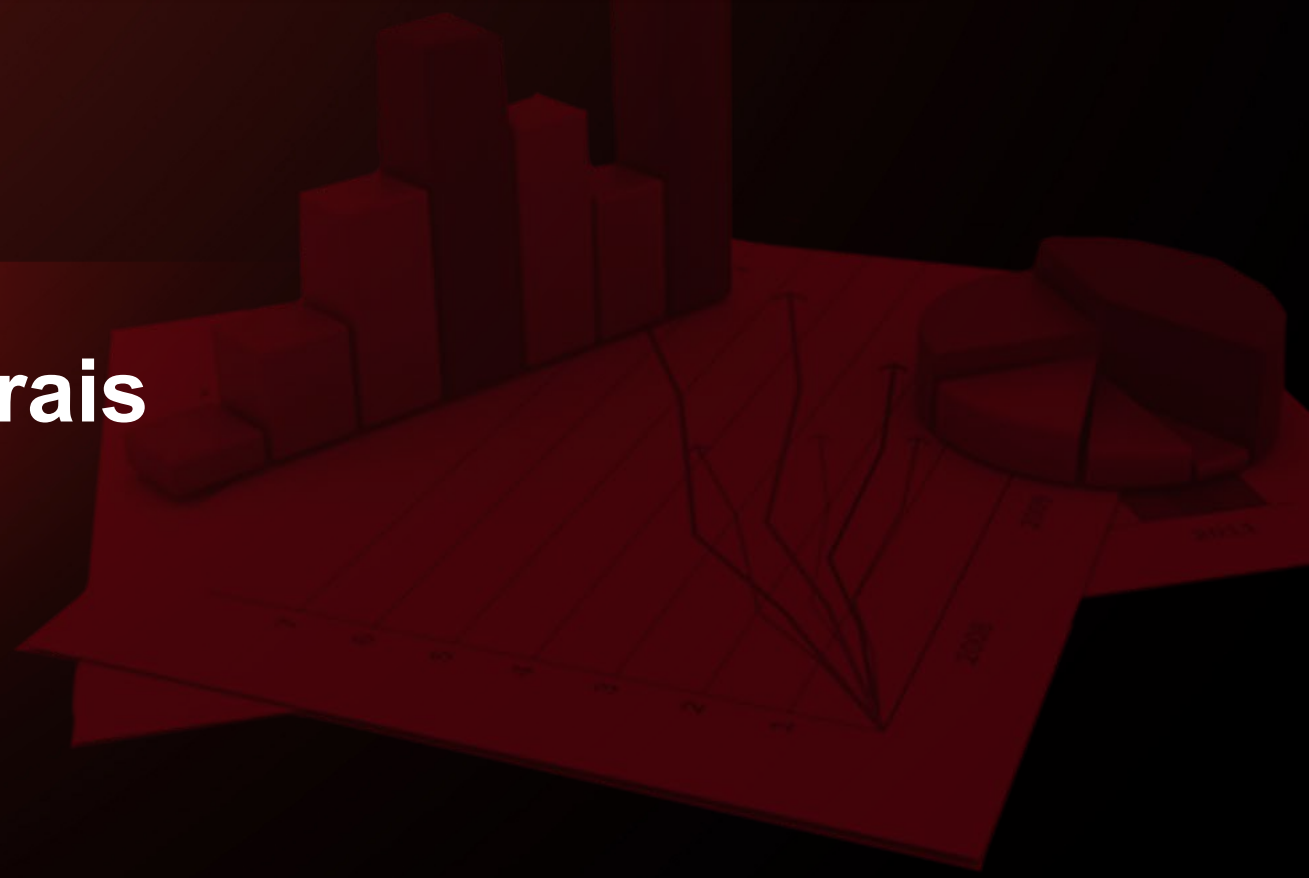
MODELAGEM E INFERÊNCIA ESTATÍSTICA

Métodos de regressão gerais e Modelo Logístico



O QUE VOU ESTUDAR HOJE?

Modelos não lineares

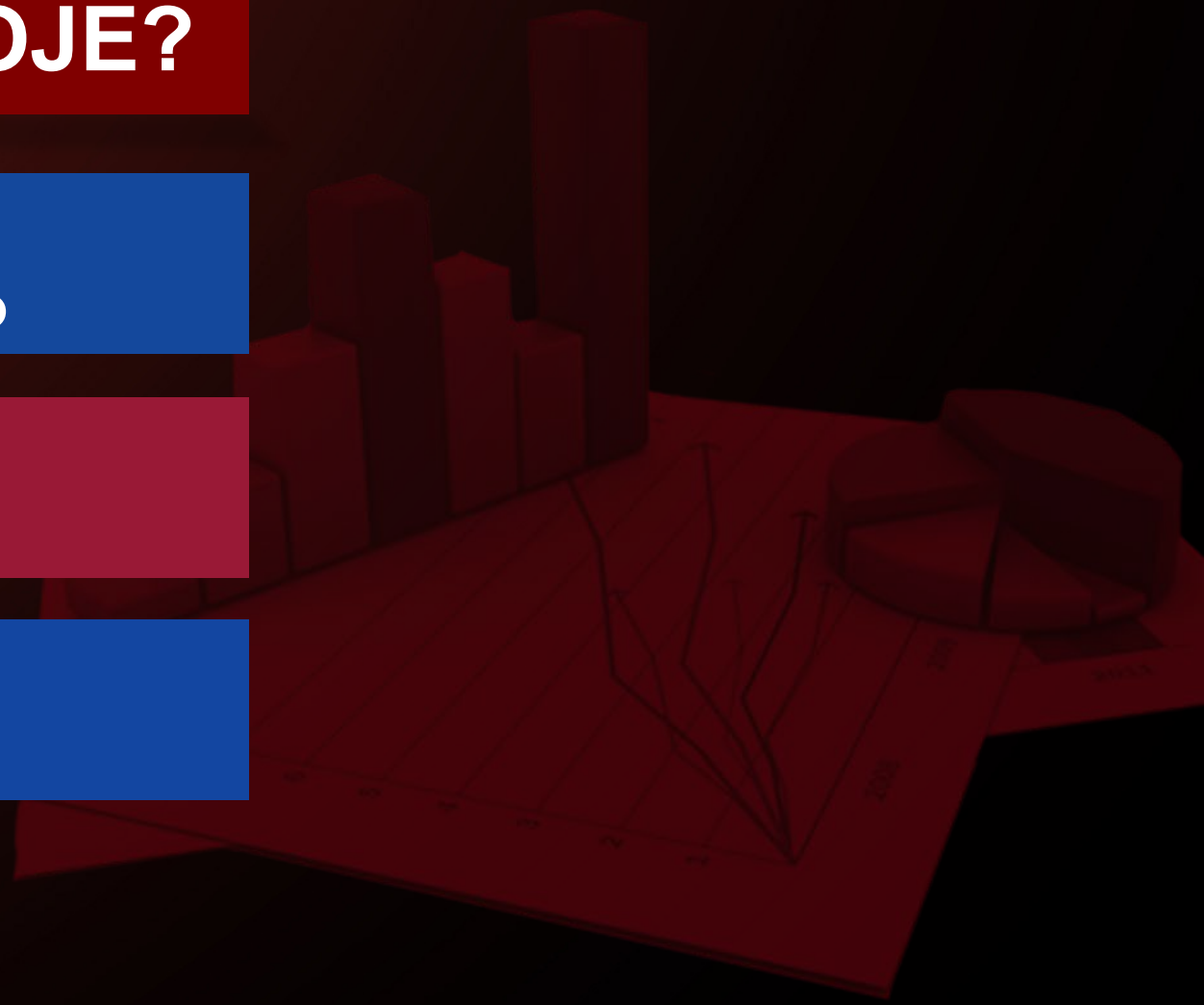
- Erro multiplicative e aditivo

Métodos de regressão gerais

- Reamostragem

Modelo Logístico

- Exemplo



MODELOS NÃO LINEARES PROBABILÍSTICOS

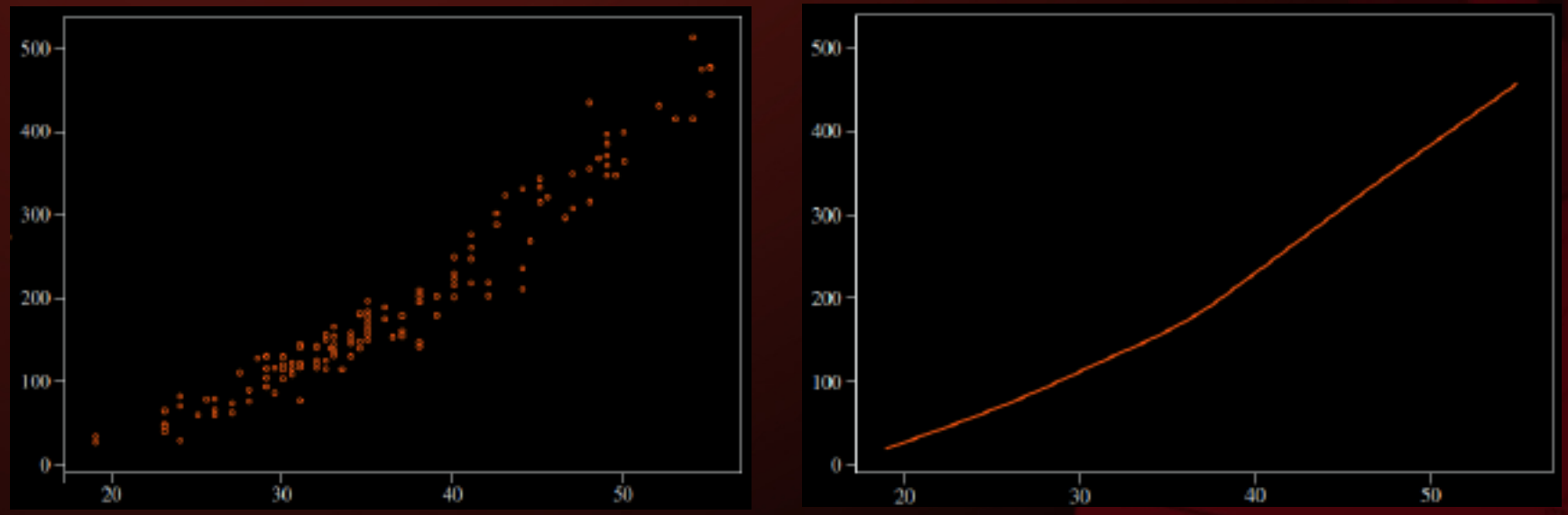
Função	Modelo probabilístico	Transformação para linearizar	Forma linear
Exponencial $y = \alpha e^{\beta x}$	$y = \alpha e^{\beta x} \varepsilon$	$y' = \ln(y)$	$y' = \ln(\alpha) + \beta x + \log(\varepsilon)$
Potência $y = \alpha x^{\beta}$	$y = \alpha x^{\beta} \varepsilon$	$y' = \log(y)$ e $x' = \log(x)$	$y' = \log(\alpha) + \beta x' + \log(\varepsilon)$
Logarítmica $y = \alpha + \beta \log(x)$	$y = \alpha + \beta \log(x) + \varepsilon$	$x' = \log(x)$	$y = \alpha + \beta x' + \varepsilon$
Recíproca $y = \alpha + \beta \frac{1}{x}$	$y = \alpha + \beta \frac{1}{x} + \varepsilon$	$x' = \frac{1}{x}$	$y = \alpha + \beta x' + \varepsilon$

ε multiplicativo

ε aditivo

MÉTODOS DE REGRESSÃO GERAIS

Reamostragem



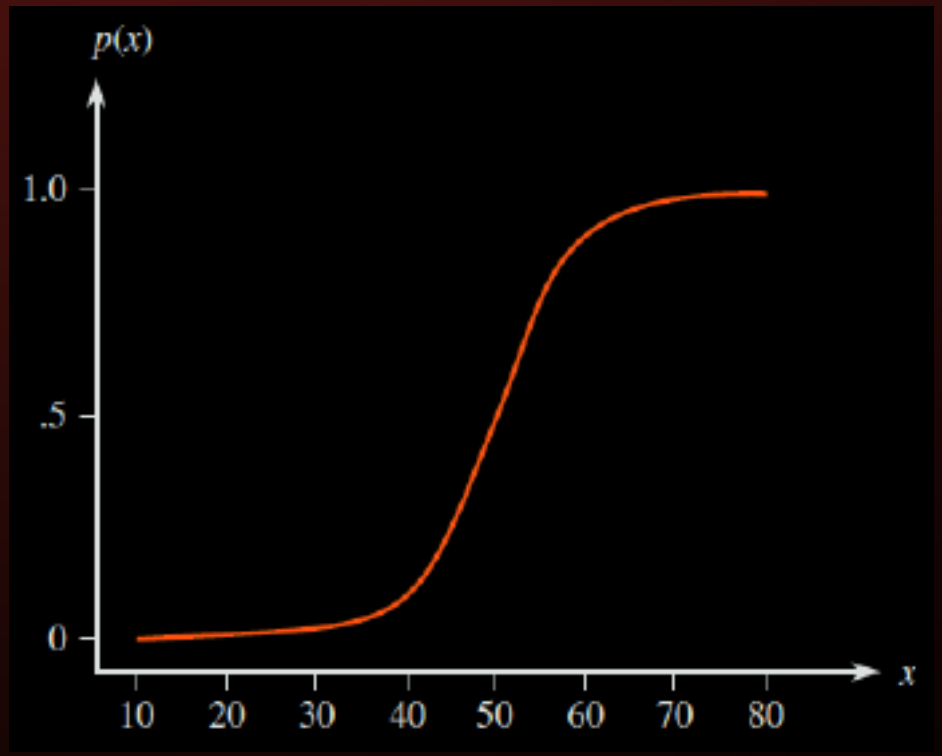
Fonte: (DEVORE, 2018, p. 520-521)

- Jackknife.
- Bootstrap (paramétrico e não paramétrico).

MODELO LOGÍSTICO

- **Resposta dicotômica *p* com valores possíveis 1 e 0 correspondendo ao sucesso e à falha.**

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$



Fonte: (DEVORE, 2018, p. 522)

MODELO LOGÍSTICO

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

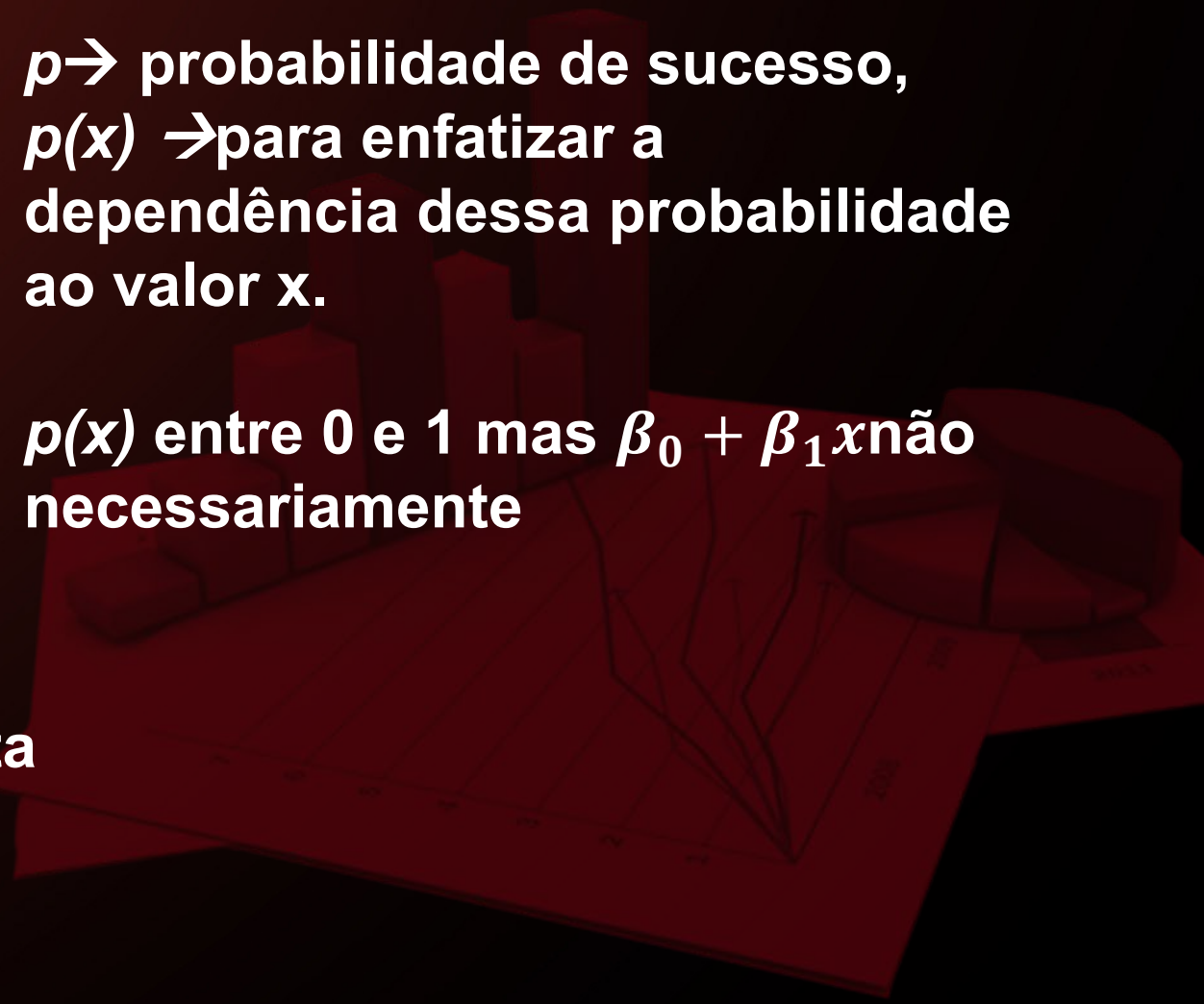
Razão de *chances*

Logaritmo da razão de *chances* é uma reta

(função linear do preditor)

p→ probabilidade de sucesso,
p(x) →para enfatizar a dependência dessa probabilidade ao valor x.

p(x) entre 0 e 1 mas $\beta_0 + \beta_1 x$não necessariamente



MODELO LOGÍSTICO

Exemplo 1

O termo cifose refere-se a uma grave curvatura protuberante da coluna vertebral que necessita de cirurgia corretiva. Um estudo realizado para determinar os fatores de risco da cifose relatou as idades a seguir (meses) para 40 indivíduos no momento da cirurgia; os primeiros 18 indivíduos tiveram cifose e os 22 restantes, não.

Cifose	12	15	42	52	59	73	82	91	96
	105	114	120	121	128	130	139	139	157
Sem Cifose	1	1	2	8	11	18	22	31	37
	61	72	81	97	112	118	127	131	140
	151	159	177						

MODELO LOGÍSTICO

Exemplo 1

Use o resultado da regressão logística do Python, abaixo, para determinar se a idade parece ter um impacto significativo sobre a existência da cifose.

```
Optimization terminated successfully.
         Current function value: 0.681303
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   40
Model:                          Logit   Df Residuals:                       38
Method:                           MLE   Df Model:                            1
Date:                Tue, 08 Mar 2022   Pseudo R-squ.:                0.009934
Time:                        20:20:37   Log-Likelihood:                -27.252
converged:                       True   LL-Null:                       -27.526
Covariance Type:            nonrobust   LLR p-value:                    0.4596
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5727      0.602     -0.951      0.342      -1.753       0.608
x              0.0043      0.006      0.734      0.463      -0.007       0.016
==============================================================================
```

MODELO LOGÍSTICO

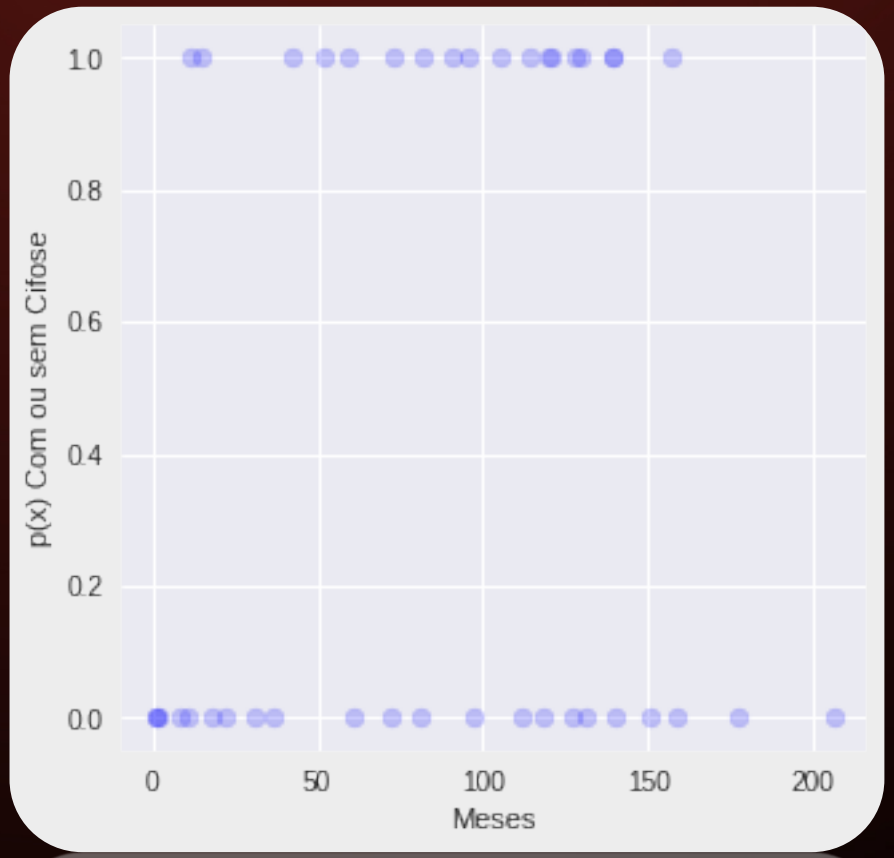
Antes de resolver o exercício

Definir os dados:
Com cifose = 1
Sem cifose = 0

```
lstx = (12,15,42,52,59,73,82,91,96,105,114,120,121,128,130,
        139,139,157,1,1,2,8,11,18,22,31,37,61,72,81,97,112,
        118,127,131,140,151,159,177,206)
lsty= (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,
       0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
# Construir o DataFrame e nomear as colunas
df = pd.DataFrame(list(zip(lstx, lsty)),
                  columns =["x","y"])
x=df['x']
y=df['y']
```

MODELO LOGÍSTICO

Antes de resolver o exercício



Determinar se a idade parece ter um impacto significativo sobre a existência da cifose.

MODELO LOGÍSTICO

Antes de resolver o exercício

Usar a biblioteca statmodel

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 import statsmodels.api as sm
5 from statsmodels.formula.api import ols
6 import seaborn as sns
7 from statsmodels.graphics.gofplots import ProbPlot
8 plt.style.use('seaborn')
9 plt.rc('axes', titlesize=10)
```

Obter os parâmetros da regressão

```
1 #adicionar uma constante preditora
2 x = sm.add_constant(x)
3 # Construir o modelo e ajustar os dados
4 model = sm.Logit(y, x).fit()
5 print(model.summary())
```

Resultados

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5727      0.602     -0.951      0.342      -1.753       0.608
x              0.0043      0.006      0.734      0.463      -0.007       0.016
==============================================================================
```

MODELO LOGÍSTICO

Resolução

Interpretação

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

$$\frac{p(x)}{1-p(x)} = e^{-0,5757+0,0043x}$$

```
Optimization terminated successfully.
         Current function value: 0.681303
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   40
Model:                          Logit   Df Residuals:                       38
Method:                           MLE   Df Model:                            1
Date:                Tue, 08 Mar 2022   Pseudo R-squ.:                0.009934
Time:                        20:20:37   Log-Likelihood:                -27.252
converged:                       True   LL-Null:                       -27.526
Covariance Type:            nonrobust   LLR p-value:                    0.4596
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5727      0.602     -0.951      0.342      -1.753       0.608
x              0.0043      0.006      0.734      0.463      -0.007       0.016
==============================================================================
```

MODELO LOGÍSTICO

Resolução

Determinar se a idade parece ter um impacto significativo sobre a existência da cifose

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5727	0.602	-0.951	0.342	-1.753	0.608
x	0.0043	0.006	0.734	0.463	-0.007	0.016

$$\frac{p(x)}{1-p(x)} = e^{-0,5757+0,0043x}$$

Realizar um teste de hipótese para verificar a relação entre as variáveis

MODELO LOGÍSTICO

Resolução

Lembrete: Procedimento do teste de hipótese para utilidade do modelo

1. Obter a reta de regressão $y = \beta_0 + \beta_1 x$
2. Definir o valor de n (número de amostras).
3. Definir o valor de k (número de variáveis).
4. Calcular os graus de liberdade gl = n-k.
5. $H_0: \beta_1 = 0$ frente a $H_a: \beta_1 \neq 0$

$$\frac{p(x)}{1-p(x)} = e^{-0,5757+0,0043x}$$

$$n = 40$$
$$k = 2$$
$$gl = 38$$

$H_0: \beta_1 = 0$ → neste modelo $\beta_1 = 0,0043$ frente a $H_a: \beta_1 \neq 0$

MODELO LOGÍSTICO

Resolução

$$\frac{p(x)}{1-p(x)} = e^{-0{,}5757+0{,}0043x}$$

Lembrete: Procedimento do teste de hipótese para β_1

1. Definir a hipótese nula $H_0: \beta_{10} = 0$ frente a $H_a: \beta_{10} \neq 0$.
2. Definir o intervalo de confiança α se conhecido.
3. Definir o intervalo crítico (t_{crit}) na tabela t-student para um determinado α e gl.
4. Definir a estatística de teste $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$.

   Como $H_0: \beta_{10} = 0 \rightarrow t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.
5. Se $|t| \geq t_{crit}$ rejeitar H_0 em favor de $H_a: \beta_{10} \neq 0$.
6. Caso contrario comparar se p-valor < α.

$\alpha = 0{,}05$ *e* $gl = 38$

$t_{crit} = t_{(0{,}025,38)}$

```
1 #usar a tabela tstudent
2 from scipy.stats import t
3 alpha = 0.05  # nível de signif.= 5%
4 df = len(x) - 2 # graus de liberdade
5 #
6 v = t.ppf(1 - alpha/2, df)
7 tcrit=v
8 print(f'tcrit=: {v}')
```

```
tcrit=: 2.024394164575136
```

MODELO LOGÍSTICO

Resolução

Lembrete: Procedimento do teste de hipótese para β_1

1. Definir a hipótese nula $H_0: \beta_{10} = 0$ frente a $H_a: \beta_{10} \neq 0$.
2. Definir o intervalo de confiança α se conhecido.
3. Definir o intervalo crítico (t_{crit}) na tabela t-student para um determinado α e gl.
4. Definir a estatística de teste $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$.

   Como $H_0: \beta_{10} = 0 \rightarrow t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.
5. Se $|t| \geq t_{crit}$ rejeitar H_0 em favor de $H_a: \beta_{10} \neq 0$.
   Ou comparar se p-valor < α.

$$t_{crit} = t_{(0,025,38)} = 2,0244$$

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 0,734$$

Se $|t| \geq t_{crit}$ rejeitar H_0

$0,734 \geq 2,0244$??

Não rejeitar $H_0: \beta_{10} = 0$

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5727	0.602	-0.951	0.342	-1.753	0.608
x	0.0043	0.006	0.734	0.463	-0.007	0.016

MODELO LOGÍSTICO

Resolução

Outra opção: comparar se p-valor < α.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5727	0.602	-0.951	0.342	-1.753	0.608
x	0.0043	0.006	0.734	0.463	-0.007	0.016

```
1 #z: estatística de teste
2 zt=model.tvalues
3 ztb1=zt[1]
4 ztb1
```

```
0.7344029072729139
```

```
1 #usar a tabela tstudent
2 from scipy.stats import t
3 #calculate p-value  TWO TAILED
4 p_val= (1-t.cdf(x=abs(ztb1), df=len(x))) * 2
5 p_val
```

```
0.46698681622739535
```

Valor-p=0,463 e $\alpha = 0,05$→ 0,463 <$0,05$??

Não rejeitar $H_0: \beta_{10} = 0$

MODELO LOGÍSTICO

Exemplo 2

O artigo *Acceptable noise levels for construction site offices* (Building Serv. Engr. Res. Tech., 2009: 87-94) analisou as respostas de uma amostra com 77 indivíduos, aos quais foi perguntado se um nível de ruído específico (dBA) em que já havia sido exposto era aceitável ou inaceitável. Eis os dados oferecidos pelos autores do artigo:

Acceptable:

55.3	55.3	55.3	55.9	55.9	55.9	55.9	56.1	56.1	56.1	56.1
56.1	56.1	56.8	56.8	57.0	57.0	57.0	57.8	57.8	57.8	57.9
57.9	57.9	58.8	58.8	58.8	59.8	59.8	59.8	62.2	62.2	65.3
65.3	65.3	65.3	68.7	69.0	73.0	73.0				

Unacceptable:

63.8	63.8	63.8	63.9	63.9	63.9	64.7	64.7	64.7	65.1	65.1
65.1	67.4	67.4	67.4	67.4	68.7	68.7	68.7	70.4	70.4	71.2
71.2	73.1	73.1	74.6	74.6	74.6	74.6	79.3	79.3	79.3	79.3
79.3	83.0	83.0	83.0							

MODELO LOGÍSTICO

Exemplo 2

Interprete o resultado da regressão logística e esboce um gráfico da probabilidade de um nível de ruído aceitável como uma função do nível de ruído específico.

```
Optimization terminated successfully.
         Current function value: 0.353003
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   77
Model:                          Logit   Df Residuals:                       75
Method:                           MLE   Df Model:                            1
Date:                Thu, 10 Mar 2022   Pseudo R-squ.:                  0.4902
Time:                        22:17:49   Log-Likelihood:                -27.181
converged:                       True   LL-Null:                       -53.314
Covariance Type:            nonrobust   LLR p-value:                 4.849e-13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         23.0140      5.041      4.565      0.000      13.133      32.895
x             -0.3562      0.078     -4.543      0.000      -0.510      -0.203
==============================================================================
```

MODELO LOGÍSTICO

Antes de resolver o exercício

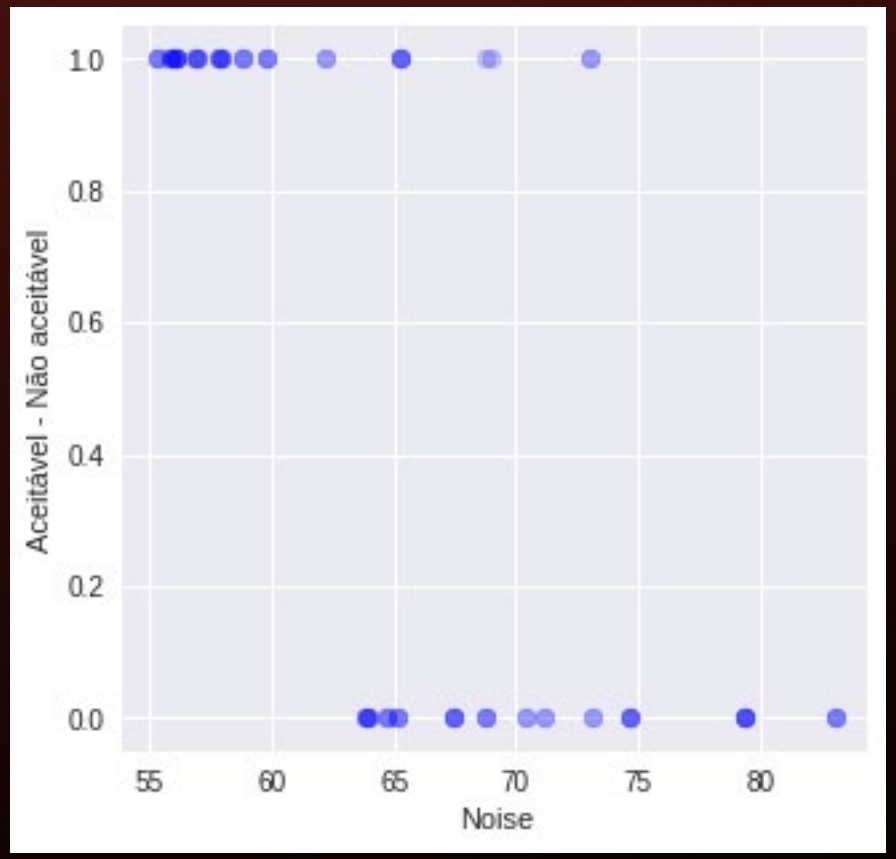
Definir os dados:
Aceitável = 1
Não aceitável = 0

```
lstxr = (55.3,55.3,55.3,55.9,55.9,55.9,55.9,56.1,56.1,56.1,56.1,56.1,56.1,56.8,
         56.8,57.0,57.0,57.0,57.8,57.8,57.8,57.9,57.9,57.9,58.8,58.8,58.8,59.8,
         59.8,59.8,62.2,62.2,65.3,65.3,65.3,65.3,68.7,69.0,73.0,73.0,63.8,63.8,
         63.8,63.9,63.9,63.9,64.7,64.7,64.7,65.1,65.1,65.1,67.4,67.4,67.4,67.4,
         68.7,68.7,68.7,70.4,70.4,71.2,71.2,73.1,73.1,74.6,74.6,74.6,74.6,79.3,
         79.3,79.3,79.3,79.3,83.0,83.0,83.0)
lstyr= (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
         1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
         0,0,0,0,0,0,0)
# Construir o DataFrame e nomear as colunas
df1 = pd.DataFrame(list(zip(lstxr, lstyr)),
                   columns =["x","y"])
x=df1['x']
y=df1['y']
df.head()
```

MODELO LOGÍSTICO

Antes de resolver o exercício

Observar o gráfico dos dados



MODELO LOGÍSTICO

Resolução

Interprete o resultado da regressão logística e esboce um gráfico da probabilidade de um nível de ruído aceitável como uma função do nível de ruído específico.

```
Optimization terminated successfully.
         Current function value: 0.353003
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   77
Model:                          Logit   Df Residuals:                       75
Method:                           MLE   Df Model:                            1
Date:                Thu, 10 Mar 2022   Pseudo R-squ.:                  0.4902
Time:                        22:17:49   Log-Likelihood:                -27.181
converged:                       True   LL-Null:                       -53.314
Covariance Type:            nonrobust   LLR p-value:                 4.849e-13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         23.0140      5.041      4.565      0.000      13.133      32.895
x             -0.3562      0.078     -4.543      0.000      -0.510      -0.203
==============================================================================
```

MODELO LOGÍSTICO

Resolução

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

$$\frac{p(x)}{1-p(x)} = e^{23{,}0140-0{,}3562x}$$

```
Optimization terminated successfully.
         Current function value: 0.353003
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   77
Model:                          Logit   Df Residuals:                       75
Method:                           MLE   Df Model:                            1
Date:                Thu, 10 Mar 2022   Pseudo R-squ.:                 0.4902
Time:                        22:17:49   Log-Likelihood:                -27.181
converged:                       True   LL-Null:                       -53.314
Covariance Type:            nonrobust   LLR p-value:                 4.849e-13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         23.0140      5.041      4.565      0.000      13.133      32.895
x             -0.3562      0.078     -4.543      0.000      -0.510      -0.203
==============================================================================
```

MODELO LOGÍSTICO

Resolução

Lembrete: Procedimento do teste de hipótese para β_1

1. Definir a hipótese nula $H_0: \beta_{10} = 0$ frente a $H_a: \beta_{10} \neq 0$.
2. Definir o intervalo de confiança α se conhecido.
3. Definir o intervalo crítico (t_{crit}) na tabela t-student para um determinado α e gl.
4. Definir a estatística de teste $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$.

   Como $H_0: \beta_{10} = 0 \rightarrow t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.
5. Se $|t| \geq t_{crit}$ rejeitar H_0 em favor de $H_a: \beta_{10} \neq 0$.
   Ou comparar se p-valor < α.

	coef	std err	z	P>|z|	[0.025	0.975]
const	23.0140	5.041	4.565	0.000	13.133	32.895
x	-0.3562	0.078	-4.543	0.000	-0.510	-0.203

Valor-p=0,000 e $\alpha = 0,05$
→0,000 <0,05

Rejeitar $H_0: \beta_{10} = 0$

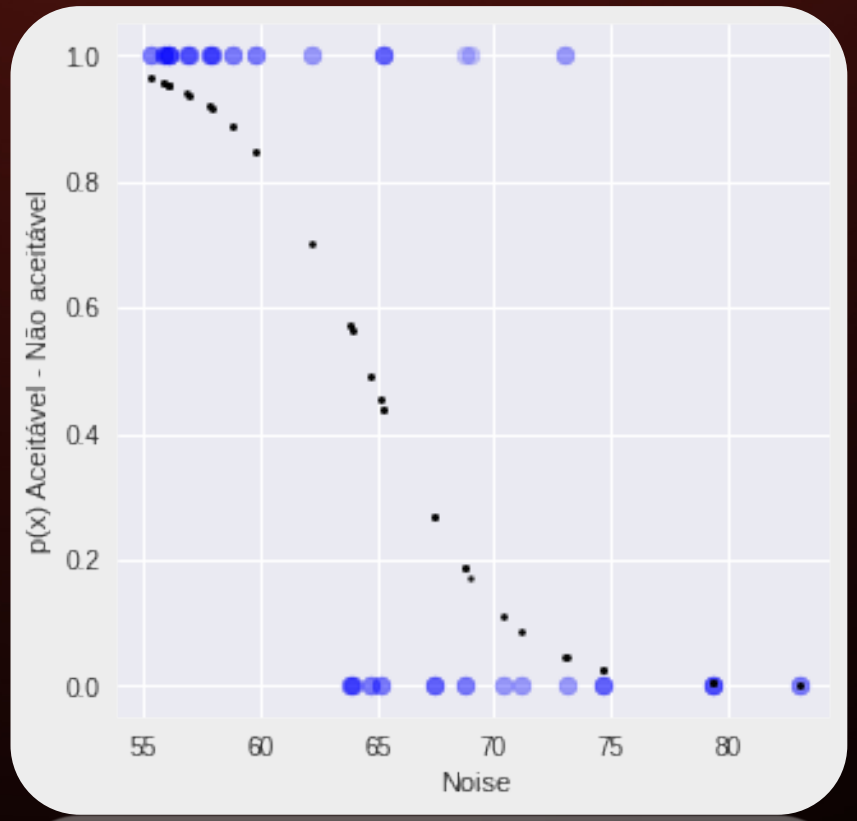
Aceitar o modelo

$$\frac{p(x)}{1-p(x)} = e^{23,0140 - 0,3562x}$$

MODELO LOGÍSTICO

Resolução

Valores observados e valores calculados




```
1 x=df['x']
2 y=df['y']
3 plt.figure(figsize=(5, 5))
4 ax = plt.axes()
5 ax.scatter(x, y, color='b', alpha=0.20)
6 ax.scatter(x, yhat, color="black", s=4)
7 ax.set_xlabel('Noise')
8 ax.set_ylabel('Aceitável - Não aceitável')
```

MODELO LOGÍSTICO

Resolução

$$\frac{p(x)}{1-p(x)} = e^{23{,}0140-0{,}3562x}$$

```
1 #Obter a razão das chances
2 odd=np.exp(modelo.params)
3 odd[1]
```

```
0.70033069911172532
```

Razão das *chances*

$$\frac{\frac{p(x+1)}{1-p(x+1)}}{\frac{p(x)}{1-p(x)}} = \frac{e^{23{,}0140-0{,}3562(x+1)}}{e^{23{,}0140-0{,}3562x}} = e^{-0{,}3562} = 0,7$$

A interpretação é que, para cada incremento de ruído, estima-se que as chances de ter um ruído não aceitável (0) irão diminuir por um fator de 0,7 (30%).

MODELO LOGÍSTICO

Resolução

Para calcular manualmente os valores estimados

$$\hat{\pi} = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

$$\hat{\pi} = \frac{e^{23,0140-0,3562x}}{1 + e^{23,0140-0,3562x}}$$

Valores observados e valores calculados

```
1 # performing predictions on the test datdaset
2 yhat = model.predict()
3 prediction = list(map(round, yhat))
4
5 # comparing original and predicted values of y
6 print('Valores observados:', list(y))
7 print('Valores previstos:', prediction)
```

```
Valores observados: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
Valores previstos: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

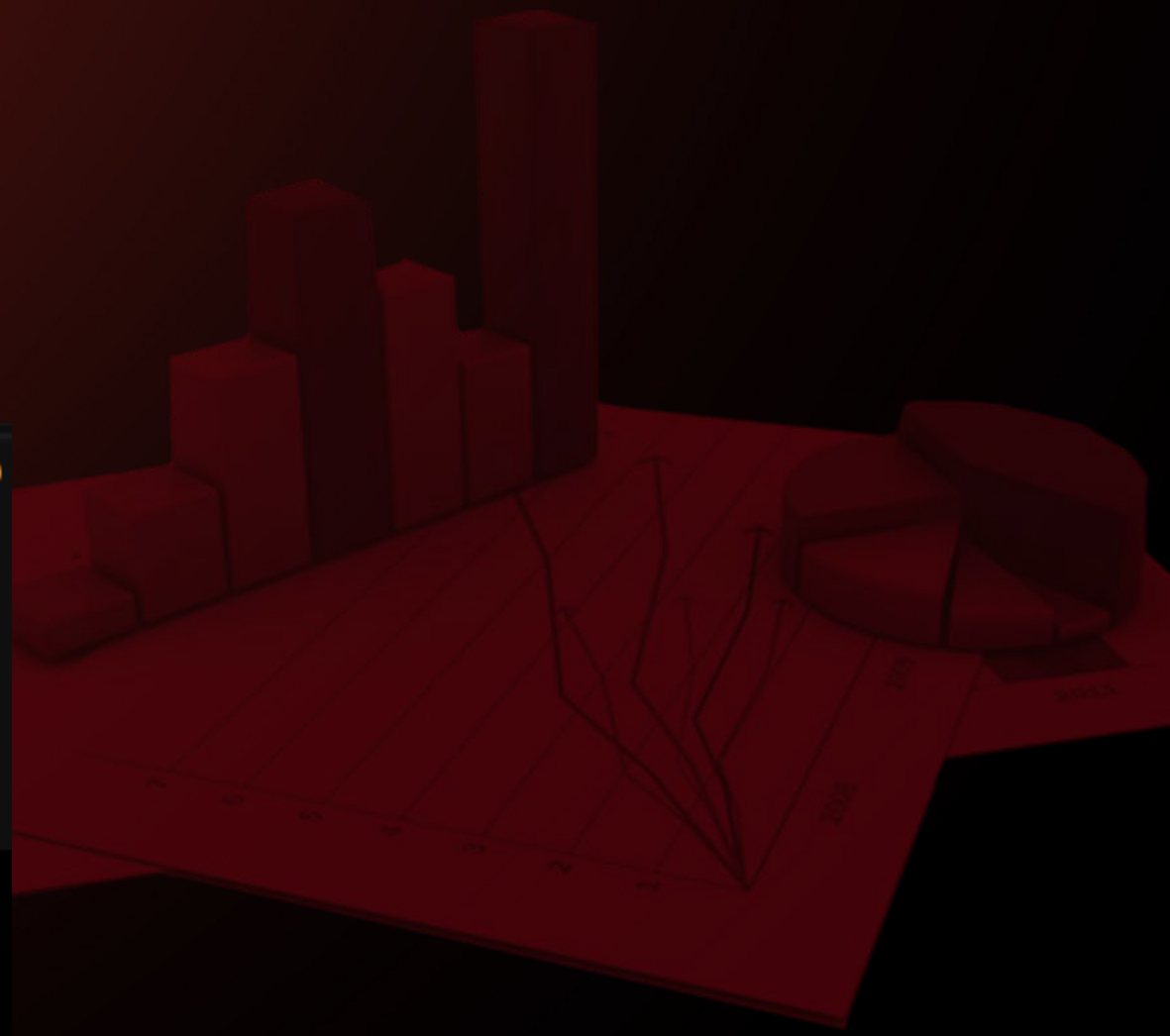
MODELO LOGÍSTICO

Comentários complementares

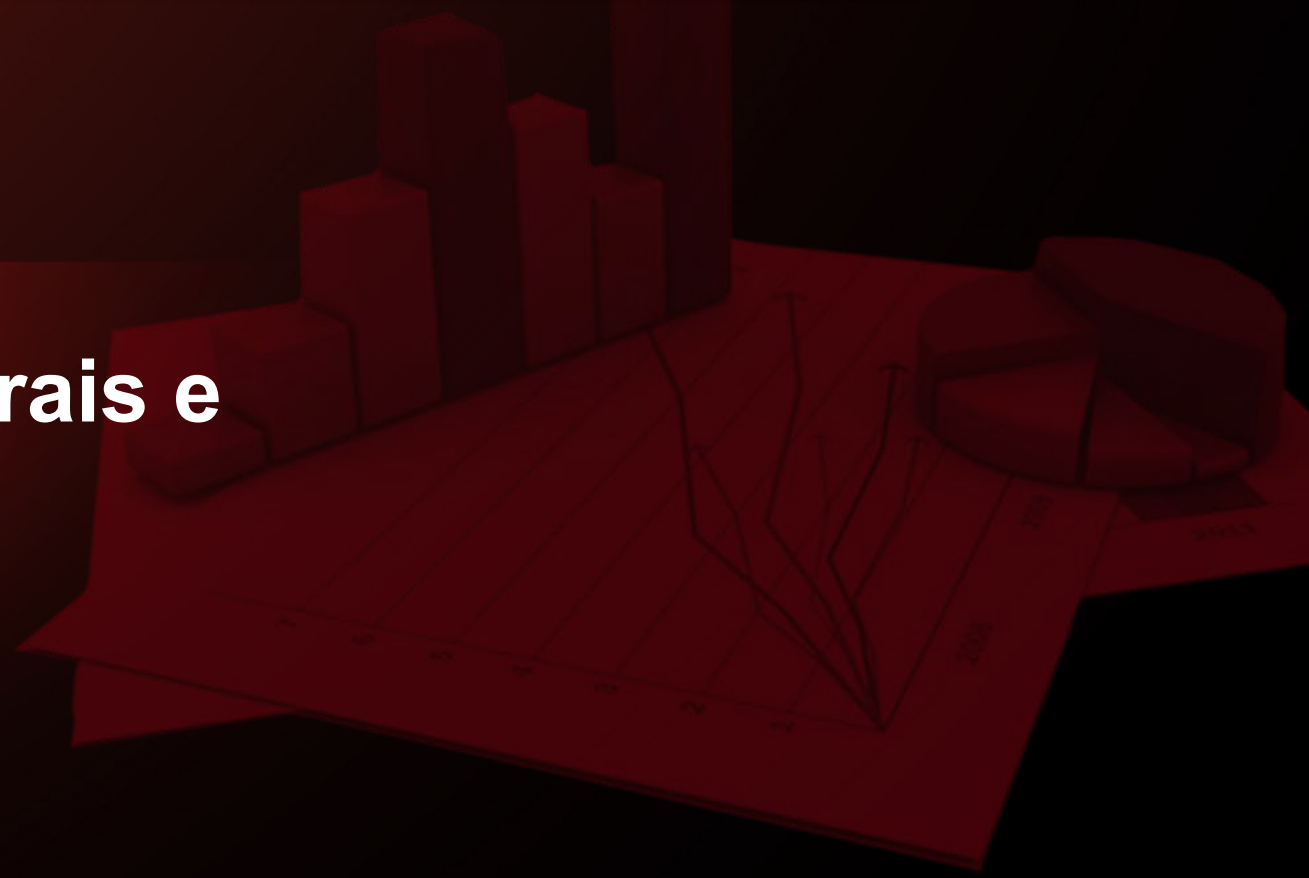
Verificar a acurácia do modelo

```
1 from sklearn.metrics import (confusion_matrix,accuracy_score)
2
3 # confusion matrix
4 cm = confusion_matrix(y, prediction)
5 print ("Confusion Matrix : \n", cm)
6
7 # accuracy score of the model
8 print('Test accuracy = ', accuracy_score(y, prediction))
```

```
Confusion Matrix :
 [[31  6]
 [ 8 32]]
Test accuracy =  0.8181818181818182
```



MODELAGEM E INFERÊNCIA ESTATÍSTICA



Métodos de regressão gerais e Modelo Logístico