Regressão

Lucas de Oliveira Santos
lucas.santos@lapisco.ifce.edu.br

Assuntos abordados

- ⇨ Introdução
- ⇨ Correlação entre variáveis
- ⇨ Regressão Linear Simples
- ⇨ Regressão Múltipla
- ⇨ Regressão Polinomial
- ⇨ Conclusões

Introdução

- ⇨ Em suma, a regressão busca modelar a relação entre variáveis do tipo entrada-saída, permitindo a previsão de valores contínuos.
- ⇨ Ao explorarmos a evolução desse método, podemos compreender melhor como eles têm contribuído para a evolução da interpretação e da previsão de dados, impactando campos tão diversos quanto medicina, finanças e ciência dos dados.

Correlação entre variáveis

⇨ Supondo um conjunto de dados com ***N amostras***, os pares ordenados ***(x, y)*** podem ser usados para representá-las.

⇨ ***x*** é a variável independente (ou explanatória) e ***y*** é a variável dependente.

Correlação entre variáveis

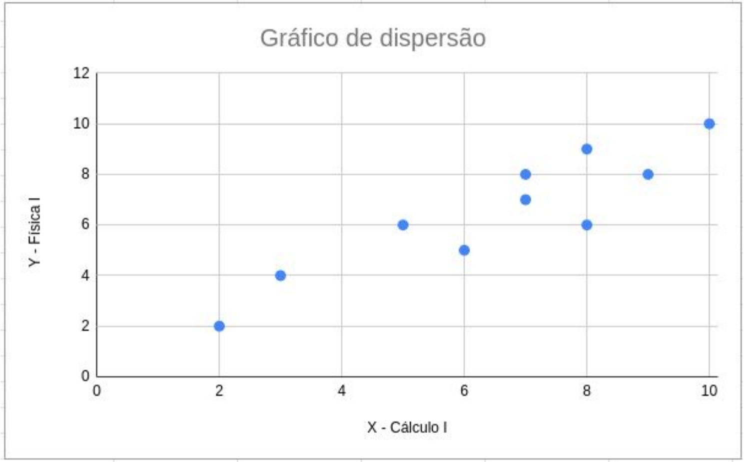
⇨ Exemplo: Coletou-se as notas de 10 alunos do IFCE nas disciplinas de cálculo I e física I. A tabela abaixo mostra a distribuição dessas notas. Procura-se entender se existe alguma correlação entre as notas obtidas entre as disciplinas e qual a sua natureza.

Notas	
Cálculo I (X)	Física I (Y)
5	6
8	9
7	8
10	10
7	7
6	5
3	4
9	8
8	6
2	2

Correlação entre variáveis

⇨ Exemplo: Coletou-se as notas de 10 alunos do IFCE nas disciplinas de cálculo I e física I. A tabela abaixo mostra a distribuição dessas notas. Procura-se entender se existe alguma correlação entre as notas obtidas entre as disciplinas e qual a sua natureza.

Notas	
Cálculo I (X)	Física I (Y)
5	6
8	9
7	8
10	10
7	7
6	5
3	4
9	8
8	6
2	2



Correlação entre variáveis

- ⇨ Para correlacionar as variáveis, utiliza-se o Coeficiente de Correlação Linear;
  - A fórmula do coeficiente é: $r = \frac{\mathbf{n}\sum xy - \sum x \sum y}{\sqrt{[\mathbf{n}\sum x^2 - (\sum x)^2][\mathbf{n}\sum y^2 - (\sum y)^2]}}$;
  - O intervalo esperado está entre -1 e 1;
  - Se r = 0, nao existe correlação entre as variáveis;
  - Se 0 < r < 0.5, existe uma correlação fraca entre as variáveis;
  - Se 0.5 < r < 1, existe uma correlação forte entre as variáveis;
  - Se r = 1, existe uma correlação perfeita entre as variáveis;
- ⇨ Para medir o quanto a variação observada em y é explicada pelas variáveis x, usa-se o r^2 que é uma métrica estatística.
  - O valor de r^2 varia entre 0 e 1 e pode ser classificado da seguinte forma:
    - Se r^2 = 0, o modelo não explica nenhuma variação nos dados;
    - Se r^2 = 1, o modelo explica toda a variação nos dados;

Correlação entre variáveis

⇨ Aplicação do Coeficiente de Correlação no exemplo das notas dos alunos.

Notas				
Cálculo I (X)	Física I (Y)	X*Y	X^2	Y^2
5	6	30	25	36
8	9	72	64	81
7	8	56	49	64
10	10	100	100	100
7	7	49	49	49
6	5	30	36	25
3	4	12	9	16
9	8	72	81	64
8	6	48	64	36
2	2	4	4	4
65	65	473	481	475

$$r = \frac{\mathbf{n}\sum xy - \sum x \sum y}{\sqrt{[\mathbf{n}\sum x^2 - (\sum x)^2][\mathbf{n}\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{10 \cdot 473 - 65 \cdot 65}{\sqrt{[10 \cdot 481 - (65)^2][10 \cdot 475 - (65)^2]}} = 0.9112$$

Como r = 0.9112, entendemos que a correlação entre as notas nas disciplinas é forte.

O valor de ***r²*** é igual a **0.8303**, isso indica que as variáveis em ***x*** explicam **83.03%** os resultados observados em ***y***.

Regressão Linear Simples

- A regressão linear simples é uma técnica estatística usada para modelar a ***relação entre duas variáveis***: uma variável independente ***x*** e uma dependente ***y***;
- Condição:
  - Assume-se que a relação existente pode ser aproximada por uma equação linear, ou seja, uma reta.
- A equação da regressão linear simples é representada pela fórmula: $Y = \beta_0 + \beta_1 \cdot X + \epsilon$
  - $\boldsymbol{\beta_0}$ é o coeficiente de interceptação;
  - $\boldsymbol{\beta_1}$ é o coeficiente de inclinação da reta;
  - $\boldsymbol{\varepsilon}$ é o termo de erro, que representa a parte não explicada pela relação linear entre ***X*** e ***Y***
- O método utilizado para calcular coeficientes $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$ é chamado de Estimador de Mínimos Quadrados - MQO. Ele requer que as condições abaixo sejam verificadas.

$$\beta_0 = \frac{\sum y}{n} - \beta_1 \cdot \frac{\sum x}{n} \qquad \beta_1 = \frac{n \sum x \cdot y - \sum x \cdot \sum y}{n \sum x^2 - (\sum x)^2}$$

Regressão Linear Simples

⇨ Algumas curiosidades sobre o Método dos Mínimos Quadrados - MQO

- Foi proposto em 1795 por **Carl Friedrich Gauss**;
- Inicialmente, o método foi aplicado no cálculo de órbitas de planetas e cometas a partir de medidas obtidas por telescópios;
- **Adrien Marie Legendre** desenvolveu de forma independente o mesmo método e o publicou primeiro em 1806;

Regressão Linear Simples

⇨ Aplicando o conceito de regressão linear nos dados dos 10 alunos selecionados.

Notas				
Cálculo I (X)	Física I (Y)	X*Y	X²	Y²
5	6	30	25	36
8	9	72	64	81
7	8	56	49	64
10	10	100	100	100
7	7	49	49	49
6	5	30	36	25
3	4	12	9	16
9	8	72	81	64
8	6	48	64	36
2	2	4	4	4
65	65	473	481	475

Encontrando os coeficientes β_0 e β_1 utilizando o MQO :

$$\beta_1 = \frac{n\sum x \cdot y - \sum x \cdot \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \cdot 473 - 65 \cdot 65}{10 \cdot 481 - 65^2} = \frac{505}{585} = 0.8632$$

$$\beta_0 = \frac{\sum y}{n} - \beta_1 \cdot \frac{\sum x}{n} = \frac{65}{10} - 0.8632 \cdot \frac{65}{10} = 6.5 - 5.6108 = 0.8892$$

⇨ A equação de regressão ajustada aos dados é dada por: $\hat{Y} = 0.8892 + 0.8632 \cdot X$

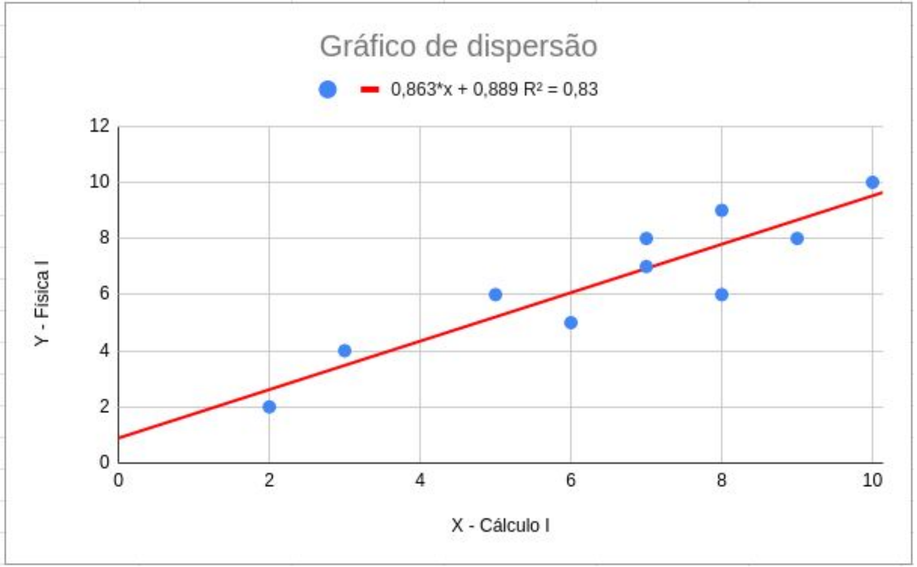
⇨ O operador ^ indica que o valor encontrado é uma aproximação, uma estimativa.

⇨ O erro de predição ε é dado por: $\epsilon = Y - \hat{Y}$

Regressão Linear Simples

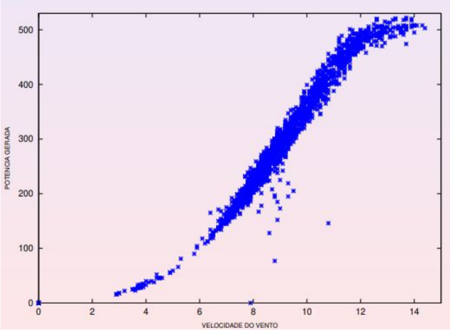
- Gráfico de dispersão com a plotagem da reta de regressão linear encontrada anteriormente.

Notas				
Cálculo I (X)	Física I (Y)	X*Y	X²	Y²
5	6	30	25	36
8	9	72	64	81
7	8	56	49	64
10	10	100	100	100
7	7	49	49	49
6	5	30	36	25
3	4	12	9	16
9	8	72	81	64
8	6	48	64	36
2	2	4	4	4
65	65	473	481	475

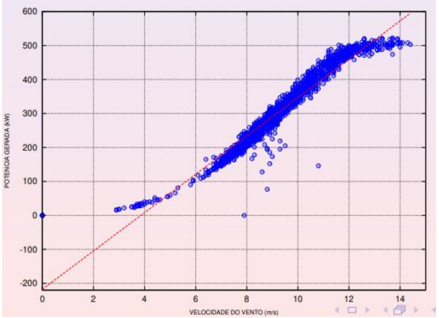


Regressão Linear Simples

- A seguir, vemos o gráfico de dispersão dos dados de um aerogerador com ***N*** = 2250.
- Aplicando-se o método MQO, obtemos os coeficientes β_0 = -217.69, β_1 = 56.44 e r^2 = 0.93.
- Considerações:
  - O alto valor de r^2 não deve ser o mais relevante na abordagem da regressão;
  - Embora o r^2 seja alto, nota-se visualmente que uma reta não ideal para representar os dados;



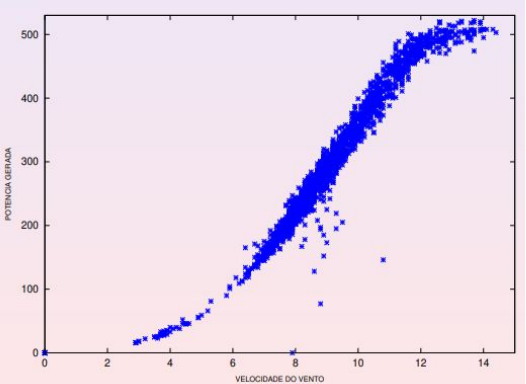
a) Plotagem dos dados



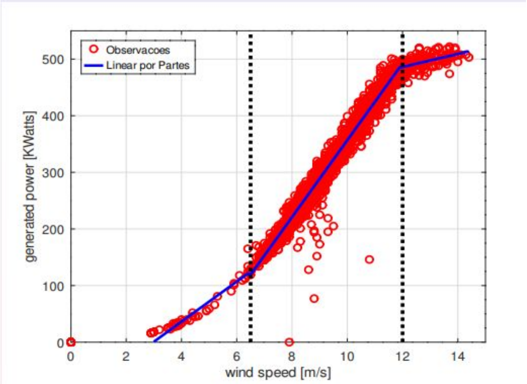
b) Plotagem da reta de regressão linear simples

Regressão Linear Simples

- ⇨ Abordagens tomadas a partir da observação dos dados plotados:
  - Utilizar a ideia de regressão linear simples por partes;
    - Deve-se dividir o gráfico de dispersão em sub-regiões onde a regressão linear seja adequada;



a) Plotagem dos dados



b) Divisão do gráfico em 3 sub-regiões e plotagem da reta de regressão linear simples

Regressão Múltipla

- ⇨ Em muitas situações do cotidiano, mais de uma variável ***x*** está se relacionando com apenas uma variável de resultado ***y***.
- ⇨ Nesse sentido, a aplicação da regressão linear passa a ser expressa pela seguinte equação: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$
- ⇨ Onde os valores representados por x_k são as ***k*** variáveis independentes, ***y*** representa a variável dependente e ε denota os erros associados aos ***k*** elementos.

Regressão Múltipla

⇨ Os coeficientes ($\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_p$) são calculados usando-se o método MQO, que é dado pela expressão a seguir:

$$J(\beta_1, \beta_2, ..., \beta_p) = \sum_{i=1}^{n} \varepsilon_i^n = \sum_{i=1}^{2} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{j,i} \right)^2$$

⇨ A função de soma dos quadrados é dada por: $S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i,1} - ... - \beta_k x_{i,k})^2$

⇨ Ela pode ser decomposta em função das derivadas parciais de cada um dos coeficientes da seguinte forma:

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1 ..., \hat{\beta}_k)}{\partial \beta_0} = \sum_{i=1}^{N} -2 \cdot \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - ... - \hat{\beta}_k x_{i,k} \right) = 0$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1 ..., \hat{\beta}_k)}{\partial \beta_1} = \sum_{i=1}^{N} -2 \cdot x_{i,1} \cdot \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - ... - \hat{\beta}_k x_{i,k} \right) = 0$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1 ..., \hat{\beta}_k)}{\partial \beta_k} = \sum_{i=1}^{N} -2 \cdot x_{i,k} \cdot \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - ... - \hat{\beta}_k x_{i,k} \right) = 0$$

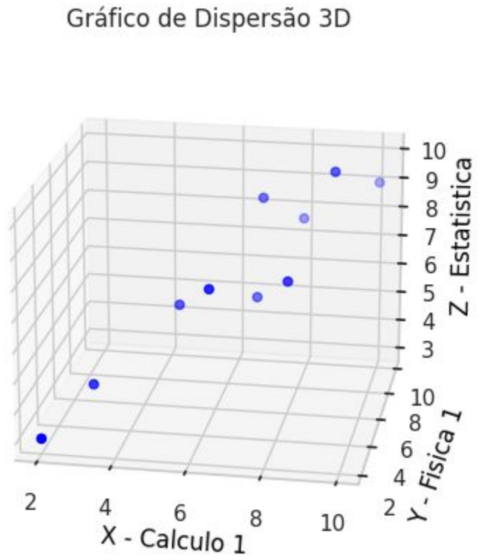
Regressão Múltipla

- ⇨ A formulação matricial desses conceitos é importante do ponto de vista computacional.
- ⇨ Nesse sentido, o modelo de regressão múltipla pode ser definido como: $y = Xb + e$
- ⇨ Expandindo os termos da matriz, temos: $\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & x_{23} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & x_{33} & \dots & x_{k3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1p} & x_{2p} & x_{3p} & \dots & x_{kp} \end{pmatrix} * \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_p \end{pmatrix}$
- ⇨ A estimativa de MQO do ponto de vista matricial é dada pela minimização da expressão: $e^T e = \sum_{i=1}^{N} e_i^2$, sendo que ***e*** é dado por: $e = y - X\beta$. Dessa forma, podemos reescrever a equação: $e^T e = (y - X\beta)^T (y - X\beta)$
- ⇨ Por fim, a estimativa de MQO é dada por: $\hat{\beta} = (X^T X)^{-1} (X^T y)$. Essa equação é conhecida como o método da pseudo-inversa.

Regressão Múltipla

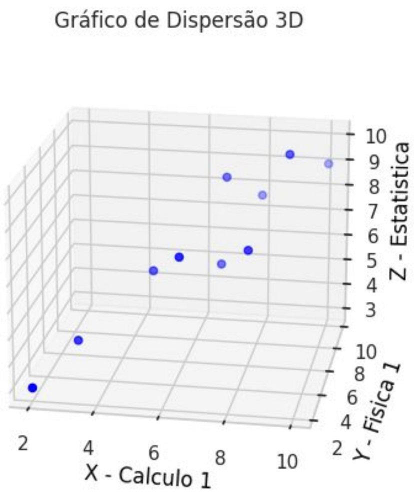
⇨ Exemplo: Coletou-se as notas de 10 alunos do IFCE nas disciplinas de cálculo I, física I e estatística. A tabela abaixo mostra a distribuição dessas notas. Procura-se entender se existe alguma correlação entre as notas obtidas entre as disciplinas e qual a sua natureza.

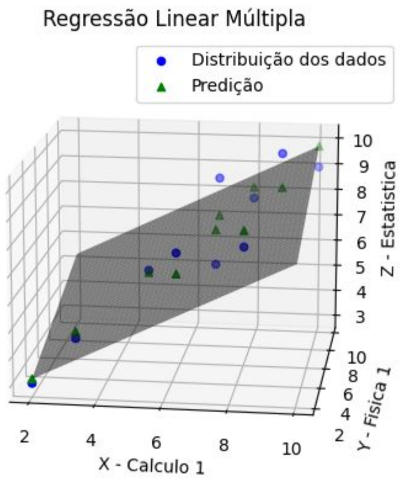
Notas		
Cálculo I (X)	Física I (Y)	Estatística (Z)
5	6	6
8	9	8
7	8	9
10	10	9
7	7	6
6	5	7
3	4	4
9	8	10
8	6	7
2	2	3
65	65	66



Regressão Múltipla

⇨ Exemplo: Coletou-se as notas de 10 alunos do IFCE nas disciplinas de cálculo I, física I e estatística. A tabela abaixo mostra a distribuição dessas notas. Procura-se entender se existe alguma correlação entre as notas obtidas entre as disciplinas e qual a sua natureza.





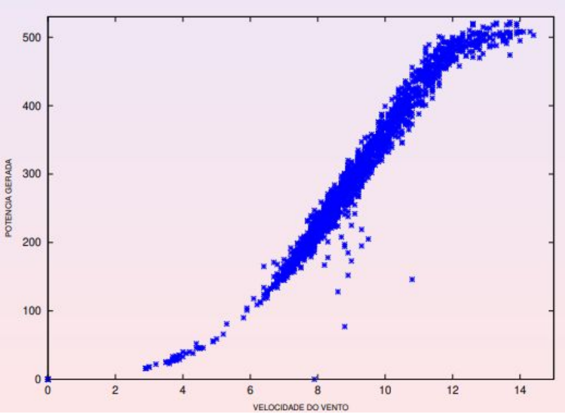
Equação do plano:

y = 1.49 + 0.57*X + 0.26*Z

Coeficiente ***r*** = 0.916 e ***r²*** = 0.84

Regressão Polinomial

- Vamos analisar abaixo o gráfico de dispersão dos dados de velocidade do vento (m/s) e potência (kW) de um aerogerador.
- Podemos inferir que a natureza da relação entre esses dados é não-linear.
- Dessa forma, o modelo de regressão linear simples ou múltipla pode não representar fidedignamente a distribuição desses dados.
- Para representar da melhor maneira possível esses dados, iremos trabalhar com a regressão polinomial de ordem k.



Regressão Polinomial

⇨ A formulação matemática da regressão polinomial é: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$

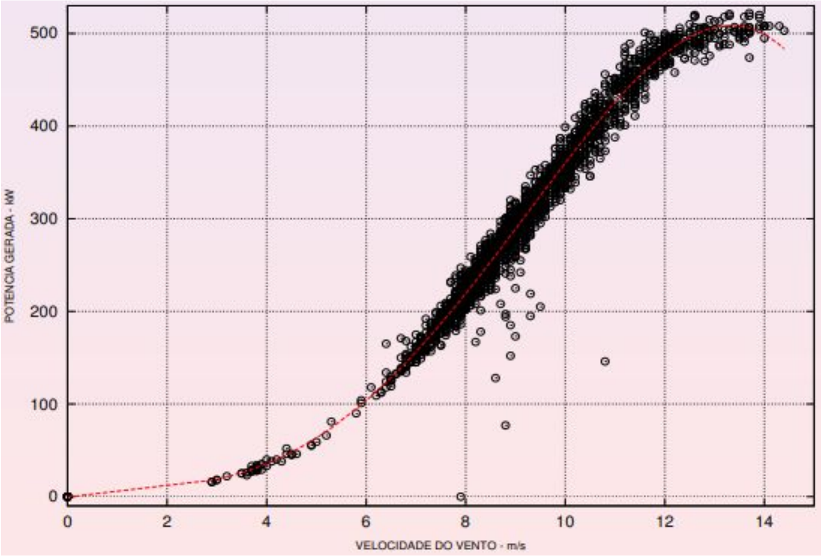
⇨ A formulação matricial é dada por:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^q \\ 1 & x_2 & x_2^2 & \dots & x_2^q \\ 1 & x_3 & x_3^2 & \dots & x_3^q \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_p & x_p^2 & \dots & x_p^q \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \dots \\ \epsilon_p \end{pmatrix}$$

⇨ De forma análoga ao problema de regressão múltipla, podemos estimar os coeficientes beta usando MQO utilizando a seguinte expressão: $\hat{\beta} = (X^T X)^{-1}(X^T y)$

Regressão Polinomial

- Utilizando os dados do aerogerador, ajustou-se o seguinte modelo polinomial de quarta ordem (***q***=4): $\hat{y} = -0.391 + 10.37x - 5.00x^2 + 1.43x^3 - 0.068x^4$, com ***r²*** = 0.974.
- A plotagem da curva do modelo sobreposto ao gráfico de dispersão é mostrada abaixo.

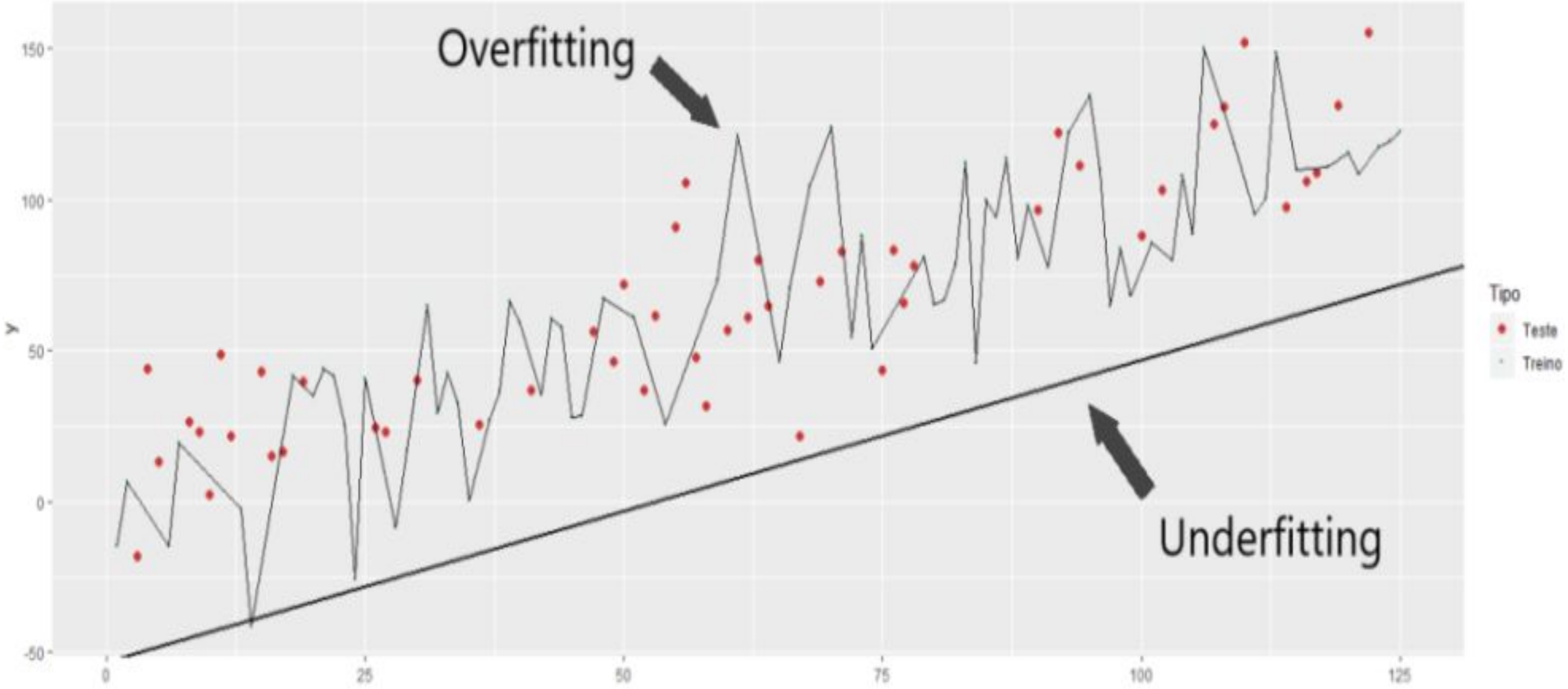


Regressão Polinomial

- ⇨ Deve-se tomar cuidado com o grau do polinômio escolhido, pois existem duas problemáticas relacionadas a ordem do modelo gerado.
- ⇨ *Overfitting:*
  - Quando o modelo de regressão tem um desempenho ótimo na **etapa de treinamento**, porém ao receber novos dados (dados de teste), o desempenho é afetado significativamente. Diz-se então que o modelo **decorou os dados na etapa de treinamento**;
  - O ***overfitting*** é comum quando o grau do polinômio é muito elevado;
- ⇨ *Underfitting:*
  - Quando o modelo de regressão não tem um bom desempenho na ***etapa de treinamento*****,** diz-se haver um ***underfitting.*** Nesse sentido, o modelo de regressão gerado pode ser descartado, pois não apresenta uma boa generalização dos dados.

Regressão Polinomial

⇨ Exemplo de overfitting e underfitting em um gráfico de dispersão qualquer.



Conclusões

- ⇨ A análise de regressão abrange uma série de técnicas voltadas para a modelagem e a investigação de relações entre duas ou mais variáveis aleatórias;
- ⇨ Podemos usar a análise de regressão para construir um modelo matemático que represente fidedignamente a relação determinística (de causa e efeito) entre duas ou mais variáveis entrada-saída;
- ⇨ O modelo de regressão gerado pode ser usado para predizer novos valores;

Referências

- [1] W. Hines, D. Montgomery, D. Goldsman, and C. Borror, “Probabilidade e estatística na engenharia, 4a edição, ed,” LTC, Rio de Janeiro-RJ, 2006.
- [2] O. Helene, Métodos dos Mínimos Quadrados. Editora Livraria da Física, 2006.
- [3] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
- [4] D. Astolfi and R. Pandit, “Multivariate wind turbine power curve model based on data clustering and polynomial lasso regression,” Applied Sciences , vol. 12, no. 1, p. 72, 2021.
- [5] S. Shokrzadeh, M. J. Jozani, and E. Bibeau, “Wind turbine power curve modeling using advanced parametric and nonparametric methods,” IEEE Transactions on Sustainable Energy , vol. 5, no. 4, pp. 1262–1269, 2014.
- [6] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics , vol. 12, no. 1, pp. 55–67, 1970.

Obrigado pela atenção!

Dúvidas?

Lucas de Oliveira Santos
lucas.santos@lapisco.ifce.edu.br