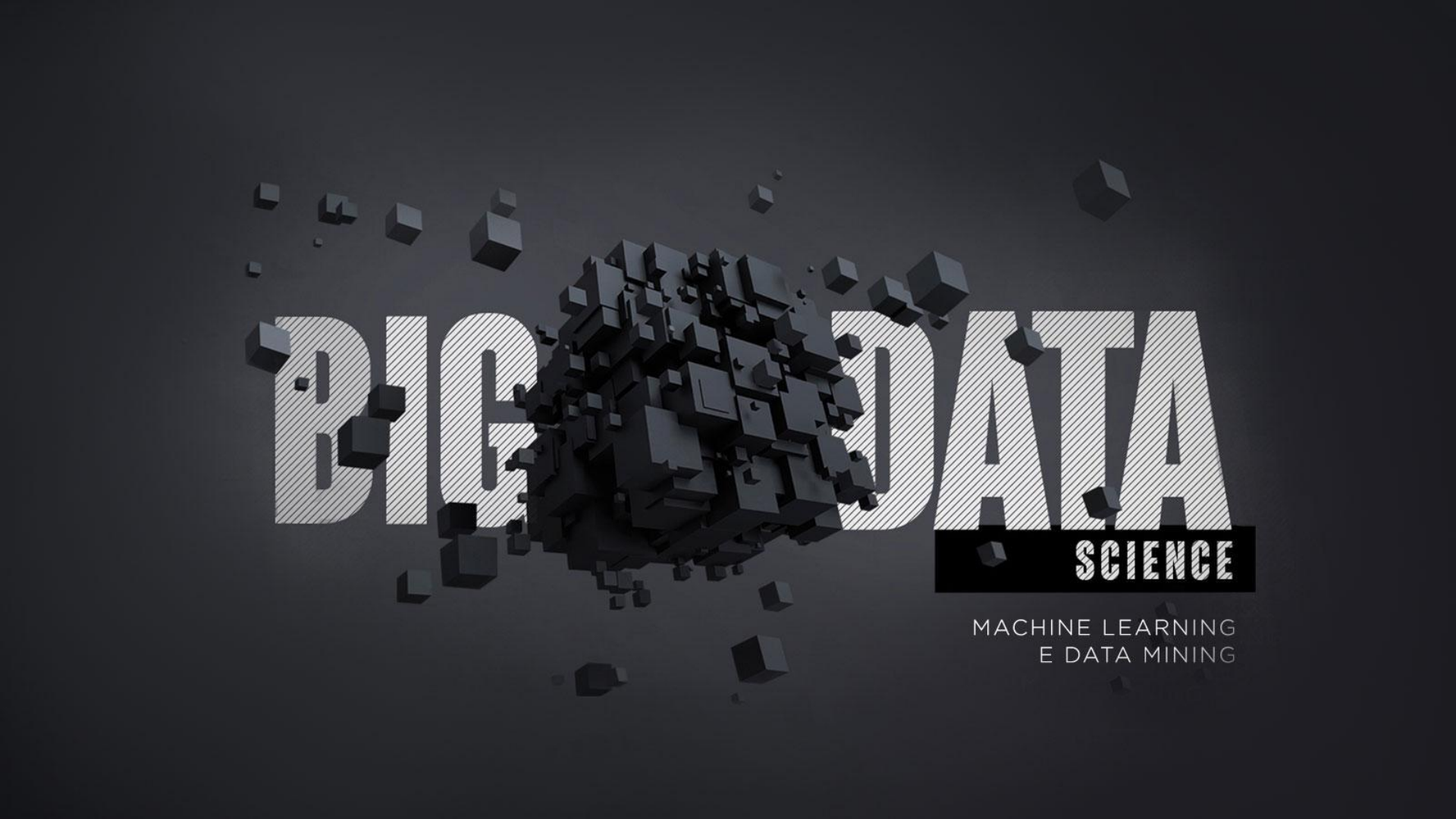

BIG DATA
SCIENCE
MACHINE LEARNING
E DATA MINING

BIG DATA SCIENCE

ESTATÍSTICA APLICADA – PARTE 2

ANÁLISE DE REGRESSÃO

DIÓGENES JUSTO

REVISÃO DA LISTA DE EXERCÍCIOS

REVISÃO DO CONTEÚDO DA PRIMEIRA AULA

MÉDIA, DESVIO PADRÃO SÃO MÉTRICAS RICAS?

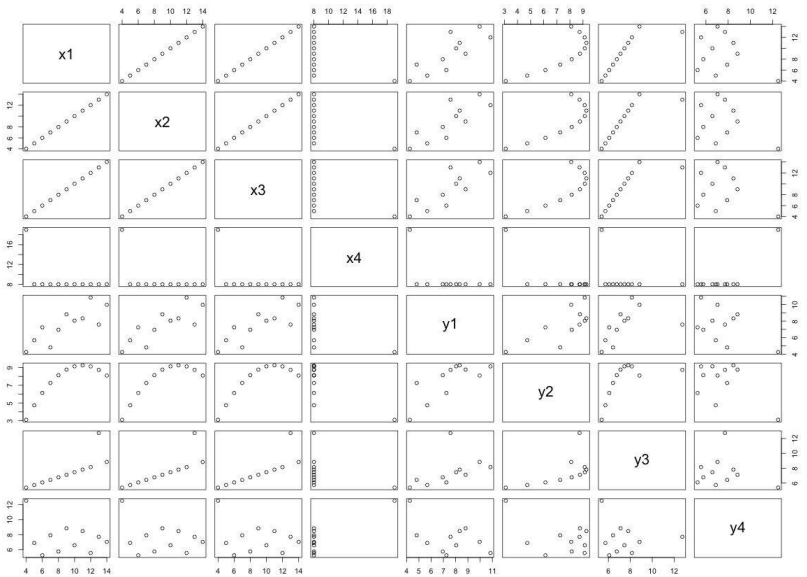
```
mean(anscombe$x1)
...
mean(anscombe$x4)
mean(anscombe$y1)
...
sd(anscombe$x1)
...
cor(anscombe$x1, anscombe$y1)
...
```

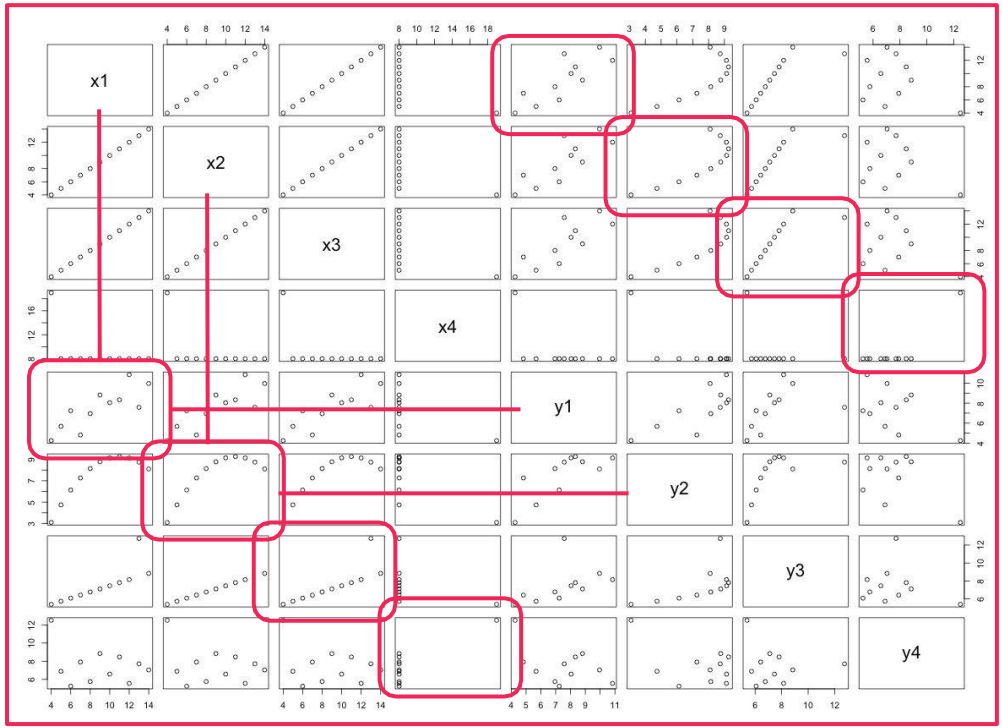
MÉDIA, DESVIO PADRÃO SÃO MÉTRICAS RICAS?

```
par(mfrow=c(2,2))
plot(anscombe$x1, anscombe$y1)
plot(anscombe$x2, anscombe$y2)
plot(anscombe$x3, anscombe$y3)
plot(anscombe$x4, anscombe$y4)
par(mfrow=c(1,1))
```

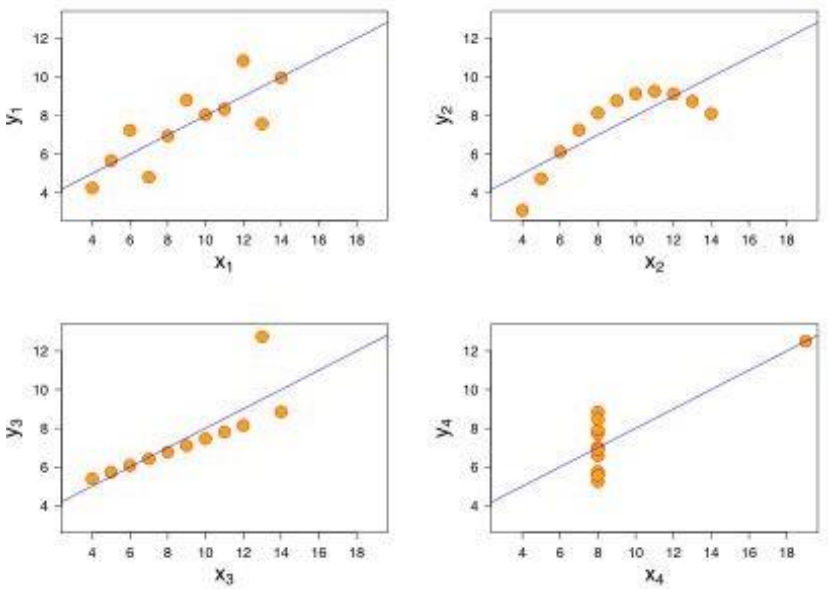
MÉDIA, DESVIO PADRÃO SÃO MÉTRICAS RICAS?

pairs(anscombe)




x1
x2
x3
x4
y1
y2
y3
y4

QUARTETO DE ANSCOMBE



https://em.wikipedia.org/wiki/anscombe's_quartet

LISTA DE EXERCÍCIOS

Média, Mediana, Moda
Score Z
Variância e Desvio Padrão

COVARIÂNCIA

CORRELAÇÃO

- Ex.1:

  R – mtcars, mpg x wt (peso do carro x consumo)

  cor(mtcars$mpg, mtcars$wt)

  plot(mtcars$mpg, mtcars$wt)

CORRELAÇÃO

Ex. 2: Excel, calcule a correlação linear do conjunto abaixo.

i	y	x
1	1	5
2	2	4
3	3	3
4	4	2
5	5	1

CORRELAÇÃO

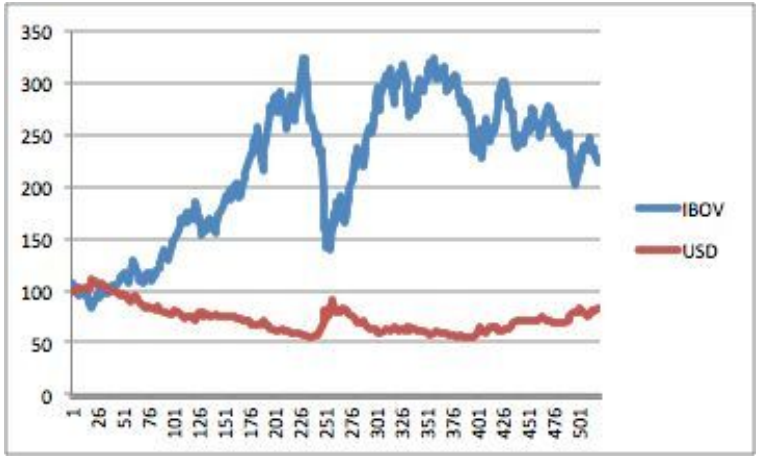
Ex. 3: Excel, calcule a correlação linear do índice iBovespa x dólar

=GoogleFinance("IBOV","price", DATE(2004, 1, 1), DATE(2013, 12, 31), "WEEKLY")

=GoogleFinance("USDBRL","price", DATE (2003, 12, 27), DATE(2013, 12, =31),"WEEKLY")

Coloque (1) ambas séries na mesma base, faça (2) gráficos de linha, (3) linha em escalas diferentes, (4) dispersão e calcule a (5) correlação linear.

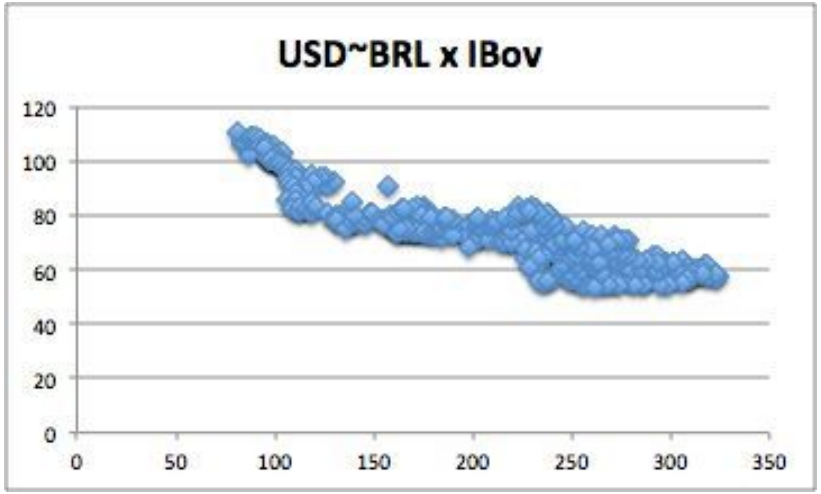
Qual a relação entre o índice iBovespa e a cotação USD x REAL?



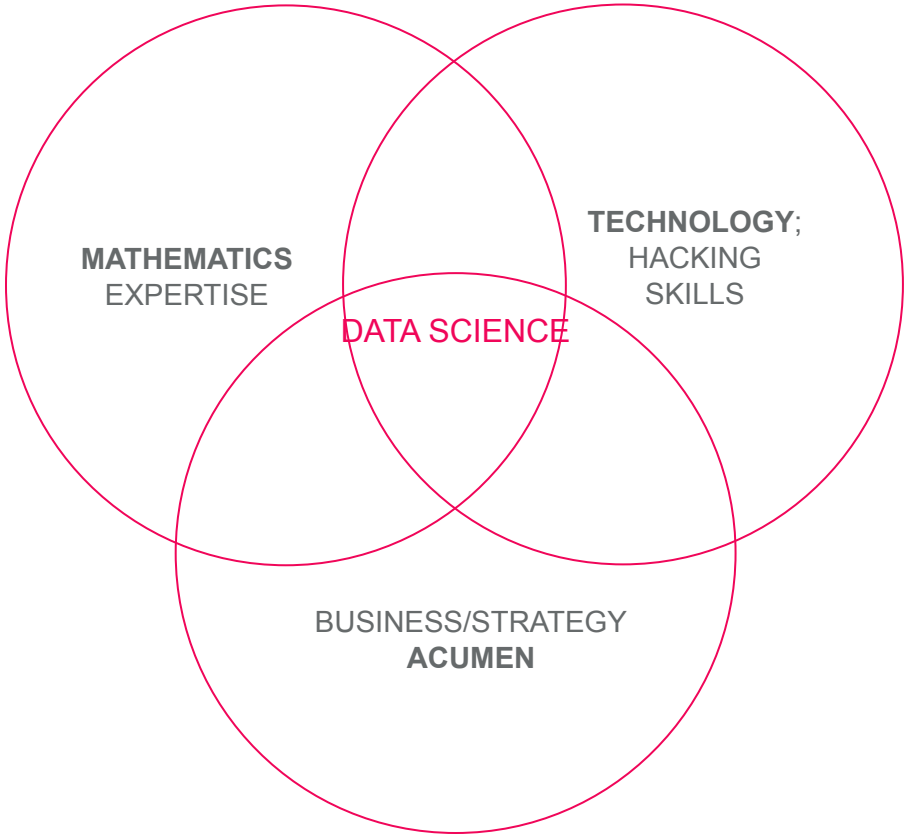
Qual a relação entre o índice iBovespa e a cotação USD x REAL?

Qual a relação entre o índice iBovespa e a cotação USD x REAL?

p˜-0,9



DATA SCIENCE


MATHEMATICS
EXPERTISE
TECHNOLOGY;
HACKING
SKILLS
DATA SCIENCE
BUSINESS/STRATEGY
ACUMEN

STEPS IN DATA ANALYSIS

- Define the **QUESTION**
- Define the **IDEAL DATA SET**
- Determine **WHAT DATA** you can access
- **OBTAIN** the data
- **CLEAN** the data
- Exploratory **DATA ANALYSIS**
- Statistical **PREDICTION/MODELING**
- **INTERPRET** Results
- **CHALLENGE** results
- Synthesize/**WRITE UP** results
- Create reproducible **CODE**

Prever o PIB

Mensal / Anual, etc, como explicar?

PIB (público),

Tráfego (ABCR) PIB (Site Itau),

Tráfego (Site ABCR)

Gráficos, Estatística descritiva

Análise de regressão

OBTENDO DADOS

google -> pib itau mensal
https://www.itau.com.br/itaubba-pt/analises-economicas/nossas-series-economicas/pib-mensal-itau-unibanco

google -> tráfego abcr download
http://www.abcr.org.br/Download.ashx?arquivo=IndiceABCR.xls

ANÁLISE EXPLORATÓRIA

- Verificar se os DADOS ESTÃO COMPLETOS
- GRÁFICOS
- ESTATÍSTICA descritiva
- CORRELAÇÃO É ALTA?

REGRESSÃO LINEAR

REGRESSÃO LINEAR

Chamaremos de modelo de regressão linear uma aplicação, com equação no seguinte formato:

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

A reta será uma aproximação, portanto, para cada ponto há uma diferença entre o ponto real e o ponto expressado por esta equação, Incluiremos, portanto, este fator de erro.

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$

PARA O MODELO:

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$

Obs 1: alfa será o mesmo para todos valores. Utilizamos a notação indicial, pois posteriormente incluiremos outros fatores alfa.

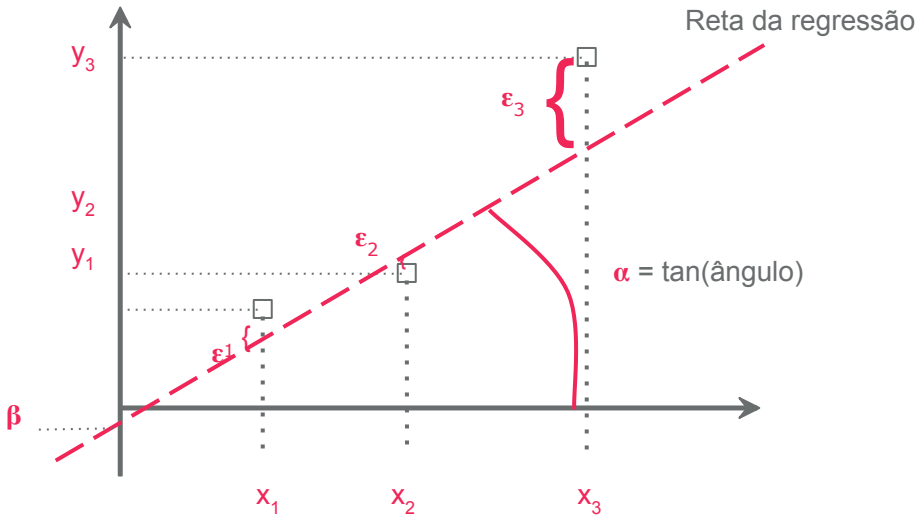
Obs 2: épslon tem índice i pois cada valor é independente, sendo correto incluir também o índice em y e x, portanto.

Obs 3: alfa, em geometria, é chamado de coeficiente angular e beta, coeficiente linear ou intercepto.

Obs 4: a variável x é dita independente e y dependente (depende de x). Ou ainda x exógena (exo=fora do modelo) e y endógena (endo=explicada dentro do modelo).

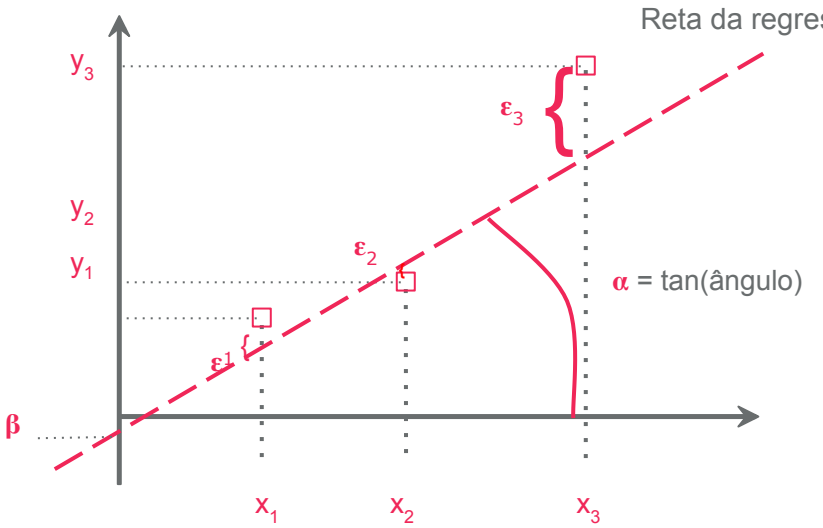
PARA O MODELO:

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$



PERG.: MAS COMO DETERMINAR OS COEFICIENTES ALFA E BETA, A PARTIR DOS DADOS?

Res.: Através de um método matemático chamado mínimos quadrados, onde procura-se minimizar os erros (epslon), daí o nome do método.



MÍNIMOS QUADRADOS

- DADO SEU RIGOR MATEMÁTICO acadêmico, a explicação do método foge do escopo da disciplina em nível de pós-graduação. Para maiores referências, sugere-se consultar:
  - http://livrommq.blogspot.com.br/

- Dentro do escopo da disciplina, utilizaremos o EXCEL como ferramenta de cálculo (que se utiliza do método em suas funções):
  - Suplemento ANÁLISE DE DADOS, Opção Regressão Linear
  - Funções: para beta, INTERCEPT (INTERCEPÇÃO) e; para alfa, SLOPE (INCLINAÇÃO)

PARA O MODELO:

$$y_i = a.x_i + þ + e_i$$

Representação em formato de equação:

$$y_1 = a.x_1 + þ + e_1$$

$$y_2 = a.x_2 + þ + e_2$$

$$y_n = a.x_n + þ + e_n$$

PARA O MODELO:

$$y_i = a.x_i + þ + e_i$$

Representação em formato tabular:

i	y_i	a	x_i	þ	e_i
1	y_1	a	x_1	þ	e_1
2	y_2	a	x_2	þ	e_2
-	-		-		-
n	y_n	a	x_n	þ	e_n

Obs: como os coeficientes alfa e beta são constantes, em geral os omitimos no formato tabular.

PARA O MODELO:

$$y_i = a.x_i + þ + s_i$$

Representação em formato tabular:

i	y_i	x_i	s_i
1	y_1	x_1	s_1
2	y_2	x_2	s_2
-	-	-	-
n	y_n	x_n	s_n

REGRESSÃO LINEAR MULTIPLA

Expandimos o modelo inserindo variáveis independentes adicionais:

$$y_i = a_1.x_{1i} + a_2.x_{2i} + þ + s_i$$

Ou ainda, generalizando para k variáveis:

$$y_i = a_1.x_{1i} + a_2.x_{2i} + \ldots + a_k.x_{ki} + þ + s_i$$

Em um modelo múltiplo (ou multivariado) são tratadas duas ou mais variáveis independentes, explicativas, ou ainda, exógenas.

MODELO AUTOREGRESSIVO

Um modelo AR utiliza a própria ”história” de uma variável para criar uma ”segunda dimensão”, dessa forma;

$$y_i = a_1.y_{i-1} + a_2.y_{i-2} + þ + s_i$$

No exemplo acima, criamos um modelo AR-2 (com duas variáveis). Abaixo, generalizamos para AR-K (modelo autoregressivo de ordem K):

$$y_i = a_1.y_{i-1} + a_2.y_{i-2} + \ldots + a_k.y_{i-k} + þ + s_i$$

Obs: aparentemente o modelo pode parecer complexo, mas veremos que sua aplicação é bem simples, na verdade, utilizando a representação tabular.

CONSTRUÇÃO 1/2

Expandiremos o modelo abaixo, para um AR-3:

$$y_i = a.y_{i-1} + þ + s_i$$

Representação em formato tabular:

i	y_i	y_{i-1}	y_{i-2}	y_{i-3}	s_i
1	y_1				s_1
2	y_2	y_1			s_2
3	y_3	y_2	y_1		s_3
4	y_4	y_3	y_2	y_1	s_4
n	y_n	y_{n-1}	y_{n-2}	y_{n-3}	s_n

CONSTRUÇÃO 2/2

Repare que a coluna y_{i-1} nada mais é do que a coluna y_i, deslocada uma linha para baixo. E y_{i-2} segue o mesmo comportamento, porém com deslocamento de duas linhas:

i	y_i	y_{i-1}	y_{i-2}	y_{i-3}	s_i
1	y_1				s_1
2	y_2	y_1			s_2
3	y_3	y_2	y_1		s_3
4	y_4	y_3	y_2	y_1	s_4
n	y_n	y_{n-1}	y_{n-2}	y_{n-3}	s_n

VARIÁVEIS DUMMY

- Uma variável dummy (também chamada indicador ou VARIÁVEL DE CATEGORIA OU BOOLEANA) é aplicada utilizando valor 0 ou 1 para indicar o que se propõe.

- É aplicável para:
  - Categorizar OUTLIERS;
  - Criar CATEGORIAS (fumante/não fumante, masc./fem., etc)
  - Sinalizar SAZONALIDADES em séries temporais
  - ENTRE OUTRAS categorizações

VARIÁVEL DUMMY PARA OUTLIER

Basta criar uma nova variável e, quando foi identificado um outlier (ver teste z), assinalar com 1.

i	y_i	x_i	d_i	s_i
1	y_1	x_1	0	s_1
2	y_2	x_2	1	s_2
3	y_3	x_3	0	s_3
n	y_n	x_n	0	s_n

Obs: no exemplo acima, somente x_2 foi identificado como outlier.

MODELO REGRESSIVO COM VARIÁVEL DUMMY

Expandiremos o modelo abaixo:

$$y_i = a_1.x_{1i} + a_2.x_{2i} + þ + s_i$$

Que ficará agora:

$$y_i = a_1.x_i + a_2.d_{1i} + þ + s_i$$

Onde $d_{1i} = \begin{cases} 0, \text{ se } x_i \text{ não for outlier} \\ 1, \text{ se } x_i \text{ for outlier} \end{cases}$

Obs: utilizamos aqui a notação indicial d_1 propositalmente, pois é comum utilizarmos várias variáveis *dummies*

VARIÁVEL DUMMY PARA SAZONALIDADES

Para captar o efeito sazonal, incluiremos uma variável dummy para cada evento sazonal.

Para o caso de identificarmos, por exemplo, que a cada mês do ano há uma sazonalidade (isto é, janeiros são similares ao longo do anos, o mesmo ocorrendo para fevereiro e todos demais meses) as dummies serão:

$$D_{jan} = \begin{cases} 1 & se\_for\_\text{mês}\_janeiro \\ 0 & caso\_\text{contrário} \end{cases}$$

$$D_{fev} = \begin{cases} 1 & se\_for\_\text{mês}\_fevereiro \\ 0 & caso\_\text{contrário} \end{cases}$$

$$\vdots$$

$$D_{nov} = \begin{cases} 1 & se\_for\_\text{mês}\_novembro \\ 0 & caso\_\text{contrário} \end{cases}$$

MODELO REGRESSIVO COM DUMMY SAZONAL

Nosso modelo será expandido para:

$$y_i = a_1.x_i + a_2.d_{1i} + a_2.d_{2i} + \ldots + a_{12}.d_{12i} + þ + s_i$$

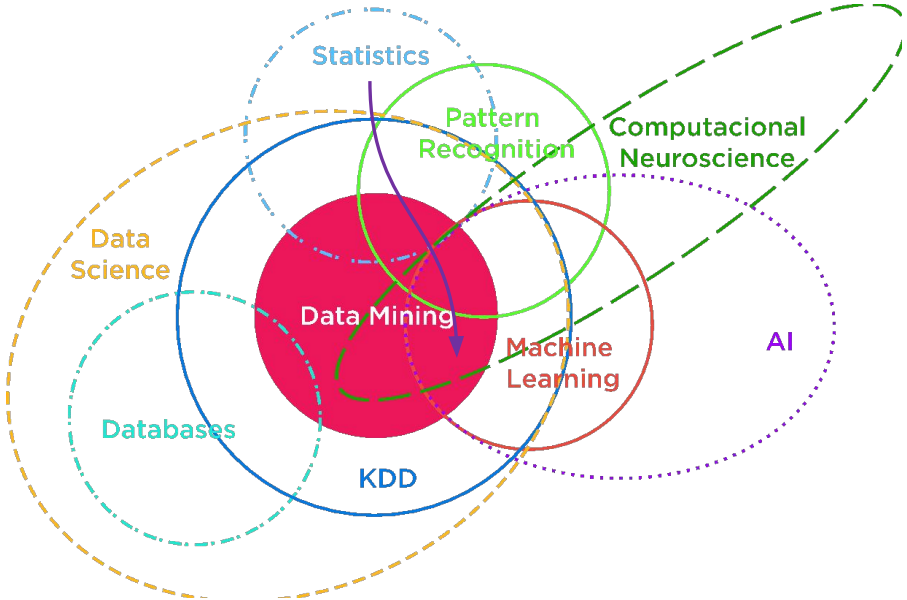
Pode parecer complexo a primeira vista, mas veremos no formato tabular que é simples.

MODELO REGRESSIVO COM DUMMY SAZONAL

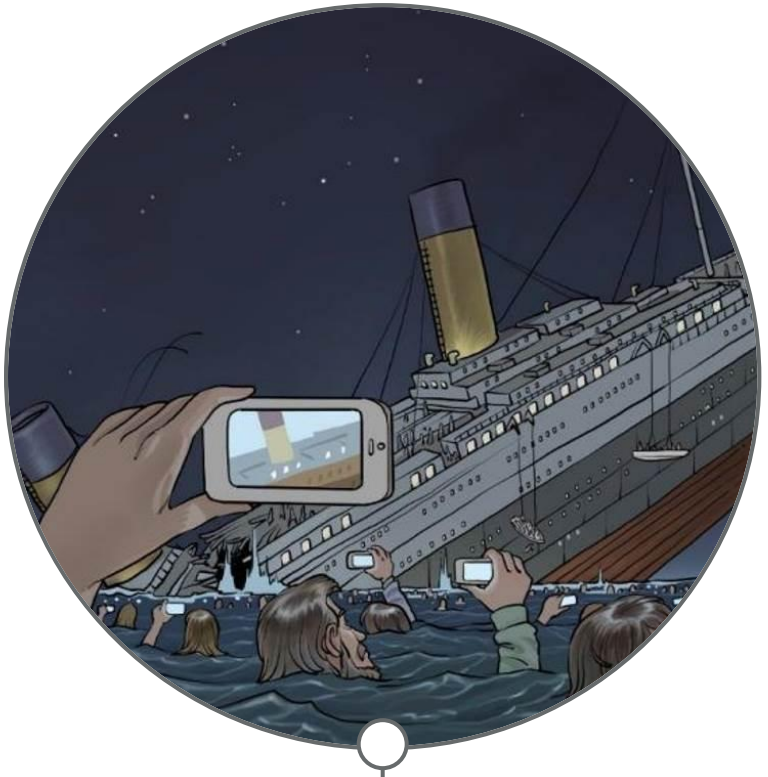
Basta criar uma coluna para cada evento sazonal

i	y_i	x_i	d_i	d_i	...	d_i	s_i
1	y_1	x_1	1	0		0	s_1
2	y_2	x_2	0	1		0	s_2
12	y_{12}	x_{12}	0	0		1	s_{12}
13	y_{13}	x_{13}	1	0		0	s_{13}
14	y_{14}	x_{14}	0	1		0	s_{14}
n	y_n	x_n	d_{1n}	d_{2n}		d_{12n}	s_n

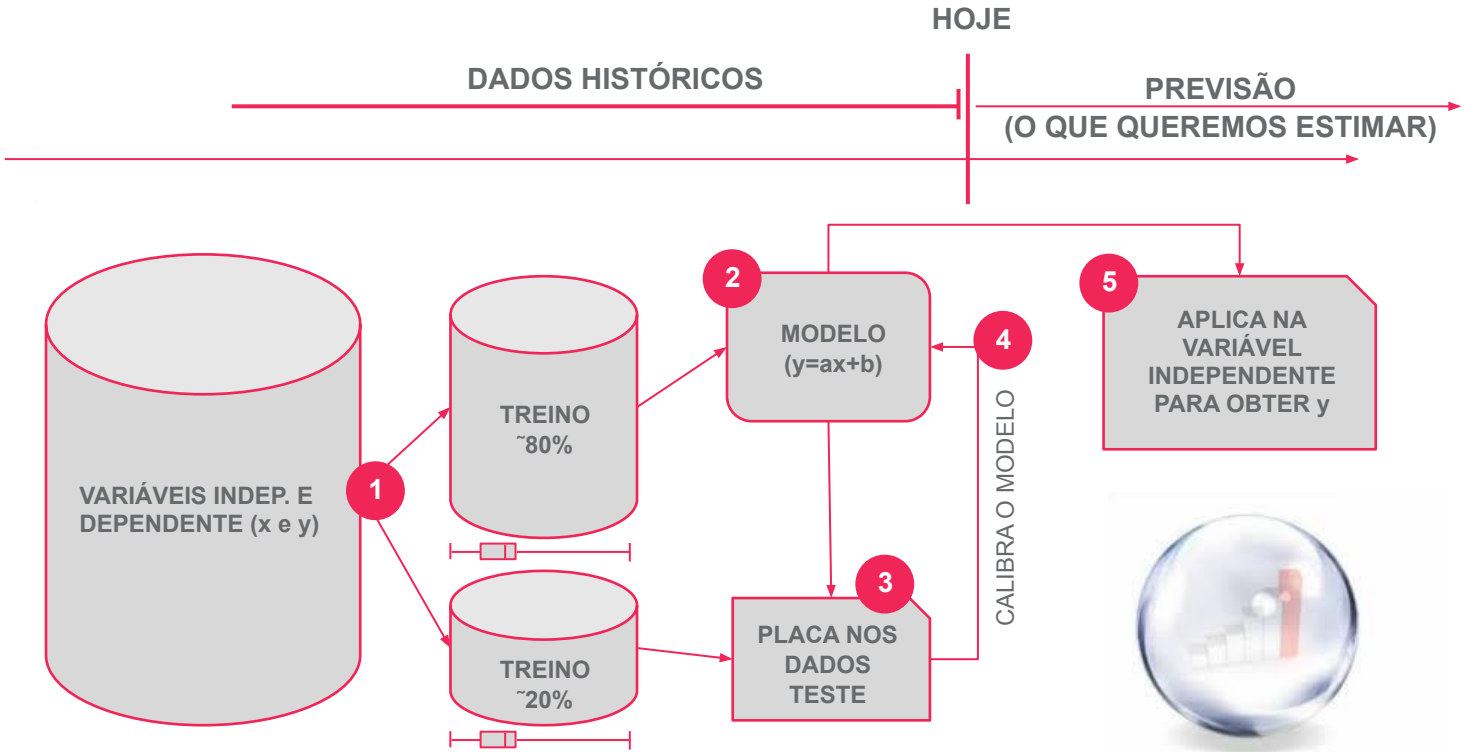
Por onde andamos no nosso ”PASSEIO”?



DESAFIO



QUAIS DADOS USAR NA ESTIMAÇÃO DO MODELO?



Obrigado

ProfDiogenes.Justo@fiap.com.br

SHIFT
FIAP