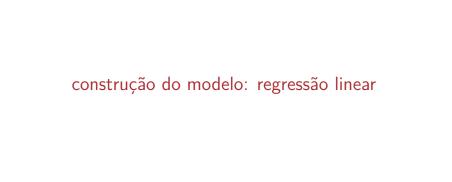
# Aula 5 - Projeto final

Jayme Anchante

1 de março de 2021



#### conceito

Modelo a relação linear entre uma variável resposta (alvo) e uma ou mais variáveis explicativas (características)

**Fonte** 

tipos

Regressão linear simples modela a relação entre uma resposta e uma característica

Regressão linear múltipla modela a relação entre uma resposta e duas ou mais características

#### otimização

- mínimos quadrados ordinários
- minimização do erro (absoluto ou quadrático)
- minimização de uma função de custo penalizadora (regressão ridge ou lasso)

## apresentação matemática

Sendo os dados  $\{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^n$  de n amostras/linhas e p variáveis/colunas. A relação linear é modelada por meio de um termo de erro  $\epsilon$ , uma variável aleatória não observável que adiciona/controla o ruído na relação de regressores e regressando tal que

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i, \quad i = 1, ..., n$$

# estimação por mínimos quadrados ordinários

Começando com a proposição inicial que

$$\hat{\beta} = \operatorname{argminS}(B) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$$

Depois de algumas transformações, chegamos em

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

estimação por mínimos quadrados ordinários: regressão simples

$$\begin{split} \hat{\beta} &= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} = \frac{\operatorname{Cov}[x, y]}{\operatorname{Var}[x]} \\ \hat{\alpha} &= \overline{y} - \hat{\beta} \, \overline{x} \;, \end{split}$$

# estimação ingênua

$$\hat{\beta} = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$$

- 1. Gerar valores (aleatórios) para os  $\beta$
- 2. Calcular a perda quadrática
- 3. Se a perda for menor, substituir os melhores parâmetros pelos atuais
- 4. Voltar ao passo 1

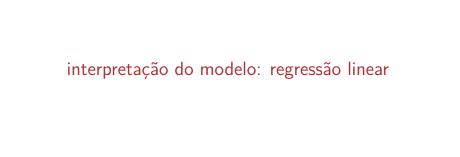
Pode ser incluído um critério de parada opcional baseado em um dos seguintes critérios: i) número de iterações; ii) erro mínimo aceitável; iii) número de iterações sem uma melhora no erro; iv) melhora marginal menor que um mínimo aceitável

#### exercício

Sendo os dados de X e y:

```
import numpy as np
rng = np.random.RandomState(42)
e = rng.random(100)
beta_0 = 3
beta_1 = 1.5
X = np.linspace(1, 100, 100)
y = beta 0 + (beta 1 * X) + e
```

- 1. Qual são os parâmetros que você deverá encontrar?
- 2. Plote os dados para explorar suas relações.
- 3. Escolha um dos métodos de estimação vistos anteriormente e estime os parâmetros?
- Compare suas respostas com sklearn.linear\_model.LinearRegression (dica: utilize o método coef\_ do estimador ajustado).



#### dados

Base de dados Guerry do pacote HistData do R. Andre-Michel Guerry (1833) foi o primeiro a sistematicamente coletar dados sociais. As variáveis utilizadas são Lottery (apostas per capita na loteria), Literacy (percentual de militares que são alfabetizados), Wealth (impostos recolhidos per capita), Region (região da França).

```
import statsmodels.api as sm
df = sm.datasets.get_rdataset("Guerry", "HistData").data
columns = ['Lottery', 'Literacy', 'Wealth', 'Region']
df = df[columns].dropna()
df.head()
```

## regressão linear

```
import statsmodels.formula.api as smf
formula = 'Lottery ~ Literacy + Wealth + Region'
mod = smf.ols(formula=formula, data=df)
res = mod.fit()
```

### resultados

print(res.summary())

## informações gerais

```
Dep. Variable:
                               Lottery
Model:
                                   01.5
Method:
                        Least Squares
Date:
                    Fri, 26 Feb 2021
Time:
                              00:56:46
No. Observations:
                                    85
Df Residuals:
                                    78
Df Model:
Covariance Type:
                             nonrobust
```

Variável dependente (alvo), data número de linhas/observações, df quer dizer degrees of freedom (graud de liberdade), tipo de covariância (existem diferentes especificações de covariância)

# informações de ajuste

```
R-squared: 0.338
Adj. R-squared: 0.287
F-statistic: 6.636
Prob (F-statistic): 1.07e-05
Log-Likelihood: -375.30
AIC: 764.6
BIC: 781.7
```

- R2: coeficiente de determinação é a proporção da variância da variável dependente que é explicada pelas variáveis explicativas.
   Ver mais.
- R2 ajustado: o mesmo que o anterior mas ajustado pelo número de colunas. Ver mais
- Estatística F: poder preditivo das variáveis explicativas.
- ▶ Probabilidade F: hipótese nula de que o modelo com intercepto e o modelo ajustado são iguais. Um valor menor que 0.05 rejeita esta hipótese com um grau de confiança de 5%.
- AIC (critério de informação de Akaike) e o BIC (critério de informação de Schwarz) são critérios para seleção de modelo. Quanto menor, melhor.

#### coeficientes

	coef	std err	t	P> t	[0.025	0.975]
Intercept	38.6517	9.456	4.087	0.000	19.826	57.478
Region[T.E]	-15.4278	9.727	-1.586	0.117	-34.793	3.938
Region[T.N]	-10.0170	9.260	-1.082	0.283	-28.453	8.419
Region[T.S]	-4.5483	7.279	-0.625	0.534	-19.039	9.943
Region[T.W]	-10.0913	7.196	-1.402	0.165	-24.418	4.235
Literacy	-0.1858	0.210	-0.886	0.378	-0.603	0.232
Wealth	0.4515	0.103	4.390	0.000	0.247	0.656

- Coeficiente é o valor do beta, o parâmetro que estamos buscando
- t: teste t cuja hipótese nula é de que o verdadeiro valor do parâmetro é zero
- P>|t|: caso seja menor que 0.05 rejeitamos a hipótese de que o parâmetro é zero (insignificante)
- Poderíamos escrever a equação como sendo:

$$y = 38.65 - 15.4Reg[= E] - 10Reg[= N] + ... - 0.18Lit + 0.45Wel$$

## coeficientes: intercepto

```
coef std err t P>|t| [0.025 0.975]

Intercept 38.6517 9.456 4.087 0.000 19.826 57.478
```

- Intercepto é o valor do alvo caso todas as variáveis independentes sejam zero
- ► Interpretação: caso região, alfabetização e riqueza sejam 0, o valor de apostas per capita é 38.65

# coeficientes: região

```
Region[T.E]
             -15.4278
                          9.727
                                   -1.586
                                              0.117
                                                        -34.793
                                                                     3.938
Region[T.N]
           -10.0170
                          9.260
                                   -1.082
                                              0.283
                                                        -28.453
                                                                     8.419
Region[T.S]
            -4.5483
                          7.279
                                   -0.625
                                              0.534
                                                        -19.039
                                                                     9.943
Region[T.W]
             -10.0913
                          7.196
                                   -1.402
                                              0.165
                                                        -24.418
                                                                     4.235
```

- Para fazermos interpretações, precisamos levar em conta sempre a categoria base (oculta)
- Pessoas da região S jogam, em média, -4.5 francos que pessoas da categoria base (região central), tudo o mais constante (ceteris paribus)

# coeficientes: riqueza

```
Literacy -0.1858 0.210 -0.886 0.378 -0.603 0.232 Wealth 0.4515 0.103 4.390 0.000 0.247 0.656
```

- Para variáveis numéricas, a interpretação é em função da própria unidade da variável
- A cada um franco a mais de riqueza per capita, a quantidade de francos colocados em aposta per capita é, em média, de 0.45, tudo o mais constante.
- Vemos que a variável alfabetização não é significativa a 5%, ao contrário da riqueza.



#### formato

- 1. 20min de apresentação por pessoa
- 2. 5min de perguntas da platéia
- 3. 10min de resposta do apresentador



# a vez de vocês





### assuntos pertinentes

- nuvem, devops, git, CI/CD, agilidade
- processamento distribuído, em GPUs
- aprendizado profundo para reconhecimento de image, vídeo, som e outras aplicações
- outras linguagens de programação, R, julia, shell
- praticar, praticar, praticar...