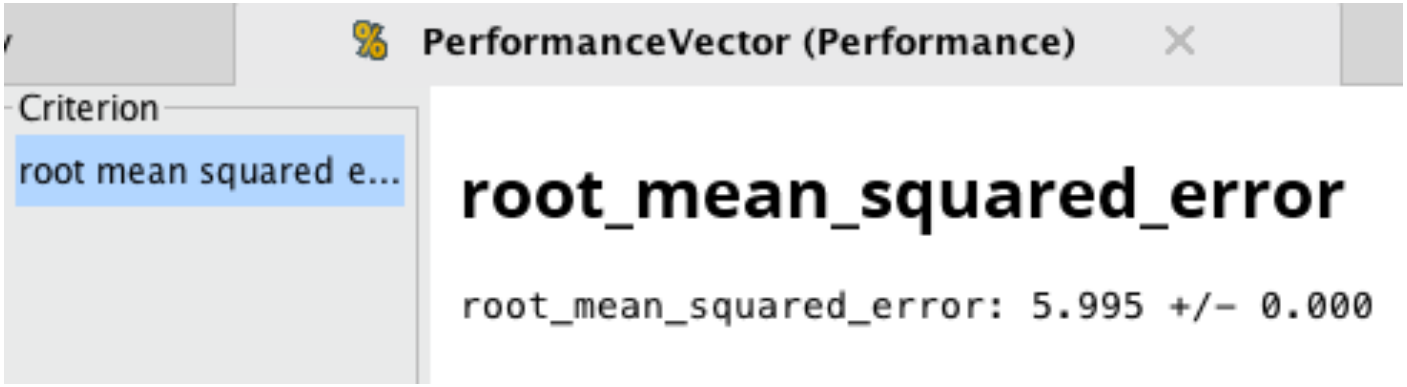
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- operador “Performance (Regression)”
- escolher medidas RMSQ

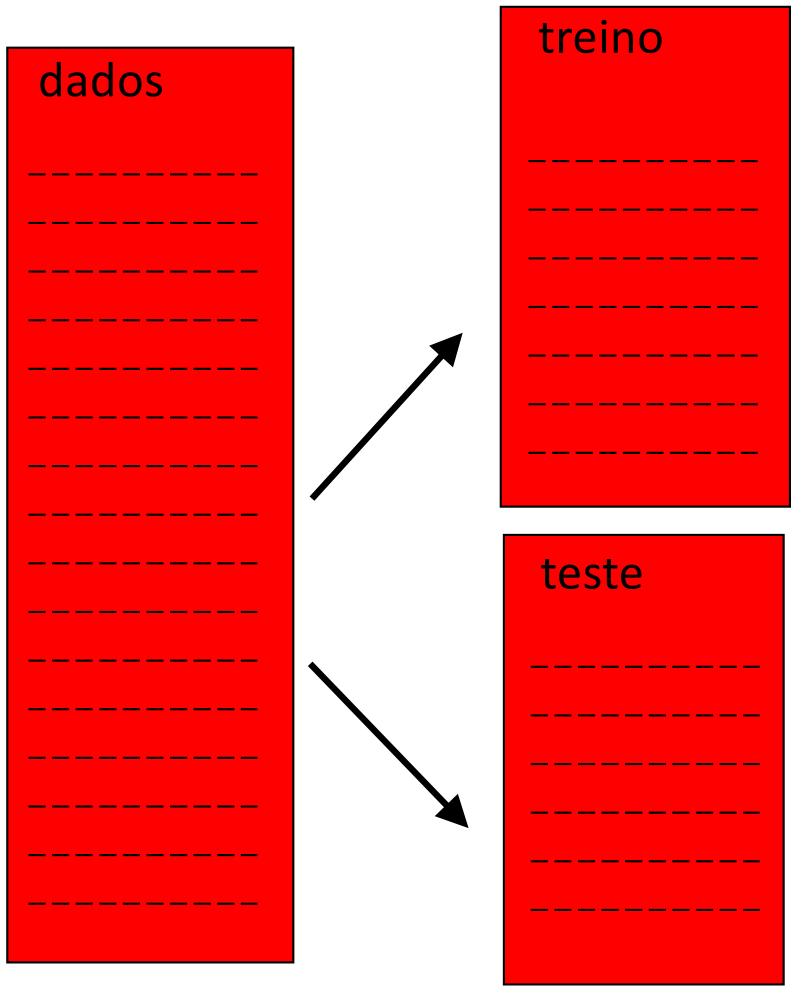


estimativa de desempenho do modelo de regressão

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

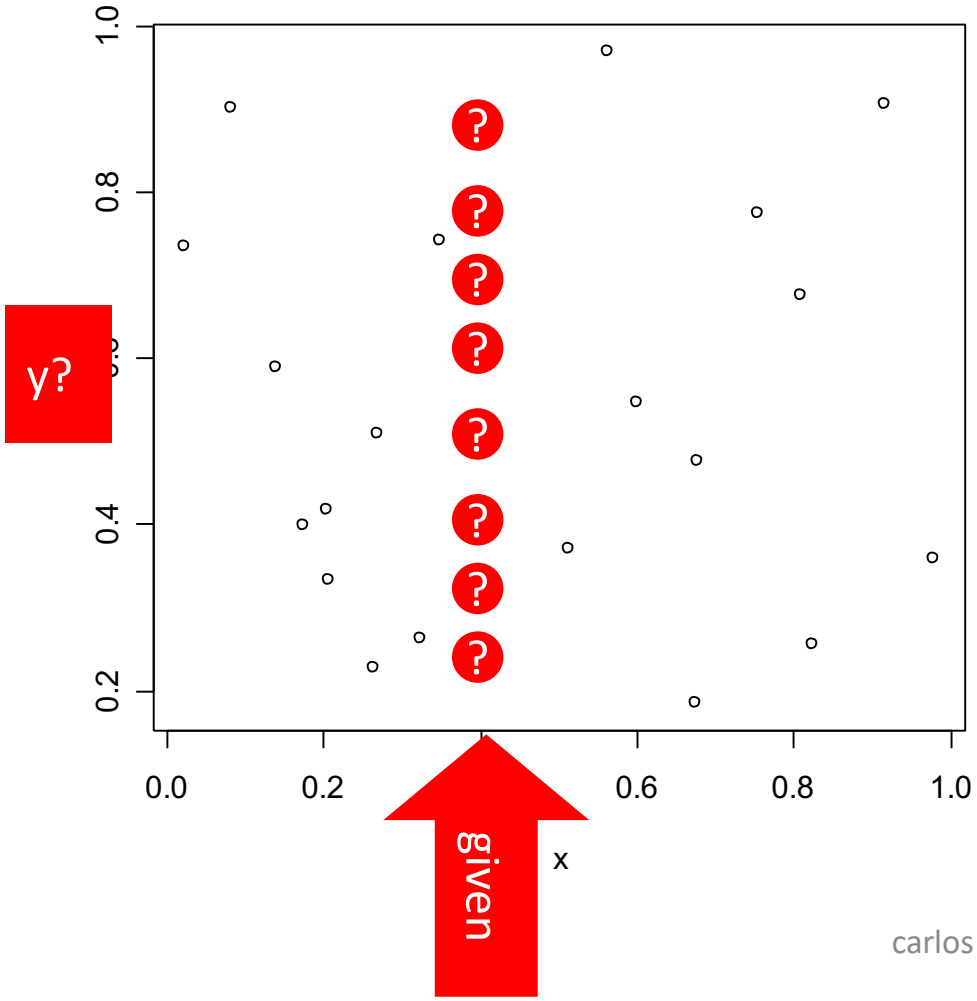


metodologia de avaliação: treino



- treino
  - para obter o modelo automaticamente
- teste
  - estimar o valor do modelo em novos casos
  - assume que os novos dados terão uma distribuição idêntica aos de treino
    - não funciona se houver alterações na distribuição: ex: inflação

short detour: brief introduction to bias



- y is customer value
- x is family income
  - and other characteristics
  - ... only 1 here for simplicity

- why is it harder to predict y now?