# MODELAGEM E INFERÊNCIA ESTATÍSTICA

## Modelos polinomiais

# O QUE VOU ESTUDAR HOJE?

- Modelo polinomial
- Variância estimada e coeficiente de determinação
- Intervalos estatísticos e procedimentos de teste
- Centralização de valores

# MODELO POLINOMIAL

O modelo de regressão polinomial de k-ésimo grau é

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

em que $\varepsilon$ é uma variável aleatória com distribuição normal

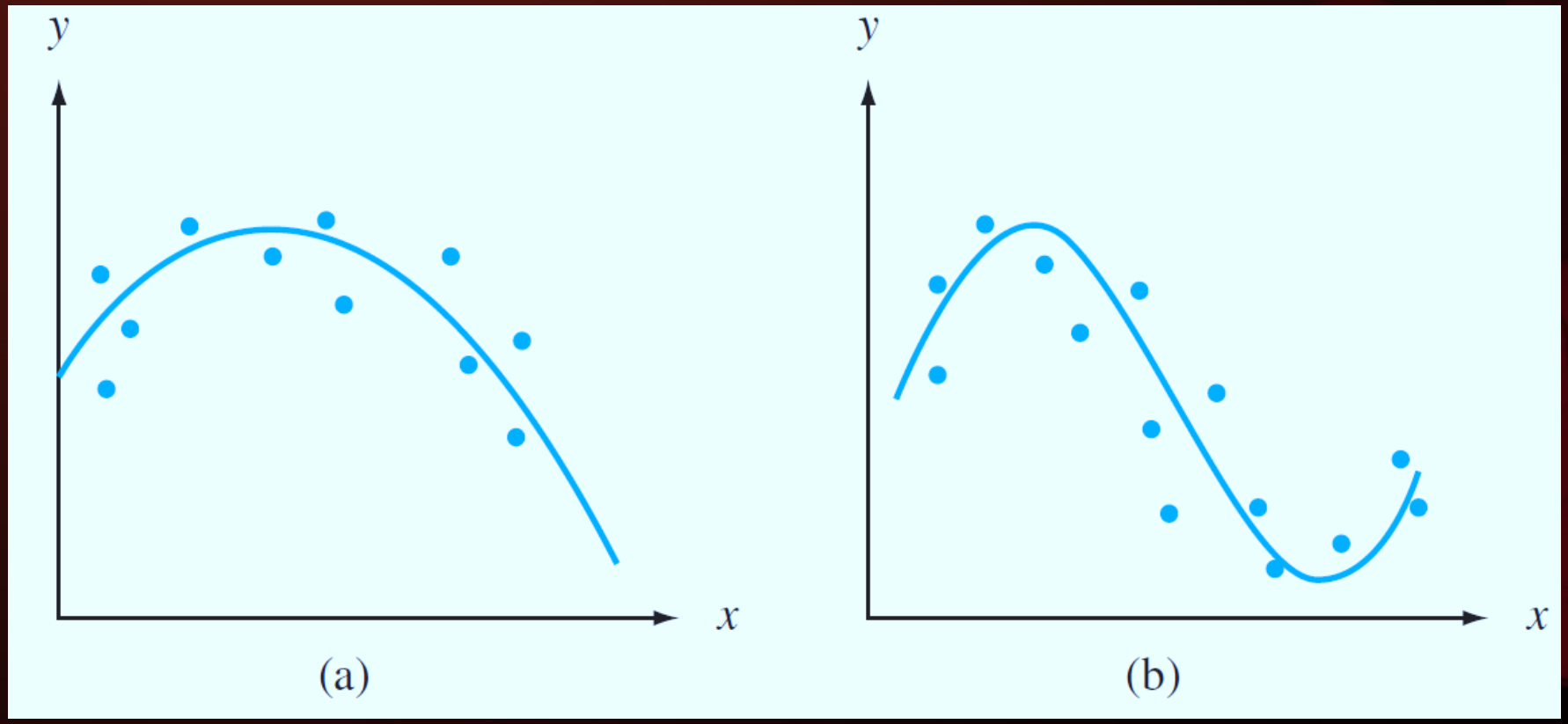
$$\mu_\varepsilon = 0 \text{ e } \sigma_\varepsilon^2 = 0$$

Para estimar os parâmetros:

- Derivadas parciais
- Sistema de equações

$$\begin{aligned}
&b_0 n + b_1 \Sigma x_i + b_2 \Sigma x_i^2 + \cdots + b_k \Sigma x_i^k = \Sigma y_i \\
&b_0 \Sigma x_i + b_1 \Sigma x_i^2 + b_2 \Sigma x_i^3 + \cdots + b_k \Sigma x_i^{k+1} = \Sigma x_i y_i \\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
&b_0 \Sigma x_i^k + b_1 \Sigma x_i^{k+1} + \cdots + b_k \Sigma x_i^{2k} = \Sigma x_i^k y_i
\end{aligned}$$

# MODELO POLINOMIAL




Fonte: (DEVORE, 2018, p. 527)

# VARIÂNCIA ESTIMADA E COEFICIENTE DE DETERMINAÇÃO

Para modelos lineares e a estimativa de $\sigma^2$ é: $\boldsymbol{\sigma^2 = s^2 = \frac{SQE}{n-2}}$

Para os modelos polinomiais: $\boldsymbol{\sigma^2 = s^2 = \frac{SQE}{n-(k+1)}}$

Coeficiente de determinação múltipla: $\boldsymbol{R^2 = 1 - \frac{SQE}{SQT}}$

Coeficiente de determinação múltipla ajustado:

$$R^2\,(\text{ajustado}) = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{\text{SQE}}{\text{SQT}} = \frac{(n-1)R^2 - k}{n-1-k}$$

# INTERVALOS ESTATÍSTICOS E PROCEDIMENTOS DE TESTE

Um IC de $100(1-\alpha)\%$ para $\beta_i$, o coeficiente de $x^i$ na regressão polinomial, é

$$\hat{\beta}_i \pm t_{\alpha/2,\, n-(k+1)} \cdot s_{\hat{\beta}_i}$$

Um teste de $H_0$: $\beta_i = \beta_{i0}$ baseia-se no valor da estatística $t$

$$t = \frac{\hat{\beta}_i - \hat{\beta}_{i0}}{s_{\hat{\beta}_i}}$$

O teste tem por base $n-(k+1)$ gl e é unilateral à direita, unilateral à esquerda ou bilateral, dependendo se a desigualdade em $H_a$ é $>$, $<$ ou $\neq$.

Fonte: (DEVORE, 2018, p. 531)

# INTERVALOS ESTATÍSTICOS E PROCEDIMENTOS DE TESTE

Seja $x^*$ um valor particular de $x$. Um IC de $100(1-\alpha)\%$ para $\mu_{Y \cdot x^*}$ é

$$\hat{\mu}_{Y \cdot x^*} \pm t_{\alpha/2, n-(k+1)} \cdot \left\{ \begin{matrix} \text{DP estimado de} \\ \hat{\mu}_{Y \cdot x^*} \end{matrix} \right\}$$

Com $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^* + \cdots + \hat{\beta}_k (x^*)^k$, $\hat{y}$ representando o valor calculado de $\hat{Y}$ para os dados especificados e $s_{\hat{Y}}$ representando o desvio padrão estimado da estatística $\hat{Y}$, a fórmula para o IC é mais parecida com a da regressão linear simples:

$$\hat{y} \pm t_{\alpha/2,\, n-(k+1)} \cdot s_{\hat{Y}}$$

Um IP de $100(1-\alpha)\%$ para um valor futuro $y$ a ser observado quando $x = x^*$ é

$$\hat{\mu}_{Y \cdot x^*} \pm t_{\alpha/2, n-(k+1)} \cdot \left\{ s^2 + \left( \begin{matrix} \text{DP estimado de} \\ \hat{\mu}_{Y \cdot x^*} \end{matrix} \right)^2 \right\}^{1/2} = \hat{y} \pm t_{\alpha/2, n-(k+1)} \cdot \sqrt{s^2 + s_{\hat{Y}}^2}$$

Fonte: (DEVORE, 2018, p. 531)

# CENTRALIZAÇÃO DE VALORES

$\beta_0$ → valor de y pra x=0

$\beta_1$ → primeira derivada da função em x = 0
(taxa instantânea da mudança de $\mu_{y,x^*}$ em x = 0)

Se $x_i$ distante de zero existe a probabilidade de obter informações imprecisas sobre os parâmetros $\beta_0, \beta_1, \beta_2, \ldots \beta_k$

$$y = \beta_0 + \beta_1(x - \bar{x}) + \beta_2(x - \bar{x})^2 + \cdots + \beta_k(x - \bar{x})^k + \varepsilon$$

Variável → $x_i'$

# REGRESSÃO POLINOMIAL

## Exemplo

O artigo Propriedades físicas da semente de cominho (“*Physical properties of cumin seed”,* J. of Agric. Engr. Res., 1996: 93-98) considerou uma regressão quadrática de y = densidade aparente sobre x = teor de umidade.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.938 |
| Model: | OLS | Adj. R-squared: | 0.896 |
| Method: | Least Squares | F-statistic: | 22.51 |
| Date: | Sat, 12 Mar 2022 | Prob (F-statistic): | 0.0156 |
| Time: | 21:12:03 | Log-Likelihood: | -20.339 |
| No. Observations: | 6 | AIC: | 46.68 |
| Df Residuals: | 3 | BIC: | 46.05 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 403.2396 | 36.453 | 11.062 | 0.002 | 287.230 | 519.249 |
| x1 | 16.1636 | 5.451 | 2.965 | 0.059 | -1.185 | 33.512 |
| x2 | -0.7063 | 0.185 | -3.813 | 0.032 | -1.296 | -0.117 |

| | | | |
|---|---|---|---|
| Omnibus: | nan | Durbin-Watson: | 2.980 |
| Prob(Omnibus): | nan | Jarque-Bera (JB): | 0.911 |
| Skew: | 0.679 | Prob(JB): | 0.634 |
| Kurtosis: | 1.658 | Cond. No. | 2.60e+03 |

```
x = (7,10.3,13.7,16.6,19.8,22)
y= (479,503,487,470,458,412)
```

# REGRESSÃO POLINOMIAL

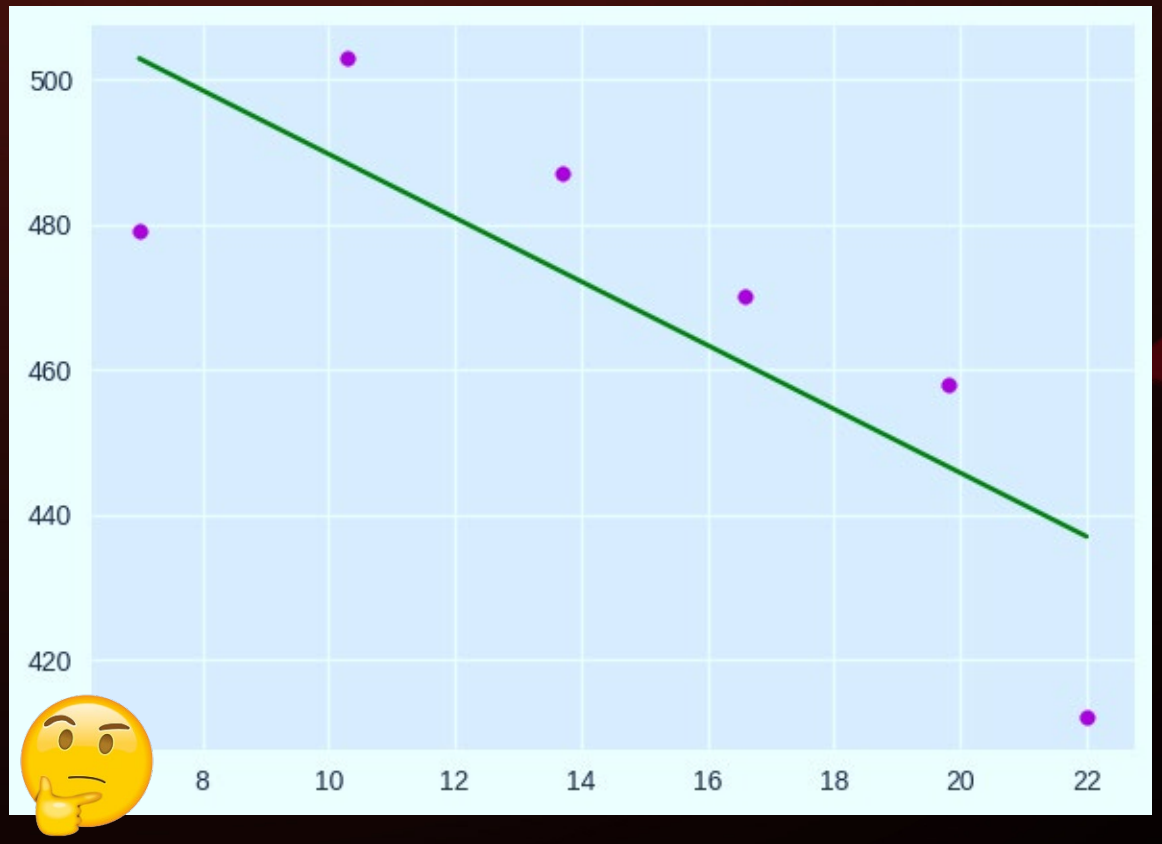
## Exemplo

Responda:

a. Um gráfico de dispersão dos dados parece consistente com modelo de regressão quadrática?
b. Que proporção da variação observada na densidade pode ser atribuída à relação do modelo?
c. Calcule um IC de 95% para a densidade média verdadeira quando o teor de umidade for 13,7, se $\hat{s}_Y = 6{,}49$.
d. Calcule um IP de 99% para a densidade média verdadeira quando o teor de umidade for 14.
e. O preditor quadrático parece fornecer informações úteis? Teste as hipóteses apropriadas no nível de significância 0,05.

# REGRESSÃO POLINOMIAL

Antes de resolver o exercício




1. Definir os dados:

```
lstx = (7,10.3,13.7,16.6,19.8,22)
lsty= (479,503,487,470,458,412)
# Construir o DataFrame e nomear as colunas
df = pd.DataFrame(list(zip(lstx, lsty)),
                  columns =["x","y"])
x=df['x']
y=df['y']
```

```
x=df['x']
yp= 533.6984 -4.3981*x
plt.plot(x, yp, color = "g")
plt.scatter(x, y, color = "m",  marker = "o", s = 30)
plt.grid(True)
x = sm.add_constant(x)
```

2. Se fosse um modelo linear

# REGRESSÃO POLINOMIAL

## Antes de resolver o exercício

1. Usar a biblioteca **statmodel**
2. Obter o polinômio quadrático

```
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)
```

```
1 x=df['x']
2 y=df['y']
3 x = sm.add_constant(x)
4 #Manter x + constant
5 from sklearn.preprocessing import PolynomialFeatures
6 #definir o valor de k, isto é o grau do polinômio
7 polynomial_features= PolynomialFeatures(degree=2)
8 xp = polynomial_features.fit_transform(x)
9 xp.shape
```

```
(6, 6)
```

# REGRESSÃO POLINOMIAL

## Antes de resolver o exercício

Reconstruir x(6,6)

```
 1 xp
```

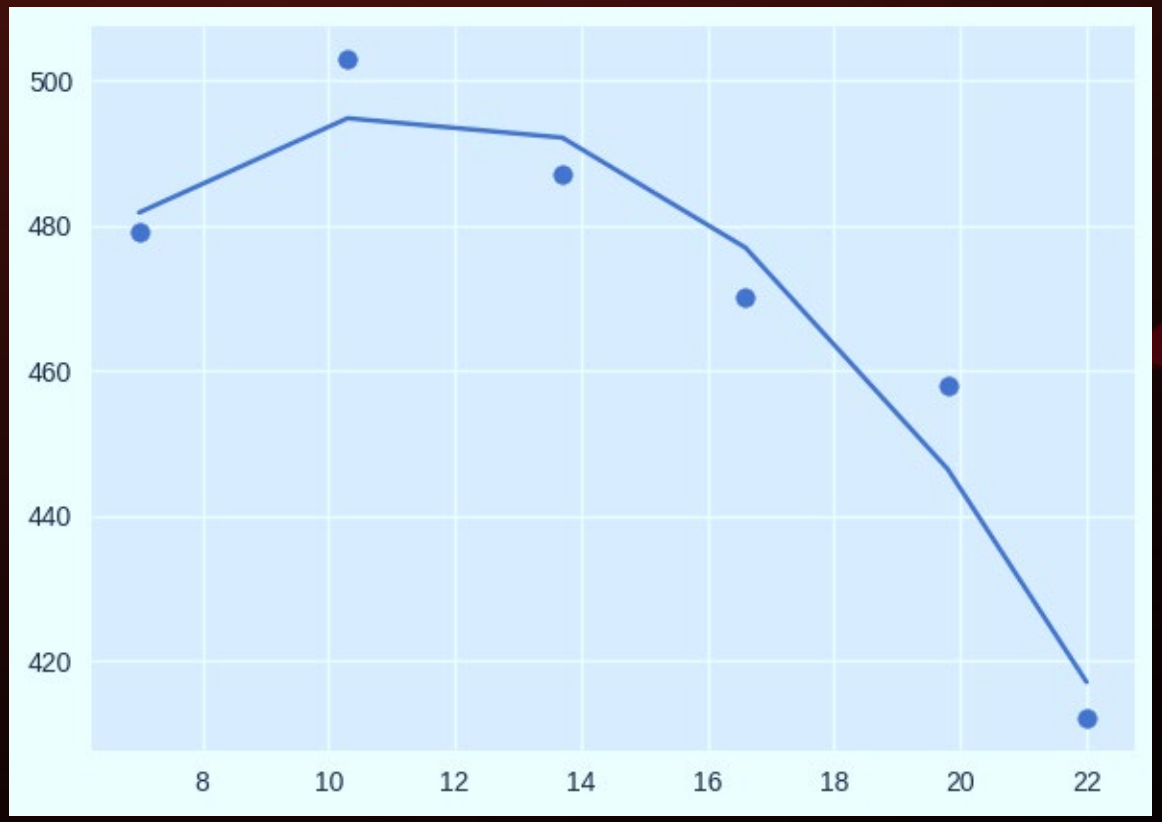
```
array([[  1.  ,   1.  ,   7.  ,   1.  ,   7.  ,  49.  ],
       [  1.  ,   1.  ,  10.3 ,   1.  ,  10.3 , 106.09],
       [  1.  ,   1.  ,  13.7 ,   1.  ,  13.7 , 187.69],
       [  1.  ,   1.  ,  16.6 ,   1.  ,  16.6 , 275.56],
       [  1.  ,   1.  ,  19.8 ,   1.  ,  19.8 , 392.04],
       [  1.  ,   1.  ,  22.  ,   1.  ,  22.  , 484.  ]])
```

```
 1 xp[:,3:6]
```

```
array([[  1.  ,   7.  ,  49.  ],
       [  1.  ,  10.3 , 106.09],
       [  1.  ,  13.7 , 187.69],
       [  1.  ,  16.6 , 275.56],
       [  1.  ,  19.8 , 392.04],
       [  1.  ,  22.  , 484.  ]])
```

# REGRESSÃO POLINOMIAL

**Resposta a)** Um gráfico de dispersão dos dados parece consistente com modelo de regressão quadrática?




```
1 #Aplicar a regressão polinomial com o novo x, isto é xp
2 modelpo = sm.OLS(y,xp[:,3:6]).fit()
3 ypred = modelpo.predict(xp[:,3:6])
4 ypred.shape
```

```
(6,)
```

```
1 x=df['x']
2 plt.scatter(x,y)
3 plt.plot(x,ypred)
```

# REGRESSÃO POLINOMIAL

**Resposta b)** Que proporção da variação observada na densidade pode ser atribuída à relação do modelo?

$$R^2 = 1 - \frac{SQE}{SQT}$$

$$R^2\,(\text{ajustado}) = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{\text{SQE}}{\text{SQT}} = \frac{(n-1)R^2 - k}{n-1-k}$$

```python
#@title resposta b)
#calcular SQE--> Baseado nos valores esperados
sqe = np.sum((ypred - y)**2)
print("SQE=", sqe)

#calcular SQT-->Baseado nos valores observados
sqt = np.sum((y - y.mean())**2)
print("SQT=", sqt)

#calcular SQR
sqr = sqt - sqe
print("SQR=",sqr)

R2=1-sqe/sqt
print("Coeficiente de determinação múltippla=",R2, "ou",R2*100, "%")

k=modelpo.df_model # grau do modelo
n=modelpo.nobs # num. amostras
R2adj=((n-1)*R2-k)/(n-1-k)

print("R2 ajustado=",R2adj, "ou",R2adj*100, "%")
```

```
SQE= 309.10898960299954
SQT= 4946.833333333333
SQR= 4637.724343730333
Coeficiente de determinação múltippla= 0.9375137651151242
R2 ajustado= 0.8958562751918736 ou 89.58562751918736 %
```

# REGRESSÃO POLINOMIAL

**Resposta b)** Que proporção da variação observada na densidade pode ser atribuída à relação do modelo?

$$R^2 = 1 - \frac{SQE}{SQT}$$

$$R^2\,(\text{ajustado}) = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{\text{SQE}}{\text{SQT}} = \frac{(n-1)R^2 - k}{n-1-k}$$

```
1 r2d=modelpo.rsquared
2 r2adjd=modelpo.rsquared_adj
3 print('R2= {0}, R2 ajustado= {1}'.format(r2d,r2adjd))
```

```
R2= 0.9375137651151243, R2 ajustado= 0.8958562751918738
```

# REGRESSÃO POLINOMIAL

**Resposta d)** Calcule um IP de 99% para a densidade média verdadeira quando o teor de umidade for 14.

- Nos modelos lineares o IC se calcula com a equação $\hat{y} \pm t_{\frac{\alpha}{2},n-2}\sqrt{s_{\hat{y}}^2 + s^2}$
- Nos modelos polinomiais o IC deve ser calculado mediante $\hat{y} \pm t_{\frac{\alpha}{2},(n-(k+1))}\sqrt{s_{\hat{y}}^2 + s^2}$

```
1 #usar a tabela tstudent pata t
2 from scipy.stats import t
3 alpha = 0.01  # significia = 1%
4 df = n-(k+1) # graus de liberdade
5 v = t.ppf(1 - alpha/2, df)
6 tt=v
7 print(f't_crit=: {v}')
```

```
t_crit=: 5.84090929975643
```

$t_{(0,005,3)} = 5,841$

$\hat{s}_Y = 6,49.$

$$y = 403,2396 + 16,1636x - 0,7063x^2$$

$$\hat{y} = \mu_{y,14} = 491,1$$

```
1 s2=sqe/(n-(k+1))
2 s=pow(s2,1/2)
3 s
```

```
10.150681251407045
```

# REGRESSÃO POLINOMIAL

**Resposta d)** Calcule um IP de 99% para a densidade média verdadeira quando o teor de umidade for 14.

- Nos modelos polinomiais o IC deve ser calculado mediante $\hat{y} \pm t_{\frac{\alpha}{2},(n-(k+1))}\sqrt{s_{\hat{y}}^2 + s^2}$

$$\hat{y} = \mu_{y,14} = 491{,}1 \text{ e } t_{(0{,}005,3)} = 5{,}841$$

$$\hat{y} \pm t_{\frac{\alpha}{2},(n-(k+1))}\sqrt{s_{\hat{y}}^2 + s^2}$$

$$491{,}1 \pm 5{,}841\left(\sqrt{6{,}49^2 + 10{,}45^2}\right)$$

IP de $\hat{y}$ (419,25 , 562,95)

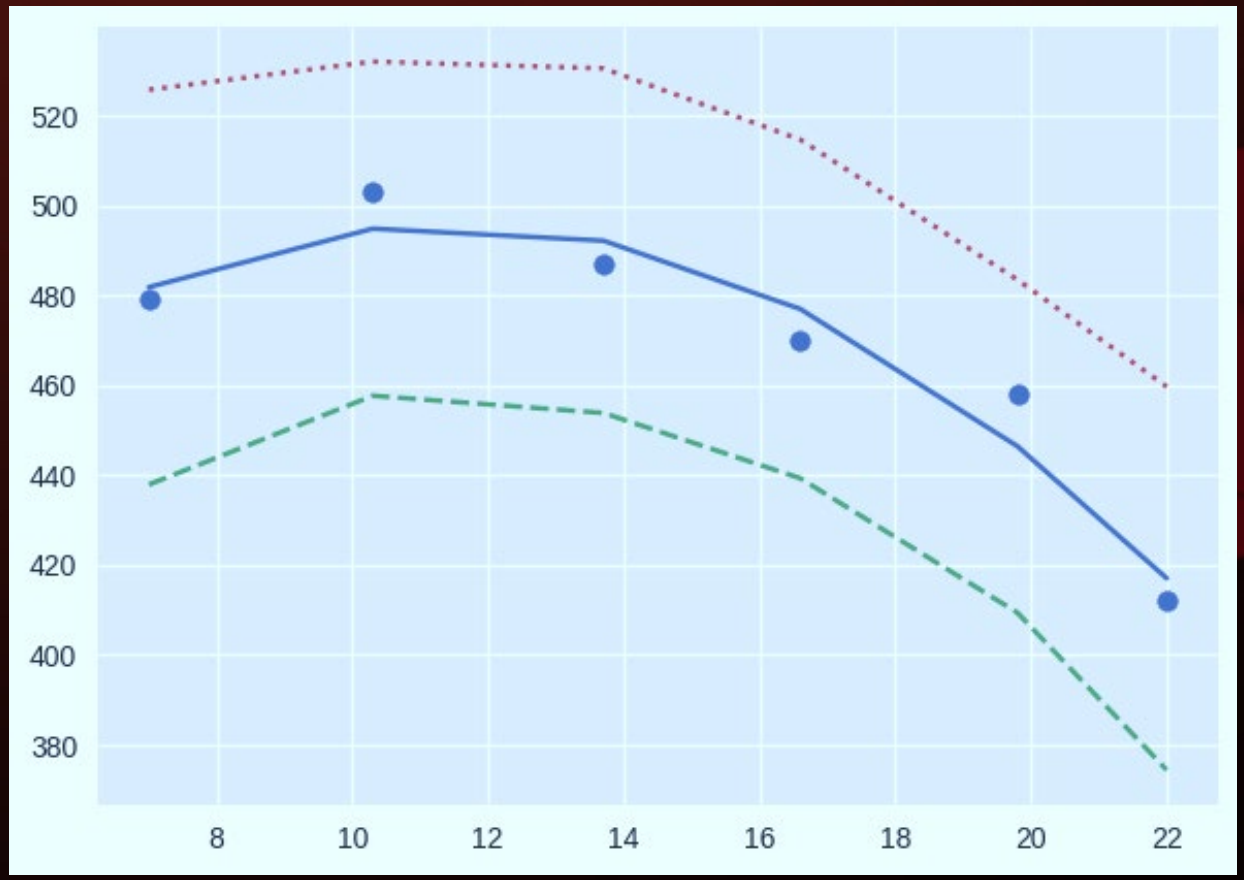
# REGRESSÃO POLINOMIAL

**Resposta d)** Calcule um IP de 99% para a densidade média verdadeira quando o teor de umidade for 14.

```
1 #Intervalos
2 from scipy import stats
3 from statsmodels.sandbox.regression.predstd import wls_prediction_std
4 _, upper,lower = wls_prediction_std(modelpo)
5 plt.scatter(x,y)
6 plt.plot(x,ypred)
7 plt.plot(x,upper,'--',label="Maior") # confid. intrvl
8 plt.plot(x,lower,':',label="Menor")
```

# REGRESSÃO POLINOMIAL

**Resposta d)** Calcule um IP de 99% para a densidade média verdadeira quando o teor de umidade for 14.




IP de $\hat{y}$ (419,25 , 562,95)

# REGRESSÃO POLINOMIAL

**Resposta e)** O preditor quadrático parece fornecer informações úteis? Teste as hipóteses apropriadas no nível de significância 0,05.

- **Definir a hipótese nula $H_0: \beta_{20} = 0$ frente a $H_a: \beta_{20} \neq 0$.**
- **Se $|t| \geq t_{crit}$ rejeitar $H_0$ em favor de $H_a: \beta_{20} \neq 0$.**
  - $t_{crit} = t_{(0,025,3)} = 5,841$
  - $|t| = 3,813$
  - $3,813 \geq 5,841$ ?? NÃO
  - Portanto REJEITAR $H_0: \beta_{20} = 0$
- **Comparar se p-valor < α.**
  - p-valor 0,032 < 0,05
  - Portanto REJEITAR $H_0: \beta_{20} = 0$ frente a $H_a: \beta_{20} \neq 0$ → Modelo quadrático.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 403.2396 | 36.453 | 11.062 | 0.002 | 287.230 | 519.249 |
| x1 | 16.1636 | 5.451 | 2.965 | 0.059 | -1.185 | 33.512 |
| x2 | -0.7063 | 0.185 | -3.813 | 0.032 | -1.296 | -0.117 |

| | | | |
|---|---|---|---|
| Omnibus: | nan | Durbin-Watson: | 2.980 |
| Prob(Omnibus): | nan | Jarque-Bera (JB): | 0.911 |
| Skew: | 0.679 | Prob(JB): | 0.634 |
| Kurtosis: | 1.658 | Cond. No. | 2.60e+03 |

# MODELAGEM E INFERÊNCIA ESTATÍSTICA

## Modelos polinomiais