COC800
Introdução à Ciência de Dados

Aprendizado Supervisionado

Rogério Espíndola espindola@ufrj.br

2025.3

Estrutura da aula

- Vizinhos mais próximos
- Modelagem bayesiana
- Modelos lineares
- Árvores e regras de decisão
- Redes neurais artificiais

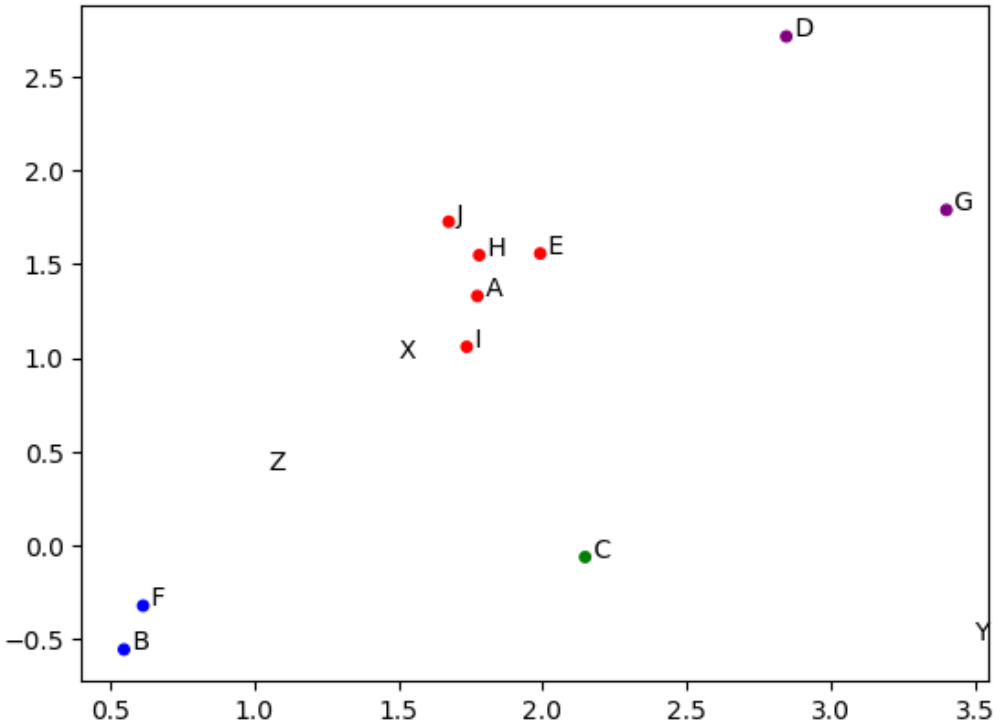
Vizinhos mais próximos

K-*nearest neighbors* (knn)

Proposto em 1951, é um dos mais utilizados algoritmos preditivos sobretudo com o surgimento dos sistemas de recomendação. Ele não produz um modelo de representação dos dados e a generalização a partir dos dados de "treinamento" é realizada no momento da tomada de decisão. Algoritmos com esse comportamento são chamados de "preguiçosos" (*lazy*).

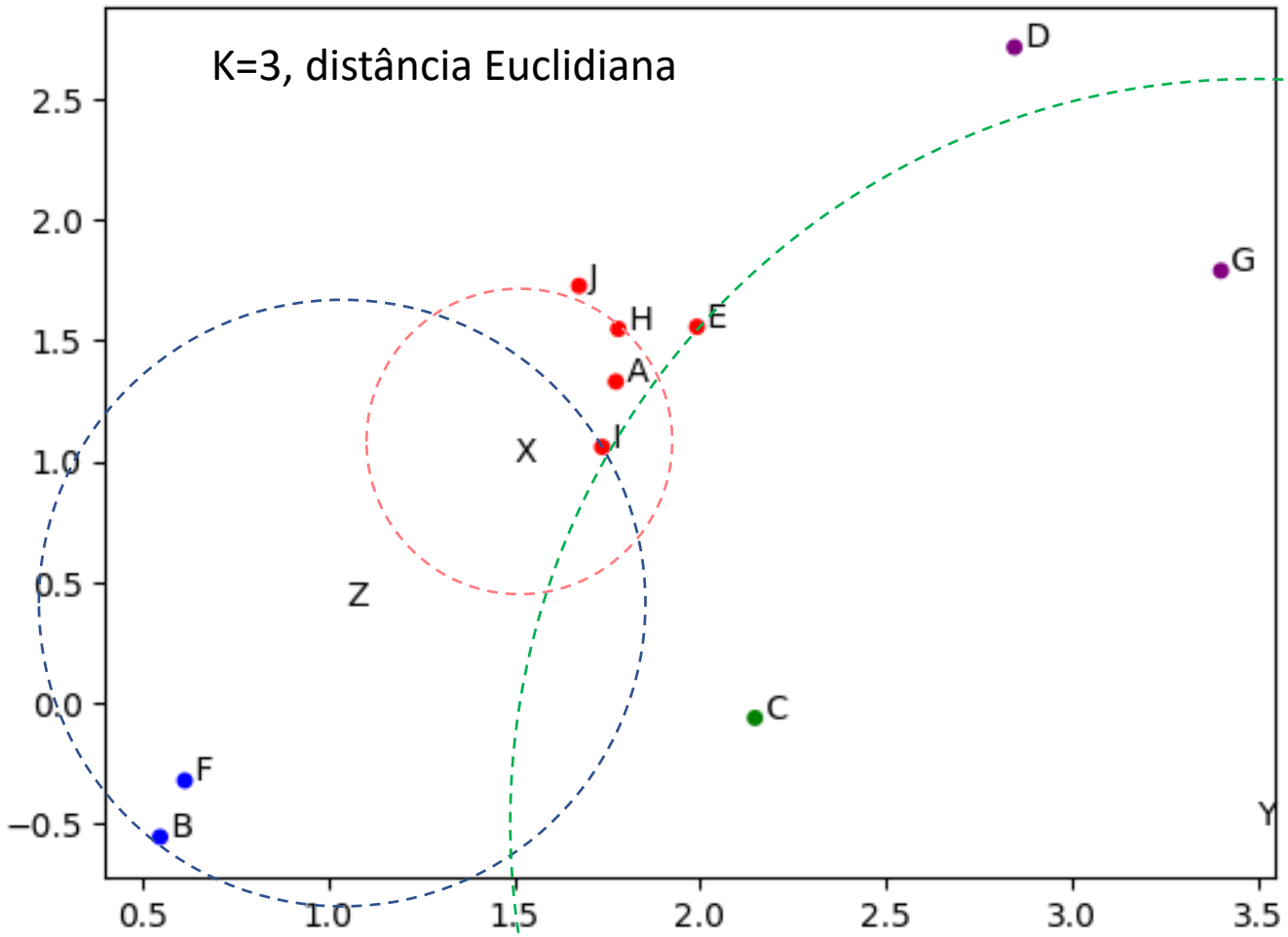
Algoritmos "preguiçosos" são interessantes em problemas cujas bases de dados mudam intensamente e são consultadas frequentemente.

O knn é um algoritmo que produz a resposta para um determinado registro a partir dos seus k registros mais próximos, realizando uma aproximação local da variável resposta. Em problemas de classificação, a saída é a classe mais frequente dentre os vizinhos. Na regressão, a saída é uma combinação dos valores da variável resposta, em geral a média aritmética deles.



Vizinhos mais próximos

K-*nearest neighbors* (knn)



sklearn.neighbors.KNeighborsClassifier

Em geral, ponderar as contribuições dos vizinhos, segundo suas proximidades ao que se quer a resposta, costuma produzir melhores resultados. Assim, registros mais próximos contribuem mais com a resposta.

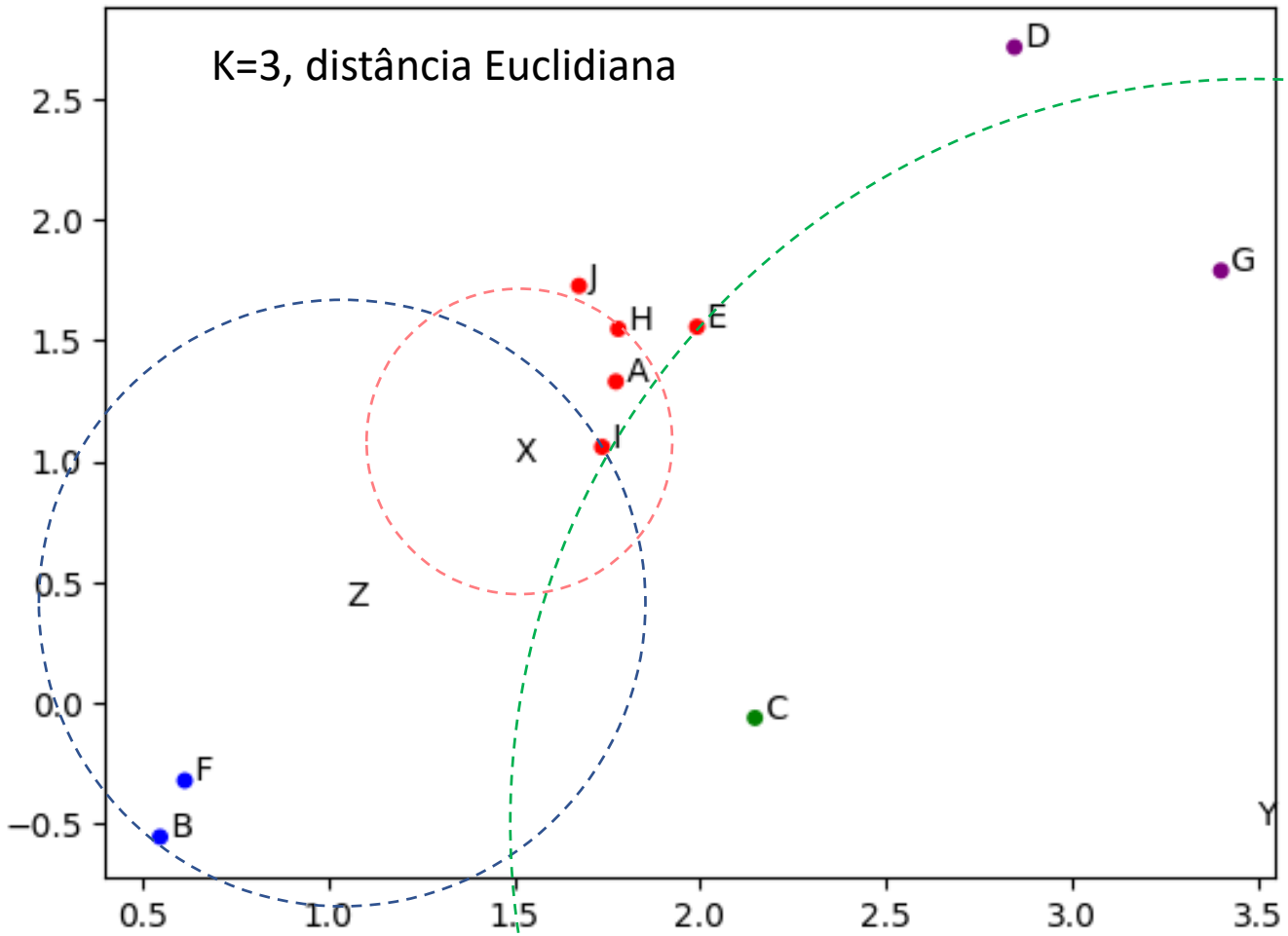
O “treinamento” apenas armazena os dados em memória.

A escolha do k depende dos dados. Valores maiores implicam em robustez a ruídos, porém piora a identificação das fronteiras de decisão.

Atributos irrelevantes devem ser excluídos para não interferirem nos cálculos das distâncias.

Vizinhos mais próximos

K-*nearest neighbors* (knn)



sklearn.neighbors.KNeighborsClassifier

A escolha da função distância a ser adotada também é importante e pode levar a uma melhor separação entre as classes.

Pode-se aplicar a análise de componentes principais para reduzir o número de atributos, tornando o cálculo de distância mais eficiente.

Podem ser arbitrados limites para os raios de observação de vizinhos evitando-se que a computação exaustiva decorrente do cálculo de todas as distâncias seja realizada.

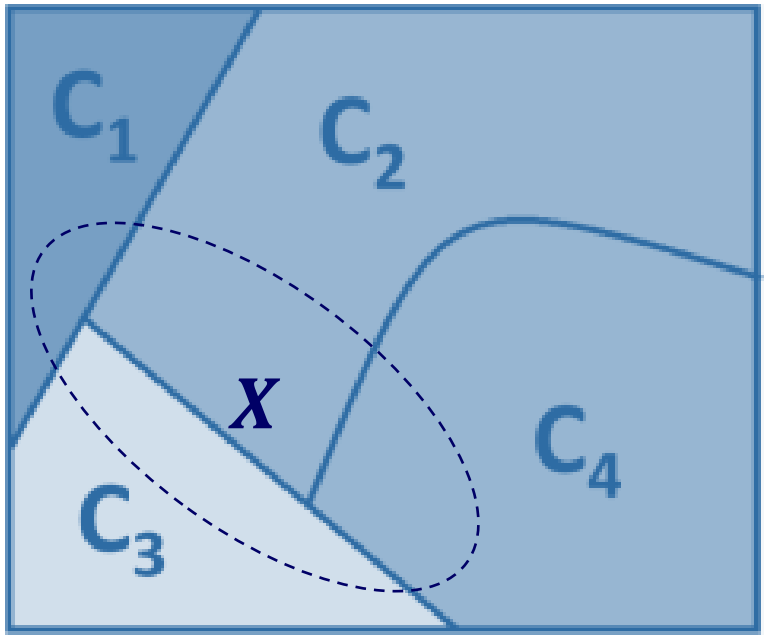
Ele é facilmente paralelizado o que promove uma grande melhoria de desempenho.

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

São algoritmos de classificação simples, rápidos e de bom desempenho, que usam o Teorema de Bayes para a tomada de decisão supondo a independência entre os atributos.

Eles consideram que cada atributo contribui de modo independente para a probabilidade de um objeto pertencer a uma classe, o que motiva a sua qualificação como “ingênuo”.



Teorema de Bayes

$$P(C_k|\boldsymbol{x}) = \frac{P(C_k \cap \boldsymbol{x})}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|C_k) \cdot P(C_k)}{P(\boldsymbol{x})}$$

em que

Informações conhecidas

Informações supostamente conhecidas

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} P(\boldsymbol{x}|C_k) \cdot P(C_k)$$

A classe com maior probabilidade $P(C_k|\boldsymbol{x}_0)$ será atribuída a um novo registro $\boldsymbol{x}_0$.

sklearn.naive_bayes

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

Como $P(\boldsymbol{x})$ é igual para toda classe, basta calcular $P(C_k|\boldsymbol{x}) = P(\boldsymbol{x}|C_k) \cdot P(C_k)$ durante a classificação.

Pela suposição simplista de independência de atributos, esses são representados por eventos independentes e, portanto:

$$P(\boldsymbol{x}|C_k) \approx \prod_{i=1}^{n} P(x_i|C_k) = P(x_1|C_k) \cdot \ldots \cdot P(x_n|C_k), \, i \text{ atributo de } \boldsymbol{x}$$

Se a variável x_i for discreta, $P(\boldsymbol{x}_i|C_k)$ é a frequência relativa de x_i em C_k.

Se a variável x_i for contínua, $P(x_i|C_k)$ é normalmente estimada pela função densidade de probabilidade de uma distribuição gaussiana:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi}\sigma_{C_k}} e^{-\frac{(x_i-\mu_{C_k})^2}{2\sigma_{C_k}^2}}$$

As diferentes abordagens ao classificador bayesiano simples referem-se, principalmente, a forma de estimar a distribuição de $P(x_i|C_k)$.

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

Exemplo:

Registro	Sexo	Curso	Idiomas	Trabalha	Classe
1	M	Eng.Civil	1	N	A
2	M	Eng.Elétrica	1	N	A
3	M	Eng.Elétrica	2	N	B
4	F	Eng.Elétrica	2	N	A
5	F	Eng.Civil	2	N	A
6	M	Eng.Elétrica	2	S	B
7	M	Eng.Elétrica	1	S	B
8	M	Eng.Elétrica	1	S	C
9	M	Eng.Civil	1	S	C
10	M	Eng.Civil	2	S	C
11	F	Eng.Civil	1	S	C
12	F	Eng.Civil	1	N	C
13	F	Eng.Civil	1	S	B
14	F	Eng.Civil	3	S	A
15	M	Eng.Civil	1	S	A

$P(C_A)=0,40$
$P(C_B)=0,27$
$P(C_C)=0,33$
$P(M/C_A)=0,50$
$P(F/C_A)=0,50$
$P(M/C_B)=0,75$
$P(F/C_B)=0,25$
$P(M/C_C)=0,60$
$P(F/C_C)=0,40$
$P(Eng.Elét/C_A)=0,33$
$P(Eng.Civil/C_A)=0,67$
$P(Eng.Elét/C_B)=0,75$
$P(Eng.Civil/C_B)=0,25$
$P(Eng.Elét/C_C)=0,20$
$P(Eng.Civil/C_C)=0,80$

$P(1/C_A)=0,50$
$P(2/C_A)=0,33$
$P(3/C_A)=0,17$
$P(1/C_B)=0,50$
$P(2/C_B)=0,50$
$P(3/C_B)=0,00$
$P(1/C_C)=0,80$
$P(2/C_C)=0,20$
$P(3/C_C)=0,00$
$P(N/C_A)=0,67$
$P(S/C_A)=0,33$
$P(N/C_B)=0,25$
$P(S/C_B)=0,75$
$P(N/C_C)=0,20$
$P(S/C_C)=0,80$

modelo produzido

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

Exemplo:

Registro	Sexo	Curso	Idiomas	Trabalha	Classe
X	F	Eng.Civil	2	S	?

$$P(C_k|\boldsymbol{x}) = \frac{P(C_k \cap \boldsymbol{x})}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|C_k) \cdot P(C_k)}{P(\boldsymbol{x})}$$

$$P(\boldsymbol{x}|C_A) = P(F|C_A) \cdot P(E.Civil|C_A) \cdot P(2|C_A) \cdot P(S|C_A) = 0{,}50 \cdot 0{,}67 \cdot 0{,}33 \cdot 0{,}33 = 0{,}036$$
$$P(\boldsymbol{x}|C_B) = P(F|C_B) \cdot P(E.Civil|C_B) \cdot P(2|C_B) \cdot P(S|C_B) = 0{,}25 \cdot 0{,}25 \cdot 0{,}50 \cdot 0{,}75 = 0{,}023$$
$$P(\boldsymbol{x}|C_C) = P(F|C_C) \cdot P(E.Civil|C_C) \cdot P(2|C_C) \cdot P(S|C_C) = 0{,}40 \cdot 0{,}80 \cdot 0{,}20 \cdot 0{,}80 = 0{,}051$$

$$P(C_A|\boldsymbol{x}) = P(\boldsymbol{x}|C_A) \cdot P(C_A) = 0{,}036 \cdot 0{,}40 = 0{,}014$$
$$P(C_B|\boldsymbol{x}) = P(\boldsymbol{x}|C_B) \cdot P(C_B) = 0{,}023 \cdot 0{,}27 = 0{,}006$$
$$P(C_C|\boldsymbol{x}) = P(\boldsymbol{x}|C_C) \cdot P(C_C) = 0{,}051 \cdot 0{,}33 = 0{,}017$$

classe C

$P(C_A)$=0,40
$P(C_B)$=0,27
$P(C_C)$=0,33

$P(M/C_A)$=0,50
$P(F/C_A)$=0,50

$P(M/C_B)$=0,75
$P(F/C_B)$=0,25

$P(M/C_C)$=0,60
$P(F/C_C)$=0,40

$P(Eng.Elét/C_A)$=0,33
$P(Eng.Civil/C_A)$=0,67

$P(Eng.Elét/C_B)$=0,75
$P(Eng.Civil/C_B)$=0,25

$P(Eng.Elét/C_C)$=0,20
$P(Eng.Civil/C_C)$=0,80

$P(N/C_A)$=0,67
$P(S/C_A)$=0,33
$P(N/C_B)$=0,25
$P(S/C_B)$=0,75
$P(N/C_C)$=0,20
$P(S/C_C)$=0,80

$P(1/C_A)$=0,50
$P(2/C_A)$=0,33
$P(3/C_A)$=0,17

$P(1/C_B)$=0,50
$P(2/C_B)$=0,50
$P(3/C_B)$=0,00

$P(1/C_C)$=0,80
$P(2/C_C)$=0,20
$P(3/C_C)$=0,00

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

Em problemas com dados reais, pode-se evitar as probabilidades nulas aplicando-se a correção de Laplace no atributo: a cada possível valor do atributo, uma unidade é adicionada a sua contagem e ao total de registros.

No exemplo anterior, se um novo registro a ser classificado tivesse o valor de um outro curso de graduação não presente nos dados originais, como Matemática, por exemplo, as contagens de Engenharia Civil e Engenharia Elétrica subiriam para 10 e 7, respectivamente, e o total seria 18.

Assim, as contagens e probabilidades afetadas, anteriores e novas, seriam as seguintes, respectivamente:

Classe	Eng. Civil	Eng. Elétrica	Matemática	Total
A	4	2	0	6
B	1	3	0	4
C	4	1	0	5

Classe	Eng. Civil	Eng. Elétrica	Matemática	Total
A	4+1	2+1	0+1	6+1+1+1
B	1+1	3+1	0+1	4+1+1+1
C	4+1	1+1	0+1	5+1+1+1

Classe	Eng. Civil	Eng. Elétrica	Matemática	Total
A	0,67	0,33	0	1
B	0,25	0,75	0	1
C	0,80	0,20	0	1

Classe	Eng. Civil	Eng. Elétrica	Matemática	Total
A	0,56	0,33	0,11	1
B	0,29	0,57	0,14	1
C	0,625	0,25	0,125	1

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

A adição de uma unidade apresentada no exemplo anterior não é obrigatória, mas é a mais comum. Pode-se considerar essa quantidade como um hiperparâmetro a ser determinado por validação cruzada. Outra possibilidade é realizar a adição de modo diferenciado por variável, conforme as características de cada uma.

Outro problema surge quando os produtórios se degeneram e, portanto, aplica-se a função logaritmo às probabilidades:

$$P(C_k|\boldsymbol{x}) = P(\boldsymbol{x}|C_k) \cdot P(C_k)$$

$$\ln\big(P(C_k|\boldsymbol{x})\big) = \ln\big(P(\boldsymbol{x}|C_k) \cdot P(C_k)\big) = \ln P(\boldsymbol{x}|C_k) + \ln P(C_k) =$$

linearidade

$$\ln \prod_{i=1}^{n} P(x_i|C_k) + \ln P(C_k) = \ln P(x_1|C_k) + \cdots + \ln P(x_n|C_k) + \ln P(C_k)$$

Modelagem bayesiana

Classificador bayesiano simples (*naïve Bayes classifier*)

Uma outra maneira de categorizar as modelagens preditivas é qualificá-las como discriminativas e generativas.

Considerando a tarefa de classificação, modelos discriminativos aprendem a determinar as fronteiras de decisão que separam as diferentes classes nos dados de entrada. Na tarefa de regressão, eles estimam diretamente a relação entre as variáveis de entrada e a variável de saída.

Já os generativos modelam a distribuição de probabilidade conjunta dos dados de entrada e de saída, estimando os processos geradores desses dados, o que permite produzir novos dados semelhantes aos que foram utilizado durante o aprendizado.

Assim, classificadores bayesianos são da categoria generativa e podem ser utilizados para gerar novas amostras de dados a calculados a partir de $P(C_k)$ e $P(\boldsymbol{x}_i|C_k)$ originais dos dados e de $P(\boldsymbol{x}|C_k)$ e $P(C_k|\boldsymbol{x})$ calculados. Considerando o exemplo apresentado, podemos usar o modelo para gerar a classe do registro abaixo fazendo um sorteio da classe a ser atribuída baseando-se nas $P(C_k|\boldsymbol{x})$ calculadas e normalizadas:

Registro	Sexo	Curso	Idiomas	Trabalha	Classe
X	F	Eng.Civil	2	S	?

$$\begin{matrix} P(C_A|\boldsymbol{x}) = 0{,}014 \\ P(C_B|\boldsymbol{x}) = 0{,}006 \\ P(C_C|\boldsymbol{x}) = 0{,}017 \end{matrix} \xrightarrow[(\div P(\boldsymbol{x}))]{\text{normalização}} \begin{matrix} P(C_A|\boldsymbol{x}) = 38\% \\ P(C_B|\boldsymbol{x}) = 16\% \\ P(C_C|\boldsymbol{x}) = 46\% \end{matrix}$$

Modelos lineares

Regressão linear

Modelagem que procura estabelecer uma relação linear entre as variáveis de entrada, independentes, e a variável de saída, dependente. Ela é simples quando há uma única independente, ou múltipla, quando há várias.

Problema real: $y = f(x_1, \dots, x_n) + \varepsilon$

- y é a variável dependente (de decisão, de resposta, explicada, endógena);
- $x_1, \dots, x_n$ são as variáveis independentes (preditoras, regressoras, explicativas);
- f é desconhecida;
- ε é um termo de erro, assumido como sendo independente de $\boldsymbol{x} = [x_1, \dots, x_n]$.

Modelo linear aproximado: $y \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$

- modelo linear populacional: a melhor aproximação linear desconhecida do modelo real original;
- $\beta_0, \dots, \beta_n$ são os parâmetros desconhecidos do modelo;
- ε é um termo de erro que abrange tudo o que $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ não conseguiu capturar.

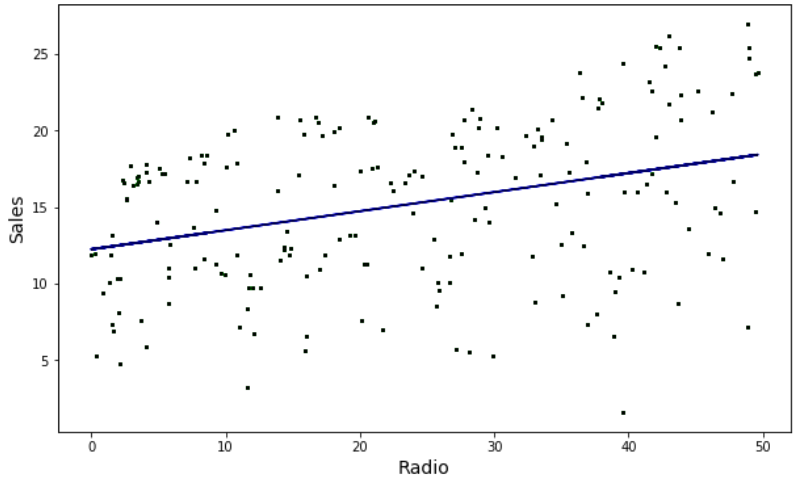
Modelos lineares

Regressão linear

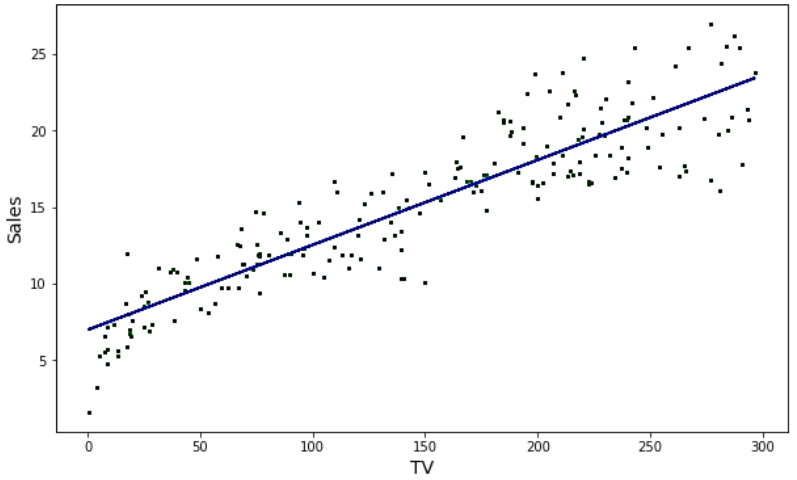
Modelo linear estimado: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_n} x_n$

- modelo estimado da aproximação $y \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ a partir dos dados disponíveis para gerá-lo;
- $\widehat{\beta_0}, \ldots, \widehat{\beta_n}$ são os parâmetros ou coeficientes que estimam $\beta_0, \ldots, \beta_n$ do modelo aproximado;
- $\widehat{\beta_0}$ é chamado de intercepto, ou seja, é o valor da estimativa da função quando os efeitos de todas as variáveis são nulos;
- Com $\widehat{\beta_0}$ nulo, assume-se que a saída é proporcional às entradas.

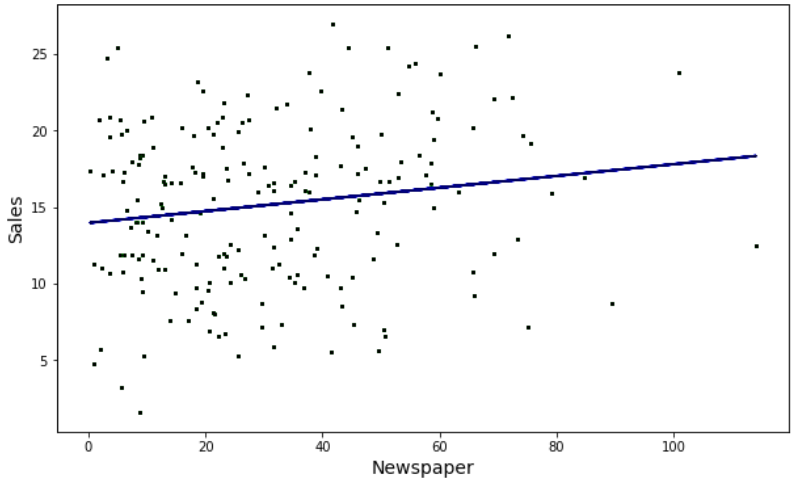
Exemplo: *Advertising dataset*, 200 registros, 4 atributos



$Sales = 12{,}236 + 0{,}124\ Radio$



$Sales = 6{,}975 + 0{,}056\ TV$



$Sales = 13{,}960 + 0{,}038\ Newspaper$

Modelos lineares

Regressão linear

Existe relação entre as variáveis dependentes e independentes?
Se os dados apontam que não há relação, descarta-se a variável.

Se houver relação, ela é linear?
Se a relação tem comportamento aproximadamente "linear", então a regressão linear é interessante. Se não, talvez seja melhor transformar algum preditor ou mesmo a variável de resposta para que a regressão linear seja melhor aplicada.

Se houver relação, qual a intensidade?
Se a intensidade é forte, o uso da relação favorece acertos do modelo. Se é fraca, o modelo pouco difere de uma decisão ao acaso.

Qual o efeito de cada variável preditora na variável de decisão?
A partir da quantificação da contribuição de cada variável pode-se estudar seu efeito e o quão correta ela é.

O quão correta é a resposta do modelo?
O uso de medidas e testes estatísticos adequados permitem saber o quanto alguma resposta do modelo tem correção e a sua confiabilidade.

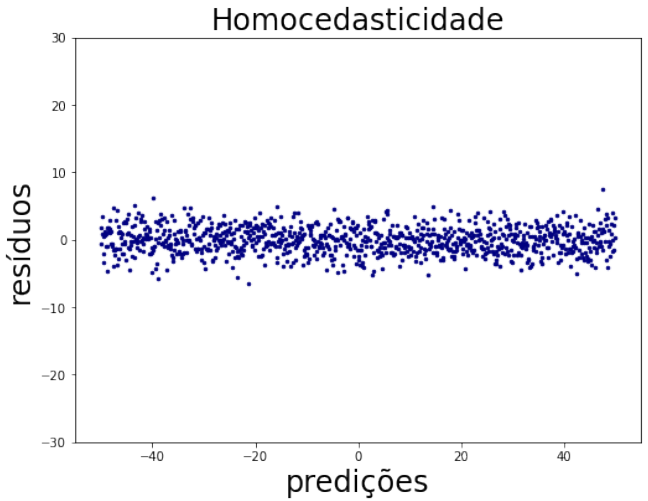
Há sinergia entre variáveis?
É possível que interações entre variáveis produzam efeitos mais fortes na resposta de um modelo.

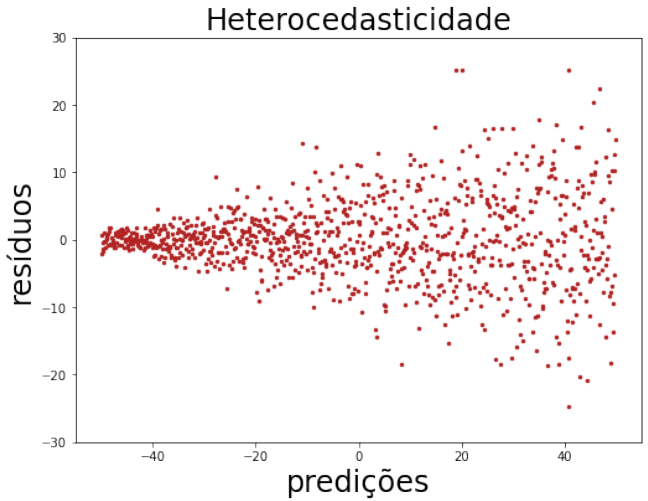
Modelos lineares

Regressão linear

Para a determinação do hiperplano que melhor descreva a relação entre as variáveis independentes e a dependente, diversos algoritmos de estimação de parâmetros podem ser utilizados e é comum que eles façam certas suposições sobre os comportamentos das variáveis:

- ausência de multicolinearidade: as variáveis preditoras não possuem fortes correlações entre si, simples ou coletivas;
- exogeneidade dos regressores: as variáveis preditoras não são correlacionadas aos erros de observação;
- linearidade de parâmetros: a variável explicada é combinação linear dos parâmetros;
- independência e normalidade de erros: os resíduos não são correlacionados e devem ser normalmente distribuídos em torno de zero;
- homocedasticidade: a variância dos resíduos não depende dos valores das variáveis.



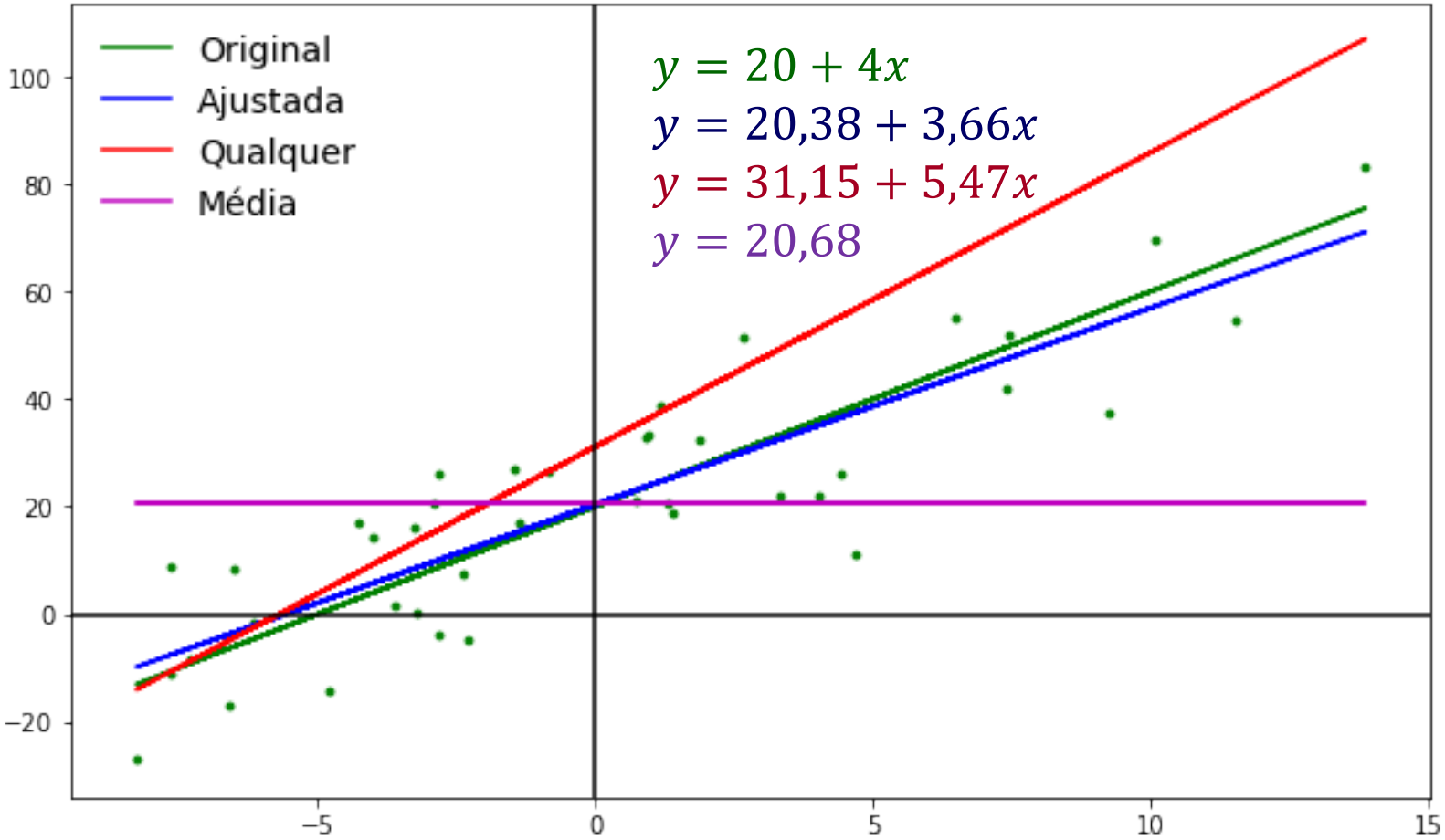


Muitas vezes, as suposições são relaxadas ou ignoradas para a aplicação de um algoritmo, promovendo a realização de operações que o tornam dependente de mais dados devido a sua maior complexidade.

Modelos lineares

Regressão linear

Exemplo: dados sintéticos, 40 registros, 2 atributos



Os passos da modelagem linear podem ser assim resumidos:

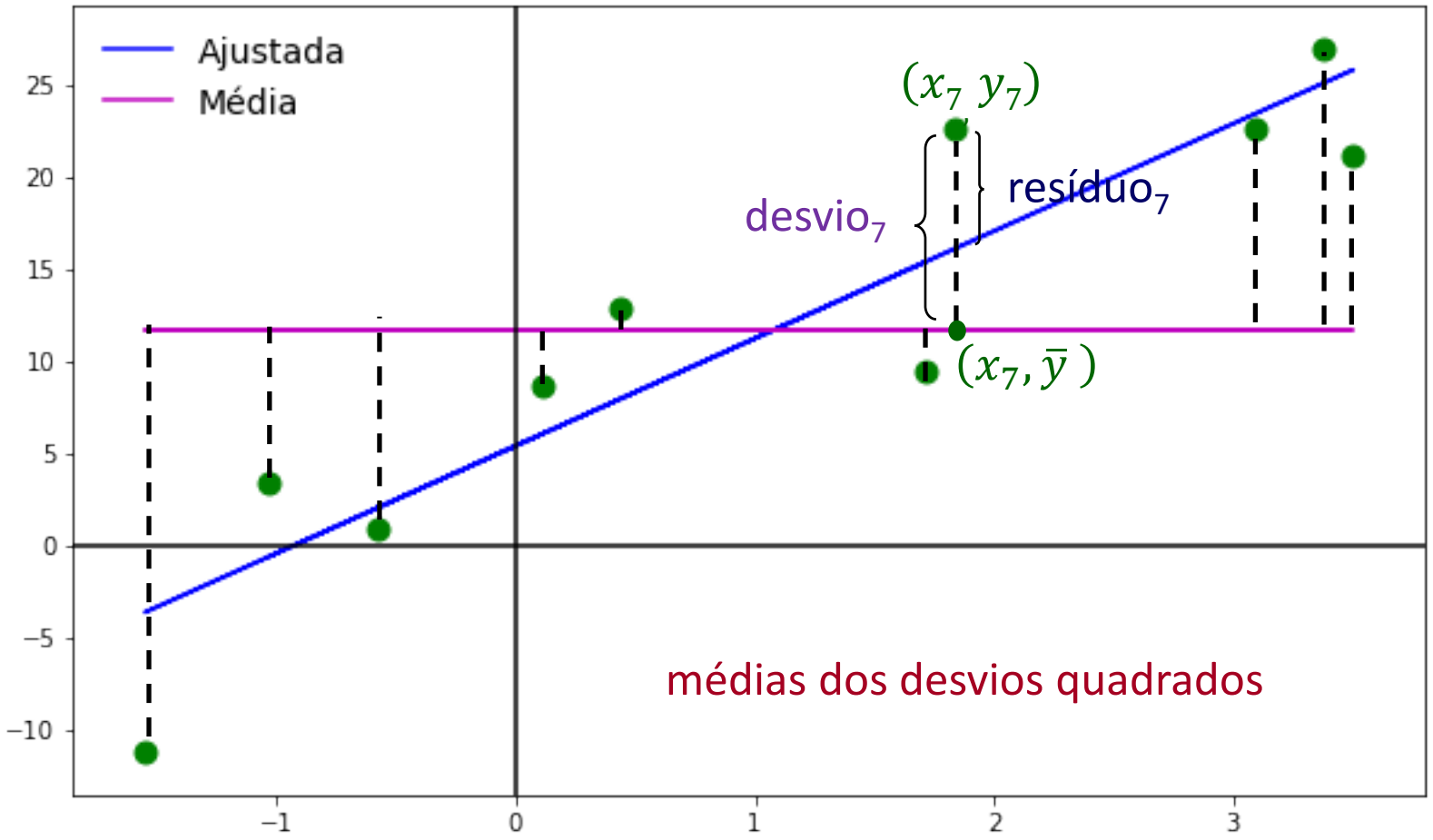
a) ajustar um modelo linear aos dados;
b) avaliar o desempenho do modelo;
c) avaliar a significância do desempenho.

statsmodels.api
sklearn.linear_model.LinearRegression

Modelos lineares

Regressão linear

Exemplo: dados sintéticos, 10 registros, 2 atributos



$$Var(y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

$$EMQ(\hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

$$Var(y) = 128{,}23$$

$$EMQ(\hat{y}) = 20{,}96$$

Qualquer reta diferente da ajustada terá EMQ maior que o dela.

Regressão linear

Como o erro médio quadrático é uma função quadrática dos coeficientes de regressão, estes podem ser calculados usando as derivadas parciais da função de erro em relação a cada um e resolvidos de forma analítica. Portanto, os coeficientes ótimos são descobertos pela minimização dos resíduos quadrados:

$$\min_{\widehat{\beta_0},\widehat{\beta_1}} EMQ(\widehat{\beta_0},\widehat{\beta_1}) = \min_{\widehat{\beta_0},\widehat{\beta_1}} \sum_{i=1}^{N} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$$

$$\frac{\partial EMQ}{\partial parâmetros} = 0 \Rightarrow \frac{\partial EMQ}{\partial \widehat{\beta_0}} = 0 \quad \text{e} \quad \frac{\partial EMQ}{\partial \widehat{\beta_1}} = 0$$

EMQ da reta melhor ajustada aos dados

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x,y)}{\sigma_x^2}$$

Forma vetorial:

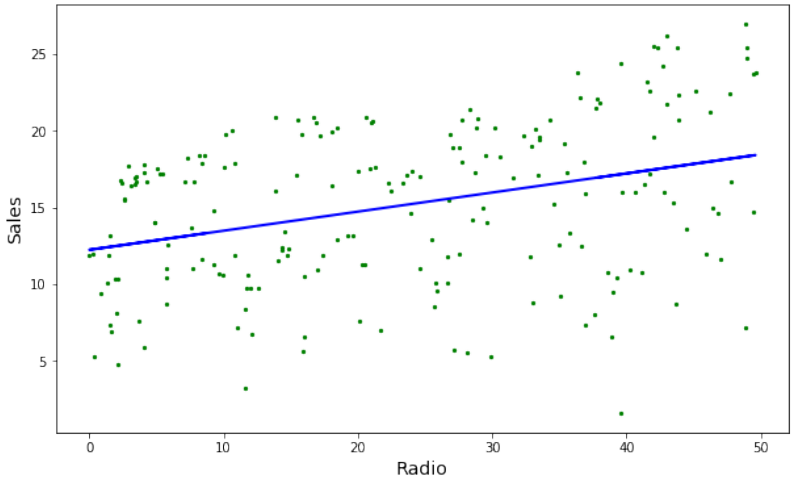
$$EMQ(\widehat{\boldsymbol{\beta}}) = \frac{1}{N}\left(\mathbf{Y} - \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{Y} - \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}^T\boldsymbol{Y}$$

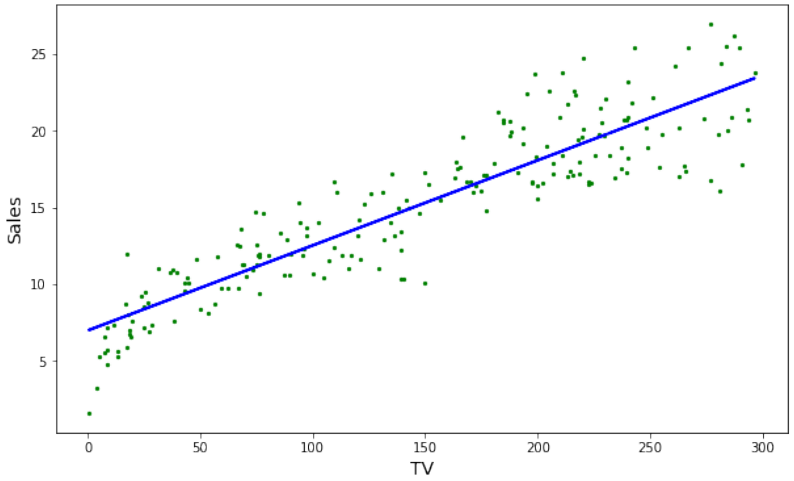
Modelos lineares

Regressão linear

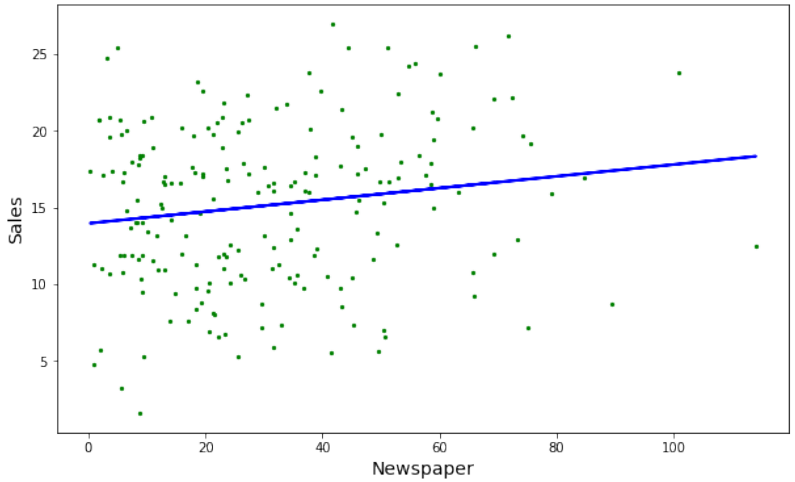
Exemplo: *Advertising dataset*, 200 registros, 4 atributos



$Sales = 12{,}236 + 0{,}124\ Radio$



$Sales = 6{,}975 + 0{,}056\ TV$



$Sales = 13{,}960 + 0{,}038\ Newspaper$

$\overline{sales} = 15{,}13$ $\quad \bar{tv} = 147{,}04,$ $\quad cov(tv, sales) = 408{,}83,$ $\quad \sigma^2_{tv} = 7370{,}95$

$\overline{radio} = 23{,}264,$ $\quad cov(radio, sales) = 27{,}43,$ $\quad \sigma^2_{radio} = 220{,}43$

$\overline{newspaper} = 30{,}554,$ $\quad cov(newspaper, sales) = 18{,}18,$ $\quad \sigma^2_{newspaper} = 474{,}31$

Regressão linear

Pode-se confiar na correção dos parâmetros encontrados?

A média aritmética é um estimador não enviesado de um conjunto de valores. Assim, as médias dos coeficientes obtidos a partir de uma grande quantidade de regressões produzirão valores muito próximos aos dos coeficientes reais/populacionais desconhecidos.

Entretanto, não serão realizadas inúmeras regressões e pode-se estimar o erro padrão de cada parâmetro, que possui as seguintes utilidades:

- informar o quanto o valor do parâmetro está distante do valor real desconhecido;
- encontrar o intervalo de confiança do parâmetro;
- ser usado em testes de hipóteses sobre o parâmetro para verificar se há relação entre a variável dependente e independente.

$$SE(\widehat{\beta_0}) = \sqrt{Var(\varepsilon)\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]} \qquad SE(\widehat{\beta_1}) = \sqrt{\frac{Var(\varepsilon)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad Var(\varepsilon) \cong \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-2}$$

Regressão linear

Os intervalos com 95% de confiança para β_0 e β_1 são, por praticidade:

$$\left[\widehat{\beta_0} - 2 \cdot SE\left(\widehat{\beta_0}\right), \widehat{\beta_0} + 2 \cdot SE\left(\widehat{\beta_0}\right)\right] \quad \text{e} \quad \left[\widehat{\beta_1} - 2 \cdot SE\left(\widehat{\beta_1}\right), \widehat{\beta_1} + 2 \cdot SE\left(\widehat{\beta_1}\right)\right]$$

Seria necessário calcular a estatística $t = \left(\widehat{\beta_i} - \beta_i\right)/SE\left(\widehat{\beta_i}\right)$ e consultar a distribuição de Student com $n - 2$ graus de liberdade para se obter o valor crítico tal que:

$$P\left(\widehat{\beta_i} - t_{crítico}(n-2) \cdot SE\left(\widehat{\beta_i}\right) \leq \beta_i \leq \widehat{\beta_i} + t_{crítico}(n-2) \cdot SE\left(\widehat{\beta_i}\right)\right) = 1 - nível\ de\ significância$$

Considerando o valor típico de 95% de confiança (5% de significância), para amostras acima de 30 elementos o valor crítico é aproximadamente 2 (de 2,048, com 28 graus de liberdade, a 1,960, com infinitos graus de liberdade). Assim, adota-se 2 como uma boa aproximação.

Distribuição t-Student: valores **tc** tais que P(-**tc** ≤ t ≤ **tc**) = 1 - p

Graus de Liberdade	90%	80%	70%	60%	50%	40%	30%	20%	10%	8%	6%	5%	4%	2%	1%	0,2%	0,1%
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	7,916	10,579	12,706	15,895	31,821	63,657	318,309	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	3,320	3,896	4,303	4,849	6,965	9,925	22,327	31,599
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	2,605	2,951	3,182	3,482	4,541	5,841	10,215	12,924
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,333	2,601	2,776	2,999	3,747	4,604	7,173	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,191	2,422	2,571	2,757	3,365	4,032	5,893	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,104	2,313	2,447	2,612	3,143	3,707	5,208	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,046	2,241	2,365	2,517	2,998	3,499	4,785	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,004	2,189	2,306	2,449	2,896	3,355	4,501	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	1,973	2,150	2,262	2,398	2,821	3,250	4,297	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	1,948	2,120	2,228	2,359	2,764	3,169	4,144	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	1,928	2,096	2,201	2,328	2,718	3,106	4,025	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	1,912	2,076	2,179	2,303	2,681	3,055	3,930	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	1,899	2,060	2,160	2,282	2,650	3,012	3,852	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	1,887	2,046	2,145	2,264	2,624	2,977	3,787	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	1,878	2,034	2,131	2,249	2,602	2,947	3,733	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	1,869	2,024	2,120	2,235	2,583	2,921	3,686	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	1,862	2,015	2,110	2,224	2,567	2,898	3,646	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	1,855	2,007	2,101	2,214	2,552	2,878	3,610	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	1,850	2,000	2,093	2,205	2,539	2,861	3,579	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	1,844	1,994	2,086	2,197	2,528	2,845	3,552	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	1,840	1,988	2,080	2,189	2,518	2,831	3,527	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	1,835	1,983	2,074	2,183	2,508	2,819	3,505	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	1,832	1,978	2,069	2,177	2,500	2,807	3,485	3,768
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	1,828	1,974	2,064	2,172	2,492	2,797	3,467	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	1,825	1,970	2,060	2,167	2,485	2,787	3,450	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	1,822	1,967	2,056	2,162	2,479	2,779	3,435	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	1,819	1,963	2,052	2,158	2,473	2,771	3,421	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	1,817	1,960	2,048	2,154	2,467	2,763	3,408	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	1,814	1,957	2,045	2,150	2,462	2,756	3,396	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	1,812	1,955	2,042	2,147	2,457	2,750	3,385	3,646
31	0,127	0,256	0,389	0,530	0,682	0,853	1,054	1,309	1,696	1,810	1,952	2,040	2,144	2,453	2,744	3,375	3,633
32	0,127	0,255	0,389	0,530	0,682	0,853	1,054	1,309	1,694	1,808	1,950	2,037	2,141	2,449	2,738	3,365	3,622
33	0,127	0,255	0,389	0,530	0,682	0,853	1,053	1,308	1,692	1,806	1,948	2,035	2,138	2,445	2,733	3,356	3,611
34	0,127	0,255	0,389	0,529	0,682	0,852	1,052	1,307	1,691	1,805	1,946	2,032	2,136	2,441	2,728	3,348	3,601
35	0,127	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	1,803	1,944	2,030	2,133	2,438	2,724	3,340	3,591
36	0,127	0,255	0,388	0,529	0,681	0,852	1,052	1,306	1,688	1,802	1,942	2,028	2,131	2,434	2,719	3,333	3,582
37	0,127	0,255	0,388	0,529	0,681	0,851	1,051	1,305	1,687	1,800	1,940	2,026	2,129	2,431	2,715	3,326	3,574
38	0,127	0,255	0,388	0,529	0,681	0,851	1,051	1,304	1,686	1,799	1,939	2,024	2,127	2,429	2,712	3,319	3,566
39	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,304	1,685	1,798	1,937	2,023	2,125	2,426	2,708	3,313	3,558
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	1,796	1,936	2,021	2,123	2,423	2,704	3,307	3,551
45	0,126	0,255	0,388	0,528	0,680	0,850	1,049	1,301	1,679	1,791	1,929	2,014	2,115	2,412	2,690	3,281	3,520
50	0,126	0,255	0,388	0,528	0,679	0,849	1,047	1,299	1,676	1,787	1,924	2,009	2,109	2,403	2,678	3,261	3,496
55	0,126	0,255	0,387	0,527	0,679	0,848	1,046	1,297	1,673	1,784	1,920	2,004	2,104	2,396	2,668	3,245	3,476
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	1,781	1,917	2,000	2,099	2,390	2,660	3,232	3,460
70	0,126	0,254	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,776	1,912	1,994	2,093	2,381	2,648	3,211	3,435
80	0,126	0,254	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,773	1,908	1,990	2,088	2,374	2,639	3,195	3,416
90	0,126	0,254	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,771	1,905	1,987	2,084	2,368	2,632	3,183	3,402
100	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,769	1,902	1,984	2,081	2,364	2,626	3,174	3,390
110	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,659	1,767	1,900	1,982	2,078	2,361	2,621	3,166	3,381
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,766	1,899	1,980	2,076	2,358	2,617	3,160	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,751	1,881	1,960	2,054	2,326	2,576	3,090	3,291

Regressão linear

Faz-se necessário também verificar se a variável independente associada ao parâmetro estimado é realmente relevante na explicação da variável dependente por meio de um teste de hipótese:

H_0: Não há relação entre x e y $(\beta_1 = 0)$ e H_1 : Há relação entre x e y $(\beta_1 \neq 0)$

É preciso verificar se $\widehat{\beta_1}$ é suficientemente distante de zero para se ter confiança que β_1 não seja zero. Supondo que não exista relação, ou seja, $\beta_1 = 0$, então $y = \beta_0 + \varepsilon$

Calcula-se a estatística $t = \left(\widehat{\beta_i} - 0\right)/SE\left(\widehat{\beta_i}\right)$ para saber a quantos desvios-padrão $\widehat{\beta_1}$ está distante de zero.

Com o nível de significância do teste e $n - 2$ graus de liberdade, consultar a distribuição de Student para se obter o valor crítico. Se t não estiver na região crítica do teste, rejeita-se a hipótese nula. Alternativamente, o valor p da distribuição t de Student, com $n - 2$ graus de liberdade deve ser menor que o nível de significância para a hipótese nula ser rejeitada.

Obviamente, se $\widehat{\beta_1}$ for distante de zero e se $SE\left(\widehat{\beta_1}\right)$ for bem pequeno, é praticamente certo de que a hipótese nula será rejeitada.

Modelos lineares

Regressão linear

Exemplo: *Advertising dataset*, 200 registros, 4 atributos

A partir da distribuição t com 99.9% de confiança e 198 graus de liberdade, pode-se observar que o valor p do coeficiente da variável TV é menor que o nível de significância do teste (0,001). Ou seja, de maneira rigorosa, a hipótese de que não há relação entre anúncios de TV e vendas seria rejeitada.

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.811
Method:	Least Squares	F-statistic:	856.2
Date:	Thu, 26 May 2022	Prob (F-statistic):	7.93e-74
Time:	21:47:47	Log-Likelihood:	-448.99
No. Observations:	200	AIC:	902.0
Df Residuals:	198	BIC:	908.6
Df Model:	1		
Covariance Type:	nonrobust		

$$sales = 6{,}9748 + 0{,}0555 \cdot TV$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.9748	0.323	21.624	0.000	6.339	7.611
TV	0.0555	0.002	29.260	0.000	0.052	0.059

Alternativamente, a região de aceitação de β_1 é $[-3{,}291;\ 3{,}291]$. Portanto, como t de β_1 está fora dela, pode-se rejeitar H_0 e afirmar que não há evidências de não haver relação entre as variáveis, ou seja, pode-se sugerir fortemente que há relação entre anúncios nas estações de TV e as vendas.

Modelos lineares

Regressão linear

Exemplo: *Advertising dataset*, 200 registros, 4 atributos

A partir da distribuição t com 99.9% de confiança e 198 graus de liberdade, pode-se observar que o valor p do coeficiente da variável $Radio$ é menor que o nível de significância do teste (0,001). Ou seja, de maneira rigorosa, a hipótese de que não há relação entre anúncios de rádio e vendas seria rejeitada.

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.122
Model:	OLS	Adj. R-squared:	0.118
Method:	Least Squares	F-statistic:	27.57
Date:	Thu, 26 May 2022	Prob (F-statistic):	3.88e-07
Time:	22:05:27	Log-Likelihood:	-603.18
No. Observations:	200	AIC:	1210.
Df Residuals:	198	BIC:	1217.
Df Model:	1		
Covariance Type:	nonrobust		

$$sales = 12{,}2357 + 0{,}1244 \cdot Radio$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.2357	0.653	18.724	0.000	10.947	13.524
Radio	0.1244	0.024	5.251	0.000	0.078	0.171

Alternativamente, a região de aceitação de β_1 é $[-3{,}291;\ 3{,}291]$. Portanto, como t de β_1 está fora dela, pode-se rejeitar H_0 e afirmar que não há evidências de não haver relação entre as variáveis, ou seja, pode-se sugerir fortemente que há relação entre anúncios nas estações de rádio e as vendas.

Modelos lineares

Regressão linear

Exemplo: *Advertising dataset*, 200 registros, 4 atributos

A partir da distribuição t com 99.9% de confiança e 198 graus de liberdade, pode-se observar que o valor p do coeficiente da variável $Newspaper$ é maior que o nível de significância do teste (0,001). Ou seja, de maneira rigorosa, a hipótese de que não há relação entre anúncios de rádio e vendas seria admitida.

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.025
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	5.067
Date:	Thu, 26 May 2022	Prob (F-statistic):	0.0255
Time:	22:10:49	Log-Likelihood:	-613.69
No. Observations:	200	AIC:	1231.
Df Residuals:	198	BIC:	1238.
Df Model:	1		
Covariance Type:	nonrobust		

$$sales = 13{,}9595 + 0{,}0383 \cdot Newspaper$$

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.9595	0.638	21.870	0.000	12.701	15.218
Newspaper	0.0383	0.017	2.251	0.025	0.005	0.072

Alternativamente, a região de aceitação de β_1 é $[-3{,}291;\ 3{,}291]$. Portanto, como t de β_1 está dentro dela, não se pode rejeitar H_0, restando afirmar que há evidências de não haver relação entre anúncios em jornais impressos e as vendas. Entretanto, se for utilizada outro critério menos rigoroso, 95% de confiança por exemplo, sendo $[-1{,}972; 1{,}972]$ a região de aceitação, pode-se rejeitar H_0.

Regressão linear

O quanto um modelo se ajusta aos dados? Avaliação da correção do modelo

O erro padrão residual (RSE) é uma estimativa do desvio padrão dos erros de um modelo de regressão. Mesmo que os parâmetros populacionais fossem conhecidos, ainda existiria variabilidade nos valores de $\boldsymbol{Y}$ devido a fatores não modelados. Ele representa, em média, o quanto as observações reais tendem a se afastar da reta de regressão verdadeira. Por definição, ele é o desvio-padrão dos resíduos:

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}} \cong \sqrt{Var(\varepsilon)}$$

em que p é a quantidade de variáveis (ou parâmetros sem o intercepto) e n a quantidade de registros.

Quanto menor o seu valor, melhor o ajuste do modelo aos dados pois os valores observados estarão próximos das previsões feitas pelo modelo. Portanto, ele é medida de "desajuste" do modelo: quanto maior o seu valor, maior esse desajuste. Ele não é uma medida padronizada, mas absoluta, o que facilita a interpretação direta, mas dificulta a comparação do seu valor entre modelos ajustados para diferentes problemas.

Modelos lineares

Regressão linear

O quanto um modelo se ajusta aos dados?

Avaliação da correção do modelo

Para medir diretamente o ajuste de um modelo aos dados, o coeficiente de determinação (R^2) mede a proporção de variabilidade da variável de decisão que é "capturada" pela variabilidade das estimativas, ou, em outras palavras, o quanto da variabilidade da saída é explicada/removida quando se usam as variáveis preditoras.

variabilidade não explicada após a regressão

$$R^2 = \frac{Var(y) - EMQ}{Var(y)} = \frac{Var(média) - Var(modelo)}{Var(média)} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

proporção de variabilidade explicada ou removida ao se utilizar as variáveis preditoras

variabilidade antes da regressão

$$R_a^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

N registros, p variáveis

Regressão linear

O quanto um modelo se ajusta aos dados?

Avaliação da correção do modelo

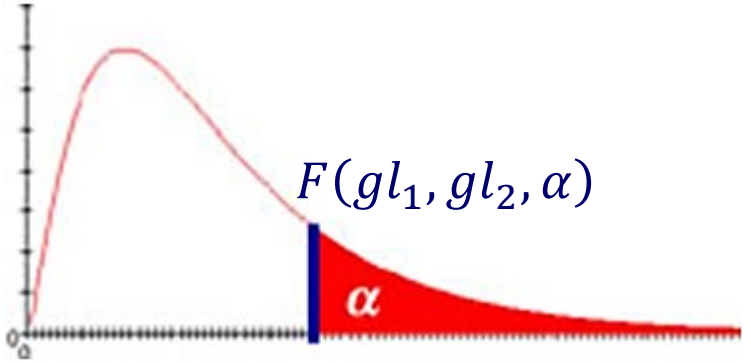
Para saber se os valores de R^2 são estatisticamente significantes, calcula-se a estatística F (Fisher), de comparação de variâncias, para estimar a quantos resíduos quadrados R^2 está distante de zero:

$$F = \frac{Var(média) - Var(modelo)}{Var(modelo)} \times \frac{n-k-1}{k} = \frac{Var(y) - EMQ}{EMQ} \times \frac{n-k-1}{k}$$

em que n é o tamanho da amostra e k é a quantidade de regressores excluindo-se o intercepto.

Dado um nível de significância α e dos graus de liberdade $gl_1 = k$ e $gl_2 = n - k - 1$, obter o valor crítico da distribuição F.

Se $F > F_{crítico}$, então o R^2 é significativo.



Modelos lineares

Regressão linear

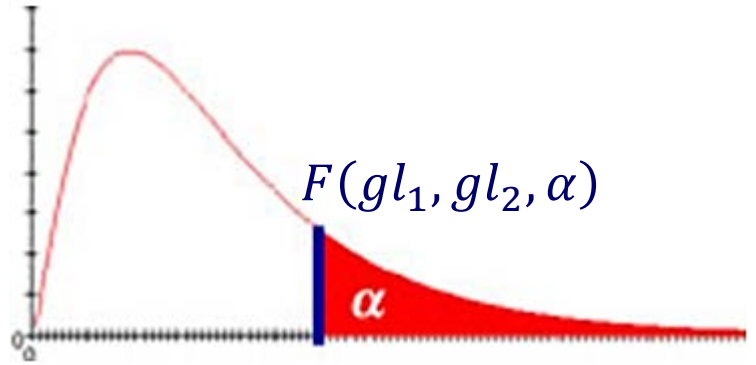
Exemplo: *Advertising dataset*, 200 registros, 4 atributos

$sales = 6{,}975 + 0{,}055\ TV$
$Var(sales) = 27{,}92, \quad RMQ = 5{,}22, \quad R^2 = 0{,}81$
$F = 861{,}03$
$F_{CRÍTICO}(1; 198; 0{,}01) = 6{,}635$

$sales = 12{,}236 + 0{,}124\ radio$
$Var(sales) = 27{,}92, \quad RMQ = 24{,}38, \quad R^2 = 0{,}12$
$F = 28{,}75$
$F_{CRÍTICO}(1; 198; 0{,}01) = 6{,}635$

$sales = 13{,}960 + 0{,}038\ newspaper$
$Var(sales) = 27{,}92, \quad RMQ = 27{,}09, \quad R^2 = 0{,}02$
$F = 6{,}07$
$F_{CRÍTICO}(1; 198; 0{,}01) = 6{,}635$

Se $F > F_{crítico}$, então o R^2 é significativo.



Regressão linear

O quanto um modelo se ajusta aos dados? Avaliação da correção do modelo

O R^2 ajustado penaliza modelos com mais variáveis, o que auxilia a seleção de modelos menos complexos e, consequentemente, menos propensos ao superajuste. Quando dois modelos apresentam valores de R^2 próximos, é importante realizar testes de hipóteses adicionais para avaliar se a diferença entre eles é estatisticamente significativa.

Como alternativa mais objetiva, pode-se utilizar os critérios AIC (Akaike *Information Criterion*) e BIC (*Bayesian Information Criterion*), que, diferentemente do R^2, comparam modelos com base no equilíbrio entre qualidade de ajuste e complexidade. Nesses critérios, modelos com menores valores de AIC ou BIC são considerados preferíveis, dispensando a necessidade de testes formais adicionais.

Ambos seguem a lógica: $Critério = Erro\ do\ modelo + Penalidade\ pela\ complexidade$

O erro do modelo é medido a partir da verossimilhança ($P(Dados\ Observados|Parâmetros)$, em que quanto maior o valor, melhor o ajuste). A complexidade é medida pela quantidade de parâmetros usados.

Modelos lineares

Regressão linear

O quanto um modelo se ajusta aos dados? Avaliação da correção do modelo

Akaike Information Criterion: $AIC(M) = -2 \cdot \ln \mathcal{L}(M) + 2 \cdot p$ modelo M, p parâmetros, n registros

Bayesian Information Criterion: $BIC(M) = -2 \cdot \ln \mathcal{L}(M) + p \cdot \ln n$ BIC é mais rigoroso que o AIC

$\mathcal{L}(M)$ é a função de máxima verossimilhança dos parâmetros do modelo, geralmente gaussiana. Dada a complexidade de cálculos e a rápida degeneração da função de verossimilhança provocada pelo produtório de probabilidades, as versões logarítmicas são as adotadas:

$$\mathcal{L}(M) = \prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}, s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot s^2}} e^{-\frac{(resíduo_i)^2}{2s^2}}$$

AIC tende a ser menos severo que o BIC quando um modelo usa mais parâmetros, ou seja, pode resultar em um valor menor (melhor) para modelos mais complexos desde que a melhoria no ajuste compense a penalização. No BIC, a penalidade cresce muito rapidamente, ultrapassando o valor de AIC a partir de 7 registros!

Modelos lineares

Regressão linear múltipla

Modelo linear estimado: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_n} x_n$

- modelo estimado da aproximação $y \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ a partir dos dados disponíveis para gerá-lo;
- $\widehat{\beta_j}$ pode ser visto como o efeito médio da variável x_j em $\hat{y}$ quando todos os demais parâmetros estão fixos;
- o uso de mais variáveis permite que o modelo se ajuste mais aos dados do que o modelo linear simples, independentemente da relevância estatística da variável.

Exemplo:
Advertising dataset,
200 registros,
4 atributos

$\alpha = 0{,}05$
$t_{crítico} = 1{,}653, gl = 196$
$F_{crítico} = 2{,}651$
$gl1 = 3, gl2 = 196$

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	605.4
Date:	Thu, 26 May 2022	Prob (F-statistic):	8.13e-99
Time:	21:30:29	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	774.7
Df Residuals:	196	BIC:	787.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6251	0.308	15.041	0.000	4.019	5.232
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012

$\alpha = 0{,}05$
$t_{crítico} = 1{,}653, gl = 197$
$F_{crítico} = 3{,}042\ gl1 = 2, gl2 = 197$

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.902
Method:	Least Squares	F-statistic:	912.7
Date:	Thu, 26 May 2022	Prob (F-statistic):	2.39e-100
Time:	21:33:56	Log-Likelihood:	-383.34
No. Observations:	200	AIC:	772.7
Df Residuals:	197	BIC:	782.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6309	0.290	15.952	0.000	4.058	5.203
TV	0.0544	0.001	39.726	0.000	0.052	0.057
Radio	0.1072	0.008	13.522	0.000	0.092	0.123

Regressão linear múltipla

Exemplo: Base de dados sobre os bairros do Rio de Janeiro, 152 registros, 24 atributos

OLS Regression Results

Dep. Variable:	Indice desenvolvimento social	R-squared:	0.664
Model:	OLS	Adj. R-squared:	0.661
Method:	Least Squares	F-statistic:	296.0
Date:	Thu, 26 May 2022	Prob (F-statistic):	2.56e-37
Time:	17:17:04	Log-Likelihood:	124.05
No. Observations:	152	AIC:	-244.1
Df Residuals:	150	BIC:	-238.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8916	0.028	31.422	0.000	0.836	0.948
Densidade domiciliar	-0.7714	0.045	-17.204	0.000	-0.860	-0.683

$\alpha = 0{,}05$

$t_{crítico} = 1{,}655, gl = 150$

$F_{crítico} = 3{,}904, gl1 = 1, gl2 = 150$

OLS Regression Results

Dep. Variable:	Indice desenvolvimento social	R-squared:	0.684
Model:	OLS	Adj. R-squared:	0.679
Method:	Least Squares	F-statistic:	161.0
Date:	Thu, 26 May 2022	Prob (F-statistic):	5.78e-38
Time:	19:43:27	Log-Likelihood:	128.70
No. Observations:	152	AIC:	-251.4
Df Residuals:	149	BIC:	-242.3
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7674	0.049	15.663	0.000	0.671	0.864
Densidade domiciliar	-0.6328	0.063	-10.073	0.000	-0.757	-0.509
Grau de verticalizacao	0.2325	0.076	3.068	0.003	0.083	0.382

$\alpha = 0{,}05$

$t_{crítico} = 1{,}655, gl = 150$

$F_{crítico} = 3{,}057, gl1 = 2, gl2 = 149$

Regressão linear múltipla

Exemplo: Base de dados sobre os bairros do Rio de Janeiro, 152 registros, 24 atributos

OLS Regression Results

Dep. Variable:	Indice desenvolvimento social	R-squared:	0.707
Model:	OLS	Adj. R-squared:	0.701
Method:	Least Squares	F-statistic:	118.8
Date:	Thu, 26 May 2022	Prob (F-statistic):	3.27e-39
Time:	19:53:32	Log-Likelihood:	134.41
No. Observations:	152	AIC:	-260.8
Df Residuals:	148	BIC:	-248.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8745	0.057	15.374	0.000	0.762	0.987
Densidade domiciliar	-0.7065	0.064	-10.958	0.000	-0.834	-0.579
Grau de verticalizacao	0.2322	0.073	3.171	0.002	0.088	0.377
Percentual de imoveis alugados	-0.1645	0.048	-3.397	0.001	-0.260	-0.069

$\alpha = 0{,}05$

$t_{crítico} = 1{,}655, gl = 150$

$F_{crítico} = 2{,}666, gl1 = 3, gl2 = 148$

$\alpha = 0{,}05$

$t_{crítico} = 1{,}655, gl = 150$

$F_{crítico} = 2.433, gl1 = 4, gl2 = 147$

OLS Regression Results

Dep. Variable:	Indice desenvolvimento social	R-squared:	0.708
Model:	OLS	Adj. R-squared:	0.700
Method:	Least Squares	F-statistic:	89.04
Date:	Thu, 26 May 2022	Prob (F-statistic):	2.80e-38
Time:	21:15:05	Log-Likelihood:	134.75
No. Observations:	152	AIC:	-259.5
Df Residuals:	147	BIC:	-244.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8668	0.058	15.010	0.000	0.753	0.981
Densidade domiciliar	-0.7182	0.066	-10.859	0.000	-0.849	-0.587
Grau de verticalizacao	0.2241	0.074	3.028	0.003	0.078	0.370
Percentual de imoveis alugados	-0.1752	0.050	-3.487	0.001	-0.274	-0.076
Percentual de area urbanizada	0.0270	0.033	0.811	0.419	-0.039	0.093

Regressão linear múltipla

Como lidar com interações entre variáveis?

Supondo que uma variável x_k tenha influência em outra variável x_l, pode-se relaxar a suposição aditiva dos modelos lineares:

$$\hat{y} = \widehat{\beta_0} + \cdots + \widehat{\beta_k} x_k + \widehat{\beta_l} x_l + \boxed{\widehat{\beta_{n+1}} x_k x_l} + \cdots + \widehat{\beta_n} x_n$$

termo de interação

Reescrevendo o modelo:

$$\hat{y} = \widehat{\beta_0} + \cdots + \widehat{\beta_k} x_k + \left(\widehat{\beta_l} + \widehat{\beta_{n+1}} x_k\right) x_l + \cdots + \widehat{\beta_n} x_n$$

Quando há interação, os coeficientes isolados das variáveis envolvidas deixam de ter interpretação direta.

Após o cálculo dos parâmetros, o teste de hipótese a respeito da relevância de $\widehat{\beta_{n+1}}$ pode indicar que a interação é estatisticamente insignificante e que o termo pode ser excluído da modelagem, ou seja, o modelo resultante é simplificado. Atenção: nunca excluir os termos originais ao manter o termo de interação.

Regressão linear múltipla

Como lidar com interações entre variáveis?

Interações como a exemplificada propiciam a ocorrência de multicolinariedade entre as variáveis envolvidas. Centralizar as variáveis originais ajuda a reduzir o problema sem afetar o desempenho da modelagem. O impacto será apenas nos valores dos parâmetros ajustados e nos erros padrão, que ficarão mais estáveis.

Outras não linearidades podem ser incorporadas, caracterizando não linearidades entre preditoras e preditas, e o modelo continua linear nos parâmetros. Por exemplo, $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}{x_1}^3 + \widehat{\beta_2}x_2 + (\widehat{\beta_3} + \widehat{\beta_4}x_2)x_3$ é um modelo linear chamado de polinômio de regressão com termo de interação.

Neste caso, pode-se fazer uma mudança de variável para melhor observar a estrutura linear: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}z_1 + \widehat{\beta_2}z_2 + \widehat{\beta_5}z_3$. Entretanto, $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}{x_1}^{\beta_0}$, por exemplo, não é linear, ou seja, não é possível reescrevê-la em uma estrutura linear. Neste caso, a regressão linear não seria a modelagem a ser adotada.

Assim, pode-se usar a modelagem linear para um ajuste não-linear aos dados desde que permaneça linear nos coeficientes. Não linearidades entre variáveis não constituem um problema.

Regressão linear múltipla

Quais variáveis são realmente importantes para o modelo?

Pode-se adotar uma estratégia de seleção de variáveis:

- Progressiva/*Forward*: a partir de um modelo básico, apenas com o intercepto, com a lista de regressores vazia, executam-se p modelos lineares simples e adiciona-se ao modelo básico a variável com menor EMQ; e, assim, mais variáveis são adicionadas ao último modelo obtido até que algum critério de parada seja atingido;

- Regressiva/*Backward*: a partir de um modelo múltiplo completo, exclui-se a variável com maior valor p do teste t; assim, variáveis são excluídas do último modelo produzido até que um critério de parada seja atingido;

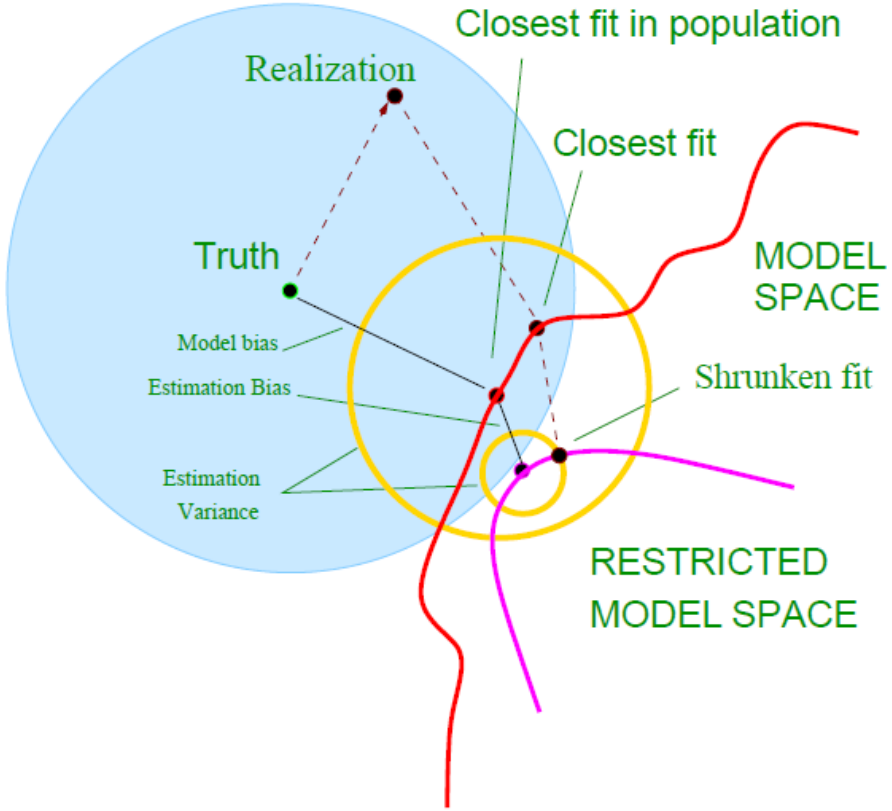
- Híbrida: começa como a progressiva e age como a regressiva caso alguma adição tenha causado aumento de valor p acima do tolerado.

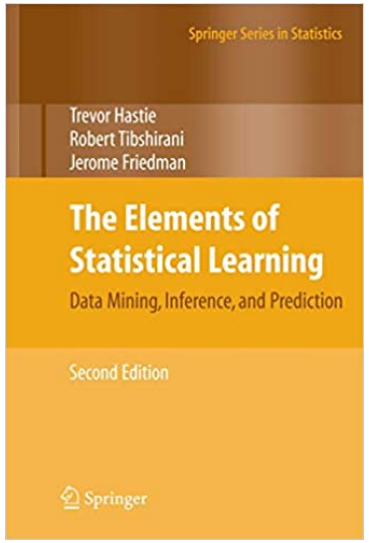
Modelos lineares

Regressão linear múltipla com regularização

Como estratégia alternativa à busca pelo dimensionamento adequado da modelagem, pode-se utilizar todas as variáveis regressoras para a geração de um modelo e adotar uma técnica que restringe ou regulariza os coeficientes do modelo.

Chamada de regularização, essa técnica torna menor, menos complexo, o espaço de busca de parâmetros de modelos e promove o encontro de modelos com melhores capacidades de generalização e com variâncias bem menores do que os que seriam obtidos sem ela.





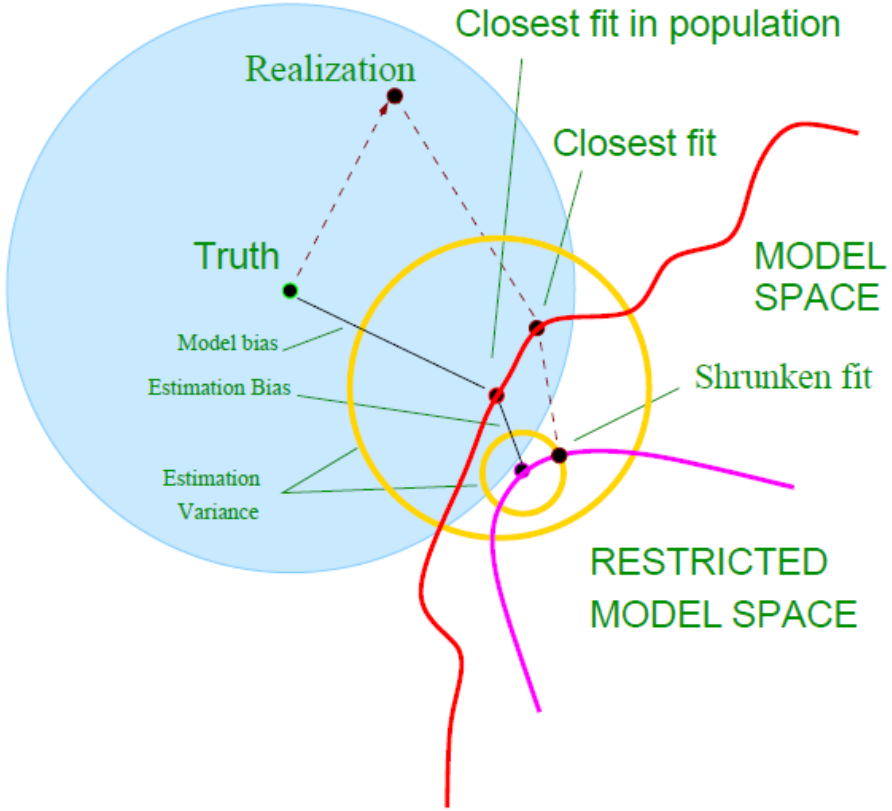
Modelos lineares

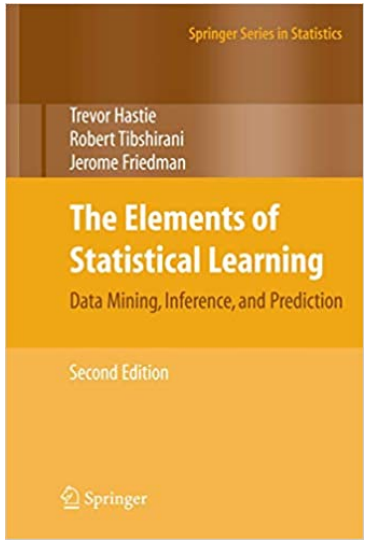
Regressão linear múltipla com regularização

Como não se conhece a real função geradora dos dados, o modelo produzido possui um erro de predição e minimizá-lo constitui um problema mal posto ou mal condicionado. Assim, resumidamente, a regularização procura tornar o problema da minimização do erro de predição em um problema menos mal posto.

Para ser bem posto ou bem condicionado, é preciso atender aos seguintes critérios:

- Existir uma função geradora e o modelo ser capaz de aproximá-la;
- A solução da minimização ser única;
- A solução deve ser estável, isto é, deve haver continuidade de comportamento diante de diferentes condições iniciais e de contorno.

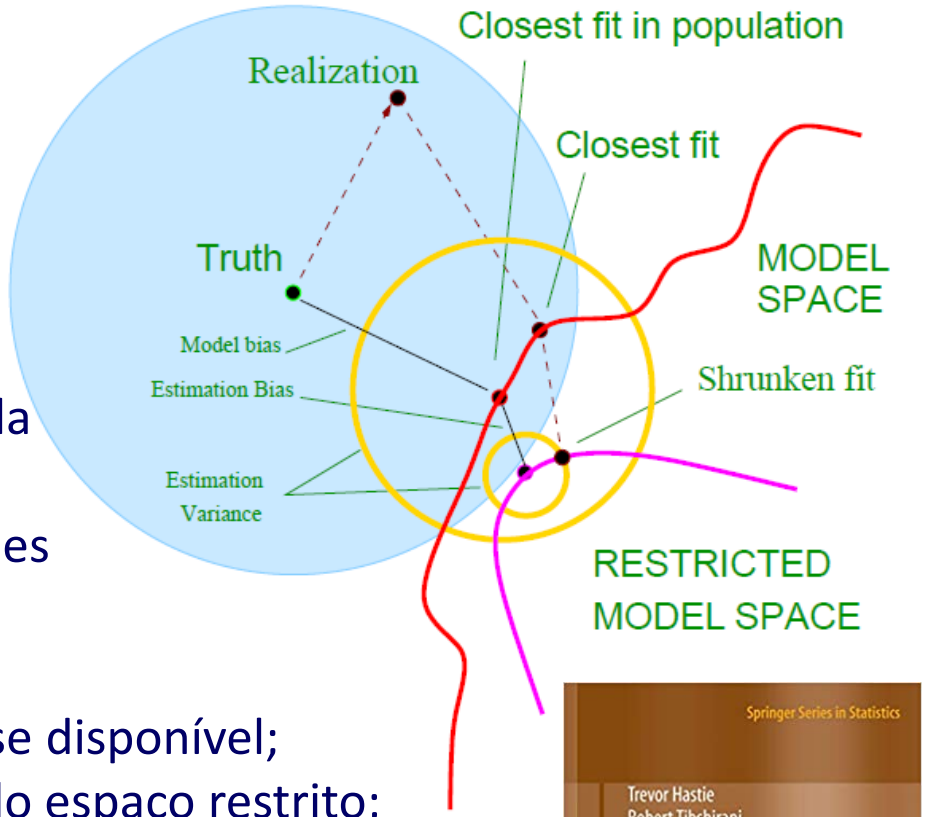




Modelos lineares

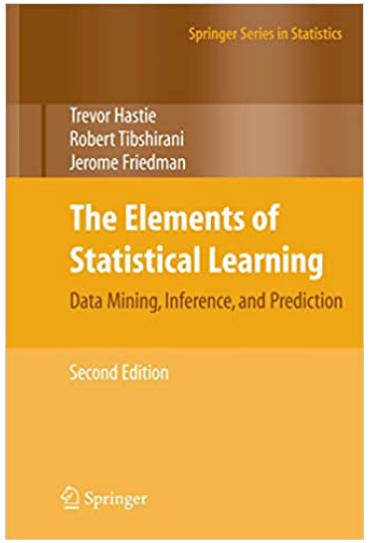
Regularização

Na figura, pode-se fazer a seguinte interpretação:



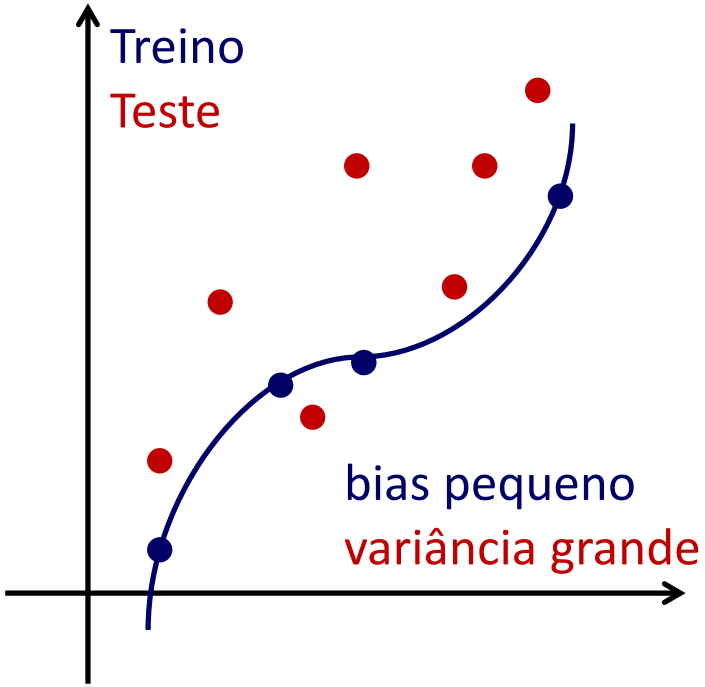
- *Truth* é a verdade dos dados, nem sempre observada;
- *Realization* é verdade observada, possivelmente com erros na coleta;
- *Model Space* é o conjunto de todas as modelagens possíveis a partir da verdade observada;
- *Restricted Model Space* é o subconjunto das modelagens com restrições (menos variáveis, domínios reduzidos, uso de regularização etc.);
- *Closest fit* é o melhor modelo ajustado aos dados observados;
- *Closest fit in population* é o modelo ideal se toda a população estivesse disponível;
- *Shrunken fit* é o modelo final ajustado aos dados observados dentro do espaço restrito;
- *Estimation Variance* refere-se à variabilidade de modelos com outros dados observados (quanto maior a complexidade, maior a varibilidade);
- *Model Bias* é o erro entre a predição do melhor modelo teórico e a verdade real;
- *Estimation Bias* é o erro adicional que surge por restringir o espaço de modelos.

Modelagens complexas podem se ajustar demais aos dados observados, tendo alta variabilidade.

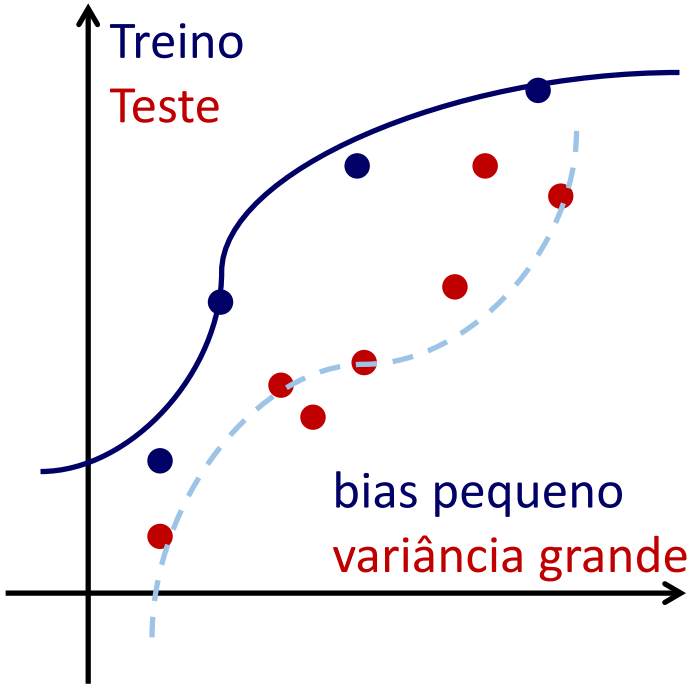


Regularização

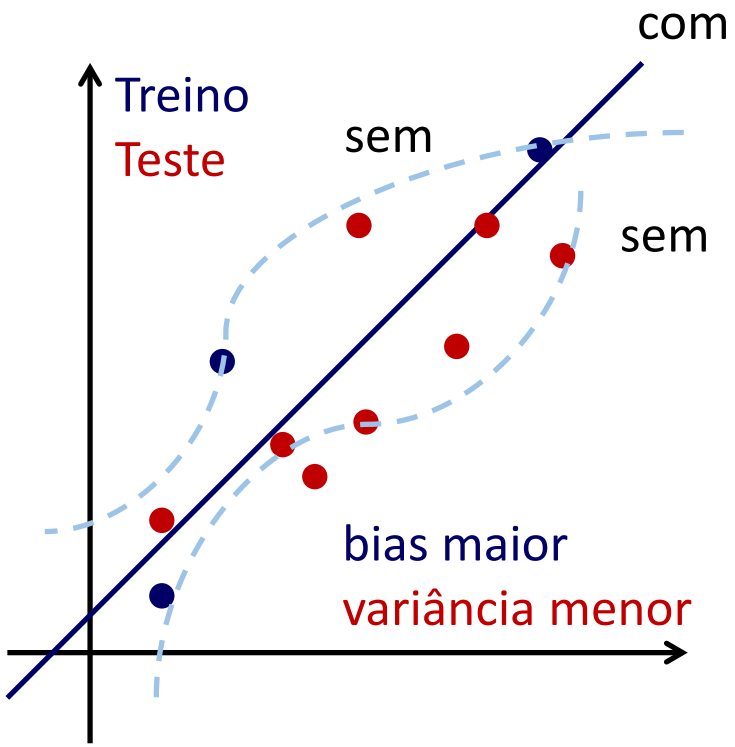
Interpretação geométrica de um superajuste



sem regularização



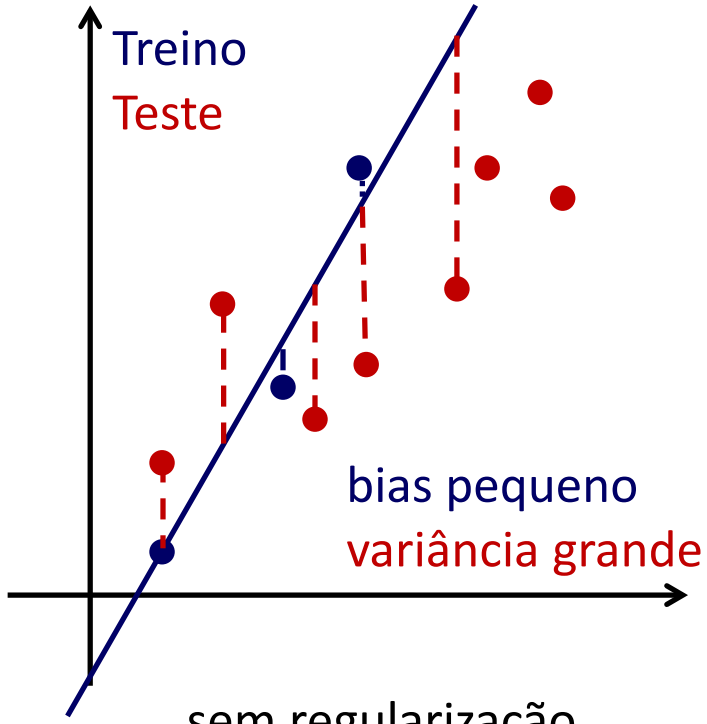
sem regularização



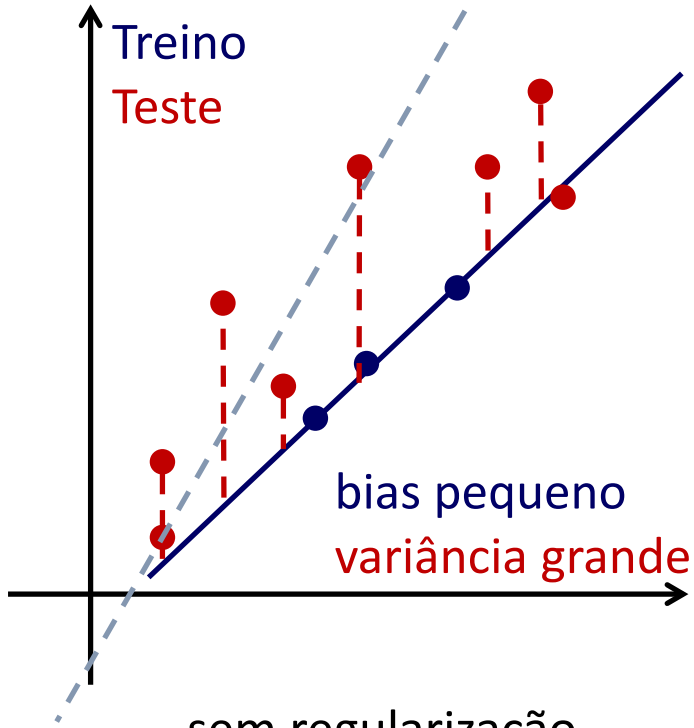
com regularização

Regressão linear múltipla com regularização

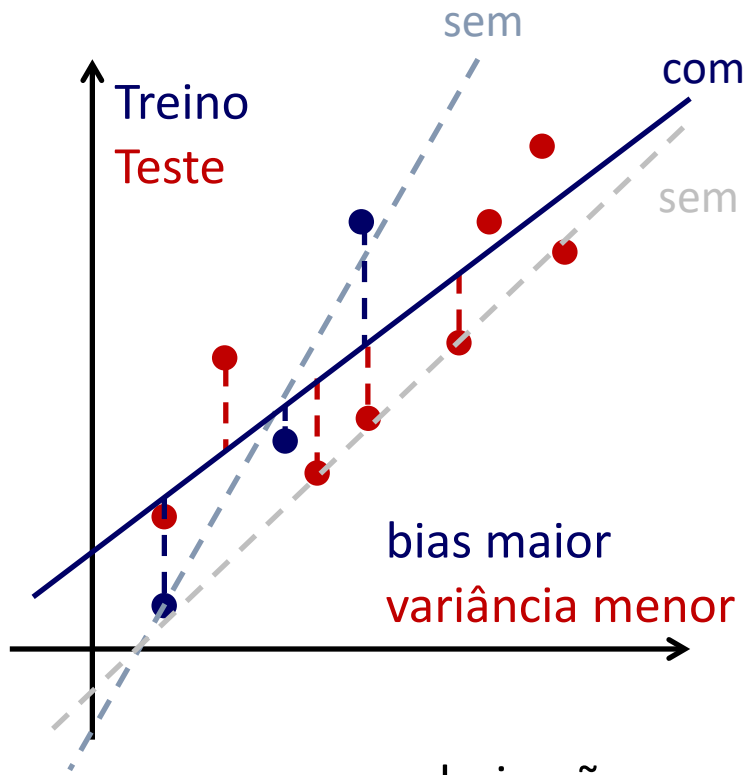
Interpretação geométrica



sem regularização



sem regularização

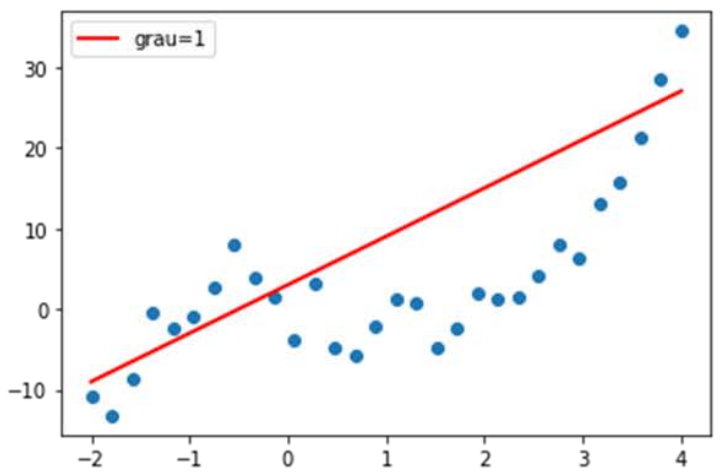


com regularização

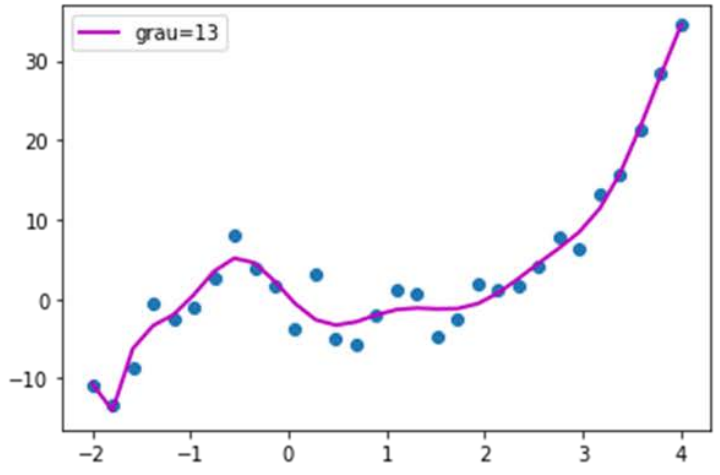
Modelos lineares

Regressão linear múltipla com regularização

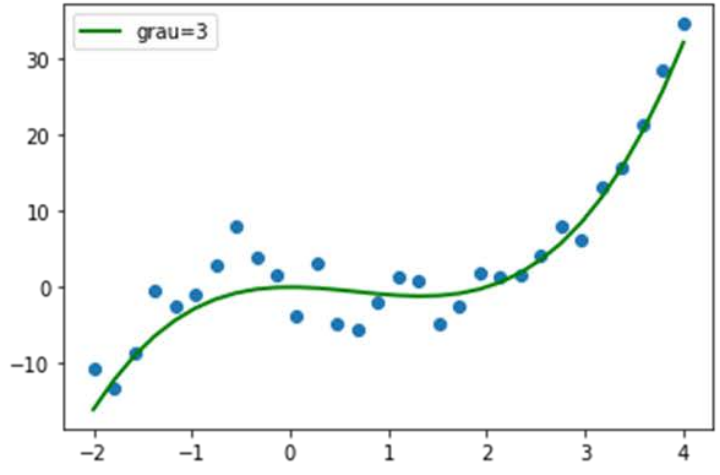
Uma observação empírica: um dimensionamento inadequado tende a provocar superajuste ou subajuste.



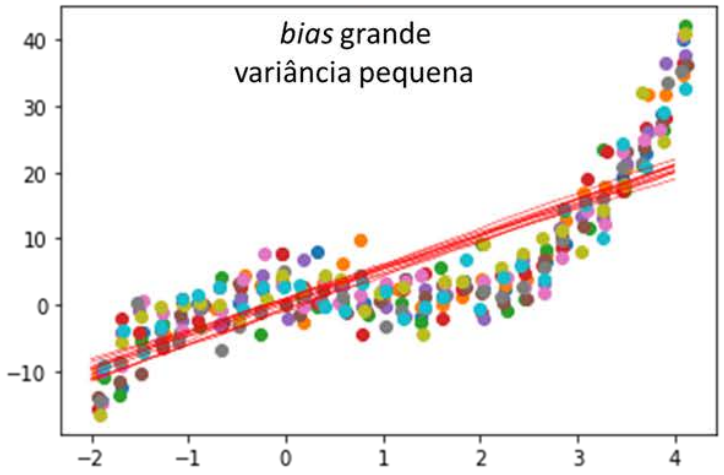
Subajustado

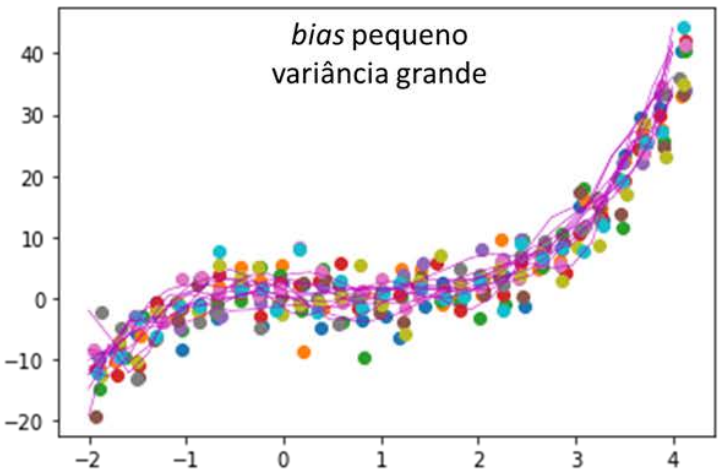


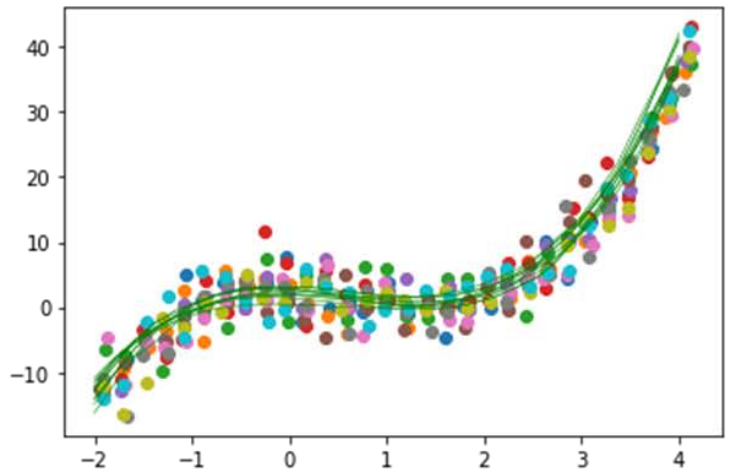
Superajustado



Bem ajustado



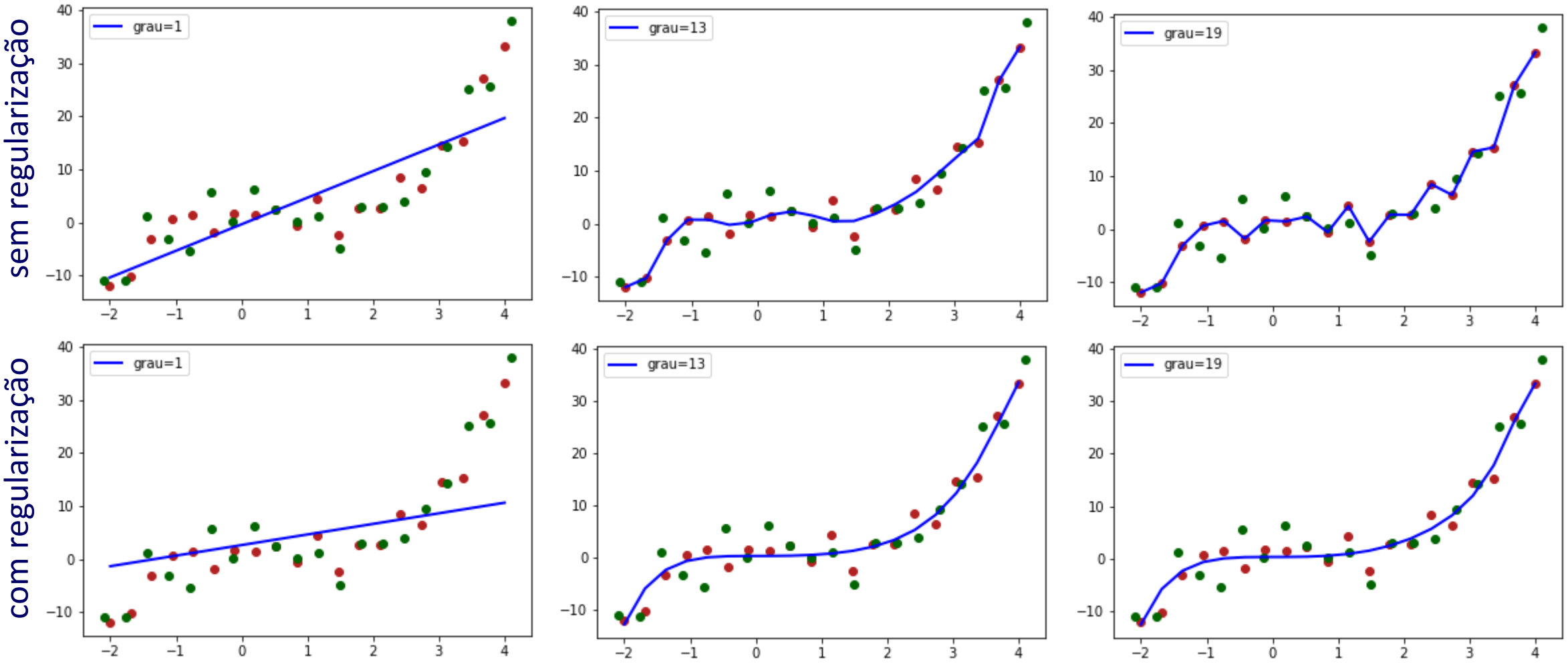




Modelos lineares

Regressão linear múltipla com regularização

Uma observação empírica: com regularização, um superdimensionamento pode ser corrigido.



Regressão linear múltipla com regularização

Mais observações empíricas:

- O superajuste está frequentemente associado a valores mais altos de coeficientes e é difícil controlá-los durante o aprendizado sem intervenção;
- Em muitos problemas, a quantidade de dados disponíveis para o aprendizado é insuficiente para a adequada geração de modelos.

Assim, as técnicas de regularização mais comuns procuram evitar o superajuste mantendo os coeficientes com valores pequenos por meio de uma penalização proporcional à soma dos valores absolutos ou quadrados de todos os coeficientes de um modelo de regressão. Muitas vezes, ambas as formas de penalização são adotadas.

$$EMQ_R = E\left((y_r - \hat{y})^2\right) + \text{termo de regularização}$$

esperança ou valor esperado

saída real

saída do modelo

Modelos lineares

Regressão linear múltipla com regularização

Ao se reduzir os domínios dos coeficientes de um modelo, alguns eventualmente poderão assumir valores nulos, o que reduz a dimensão do modelo, e promove-se uma menor variância de coeficientes em diferentes amostragens, o que resulta em menor variância das respostas de diferentes modelos.

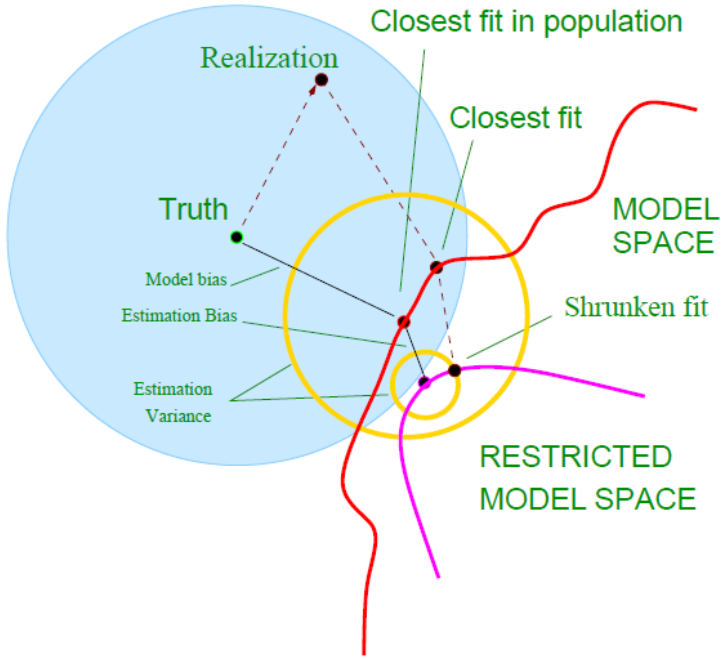
$$EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot \boldsymbol{\Sigma}_R$$

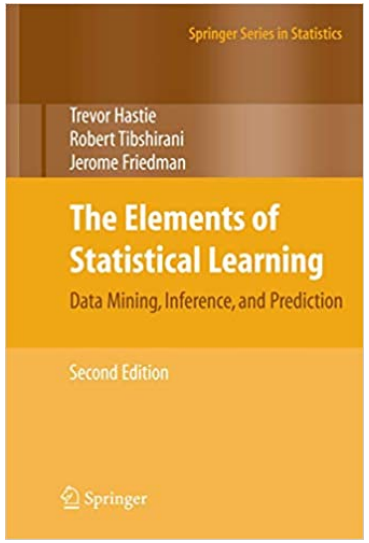
$\lambda \cdot \boldsymbol{\Sigma}_R$ → termo de regularização

$\lambda \geq 0$ é o parâmetro de regularização
$\boldsymbol{\Sigma}_R$ é o regularizador
$\lambda \in [0, +\infty)$ é arbitrado e pode ser escolhido por validação cruzada.

O efeito do termo de regularização é o de redução de possibilidades de representação das funções geradoras dos dados, tornando-as mais “similares”.

Assim, o aprendizado se torna mais estável diante de diferentes dados de treinamento apresentados.





Regressão linear múltipla com regularização

A regularização abrange todas as técnicas que adotam termos de penalização à função de erro/custo/perda de predição para evitar o superajuste.

Ao escolher o parâmetro de regularização de um modelo, é preciso encontrar um bom equilíbrio entre simplicidade e ajuste:

- se o valor for alto demais, o modelo tenderá a ser muito simples e correrá o risco de subajuste;
- se o valor for pequeno demais, o modelo tenderá a ser mais complexo do que o necessário, ocorrendo o risco de superajuste.

L2/Tikhonov: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot (\sum parâmetros^2)$ (*ridge, de cume, de crista*)

L1/LASSO: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot (\sum |parâmetros|)$ (*Least Absolute Shrinkage and Selection Operator*)

Elastic net: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda_1 \cdot (\sum parâmetros^2) + \lambda_2 \cdot (\sum |parâmetros|)$ (rede elástica)

Modelos lineares

Regressão linear múltipla com regularização

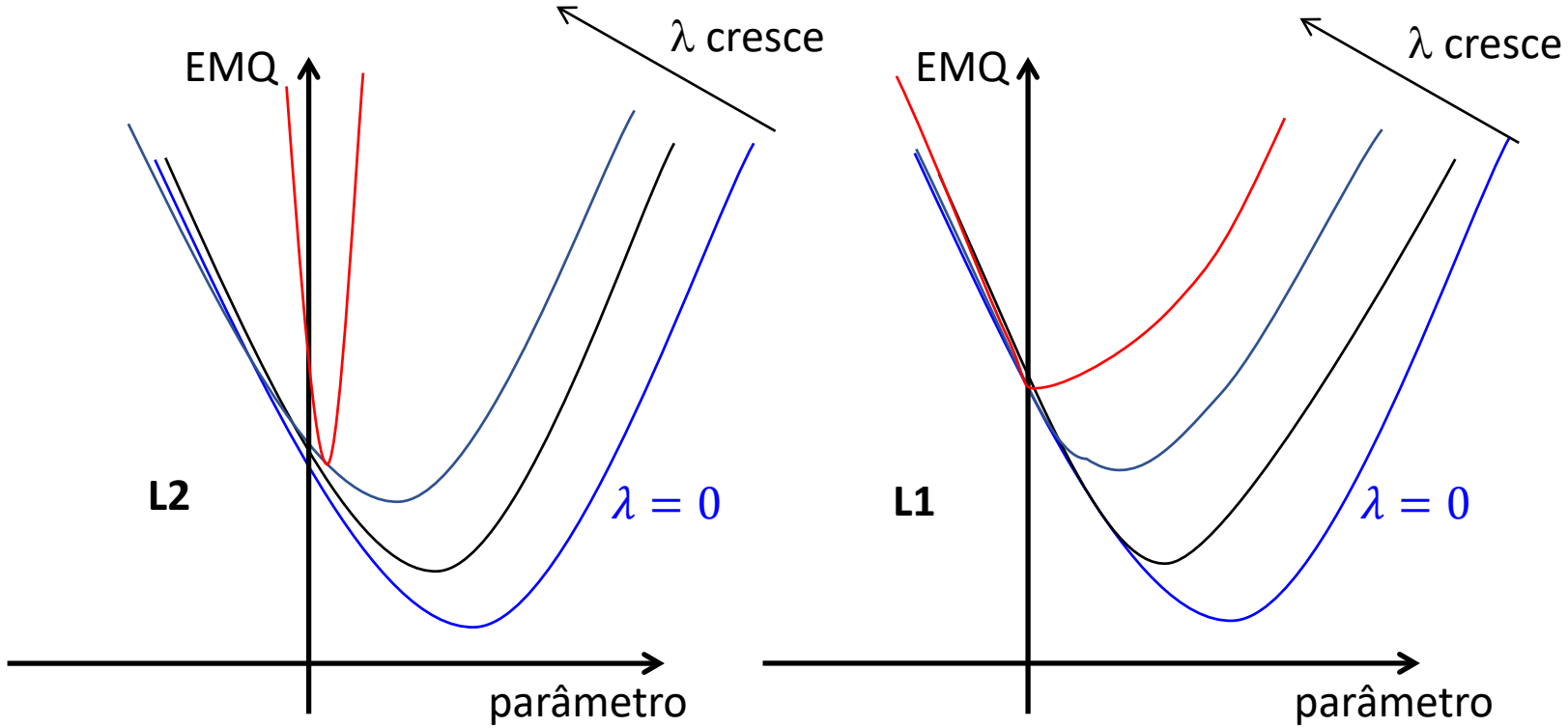
L2/Tikhonov: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot (\sum parâmetros^2)$ (*encolhimento, redução*)

L1/LASSO: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot (\sum |parâmetros|)$ (*encolhimento e seleção de atributos*)

Elastic net: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda_1 \cdot (\sum parâmetros^2) + \lambda_2 \cdot (\sum |parâmetros|)$ (ambos)

Na *ridge*, os parâmetros podem se aproximar de zero e na LASSO eles podem alcançá-lo, significando a remoção das variáveis a que os coeficientes nulos estão associados.

Assim, se há certeza que todas as variáveis são relevantes, *ridge* é a escolha a ser feita, por não realizar qualquer seleção.



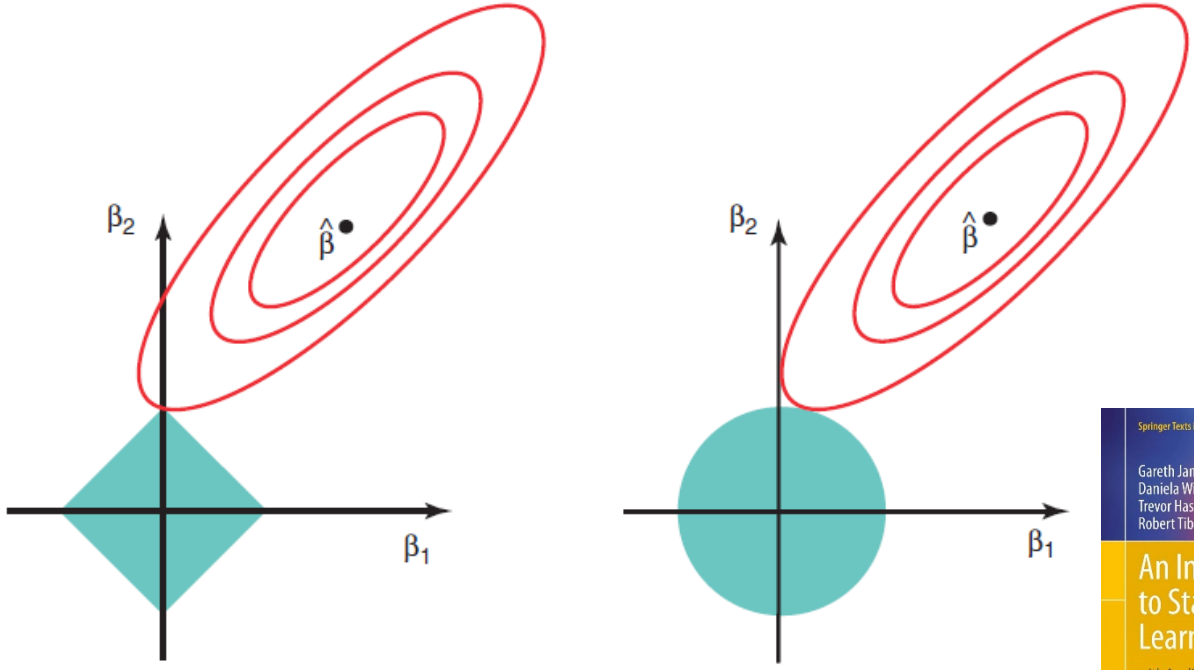
Regressão linear múltipla com regularização

Interpretação geométrica

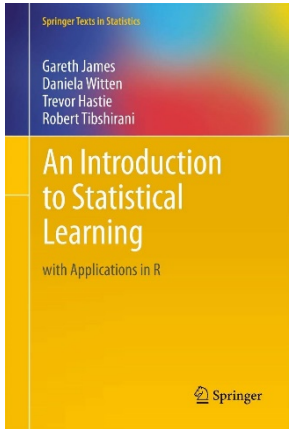
É possível mostrar que os coeficientes estimados pelas regularizações *ridge* e LASSO resolvem os seguintes problemas de otimização, respectivamente:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \cdot x_{ij} \right)^2 \text{ sujeito a } \sum_{j=1}^{p} \beta_j^2 \leq s$$

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \cdot x_{ij} \right)^2 \text{ sujeito a } \sum_{j=1}^{p} |\beta_j| \leq s$$



As estimativas dos coeficientes são dadas pelo primeiro ponto em que uma curva de contorno do erro quadrático (elipse) toca a região de restrições (losango para o LASSO e círculo para o *ridge*).



Regularização

Empiricamente, observa-se que a regularização LASSO obtém sucesso em muitos problemas e seleciona no máximo $\min(n, p)$ variáveis, n é a quantidade de registros e p a quantidade de variáveis preditoras. Isso é especialmente desvantajoso em problemas em que $p \gg n$, quando muitas variáveis relevantes podem ser descartadas, levando à instabilidade no desempenho.

Por outro lado, quando $n > p$ em problemas com variáveis correlacionadas, frequentemente a *ridge* obtém melhores resultados que a LASSO, distribuindo os valores de coeficientes de forma mais homogênea, promovendo uma generalização mais suave e estável.

A *Elastic net* procura superar as dificuldades da LASSO combinando sua atuação com a da *ridge*:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{f(\lambda_1, \lambda_2, \boldsymbol{\beta})\}$$

L2/Tikhonov: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot (\sum parâmetros^2)$ (*ridge, de cume, de crista*)

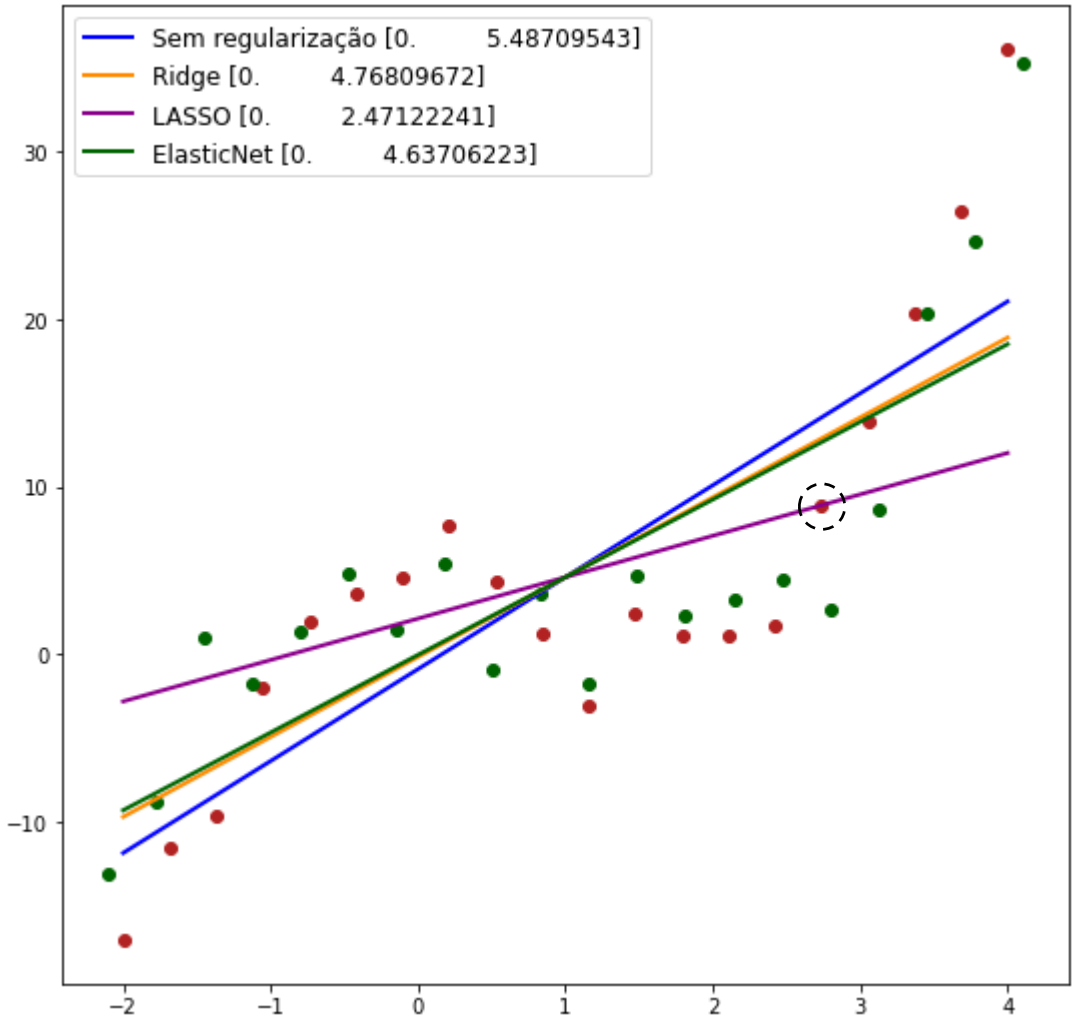
L1/LASSO: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda \cdot (\sum |parâmetros|)$ (*Least Absolute Shrinkage and Selection Operator*)

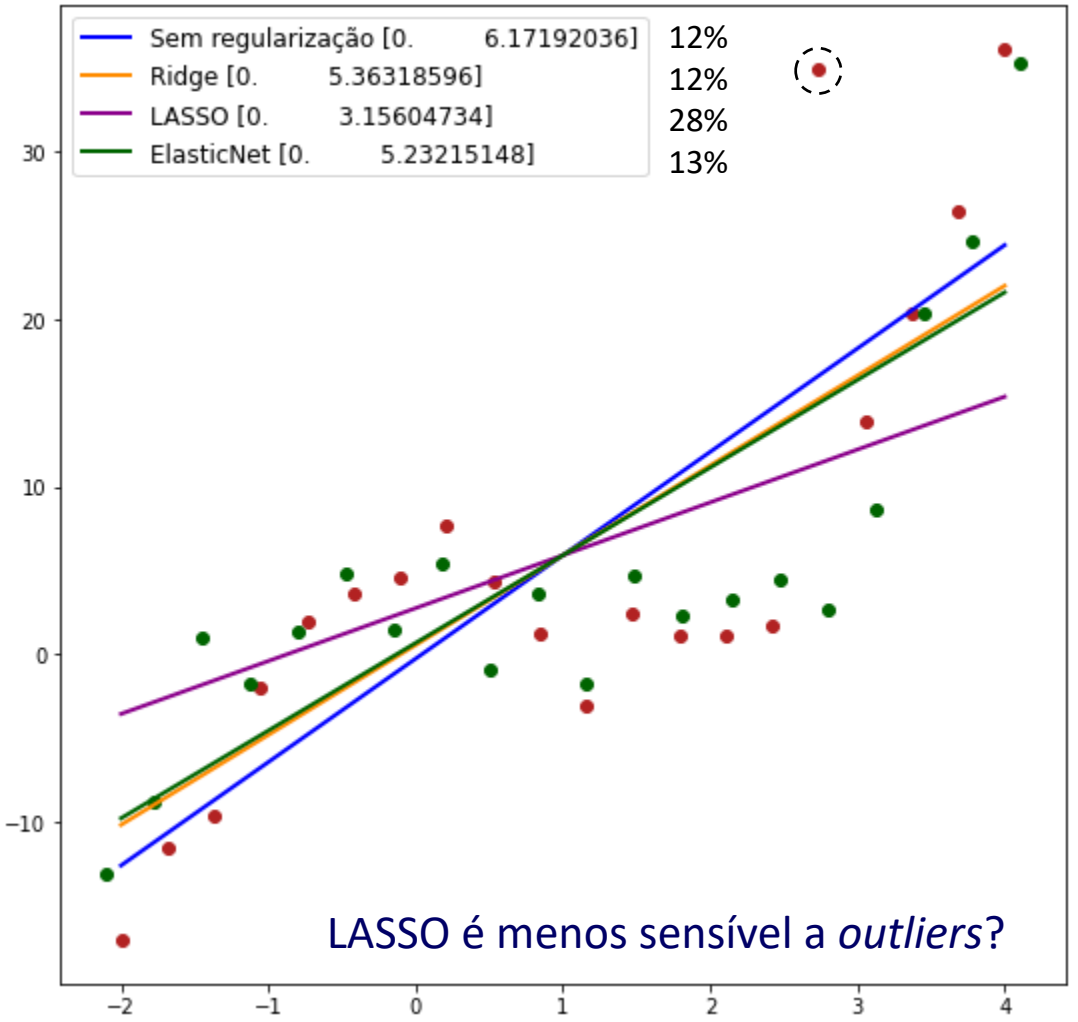
Elastic net: $EMQ_R = E\left((y_r - \hat{y})^2\right) + \lambda_1 \cdot (\sum parâmetros^2) + \lambda_2 \cdot (\sum |parâmetros|)$ (rede elástica)

Modelos lineares

Regularização

Interpretação geométrica





Regressão linear

Como lidar com variáveis discretas?

Se os possíveis valores são ordenados, basta mapeá-los no conjunto dos naturais.
Se as variáveis não são ordenadas e dicotômicas, pode-se:

$$x_i' = \begin{cases} 1, & se \;\; x_i = k \\ 0, & se \;\; x_i \neq k \end{cases},$$

sendo k um dos possíveis valores da variável original

Considerando a regressão simples:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_i' = \begin{cases} \widehat{\beta_0} + \widehat{\beta_1}, & se \;\; x_i' = 1 \\ \widehat{\beta_0}, & se \;\; x_i' = 0 \end{cases}$$

$\widehat{\beta_0}$ é a média de $\hat{y}$ quando $x_i \neq k$;
$\widehat{\beta_0} + \widehat{\beta_1}$ é a média de $\hat{y}$ quando $x_i = k$
$\widehat{\beta_1}$ é a média das diferenças entre os dois tipos

$$x_i' = \begin{cases} 1, & se \;\; x_i = k \\ -1, & se \;\; x_i \neq k \end{cases},$$

sendo k um dos possíveis valores da variável original

Considerando a regressão simples:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_i' = \begin{cases} \widehat{\beta_0} + \widehat{\beta_1}, & se \;\; x_i' = 1 \\ \widehat{\beta_0} - \widehat{\beta_1}, & se \;\; x_i' = -1 \end{cases}$$

$\widehat{\beta_0} - \widehat{\beta_1}$ é a média de $\hat{y}$ quando $x_i \neq k$;
$\widehat{\beta_0} + \widehat{\beta_1}$ é a média de $\hat{y}$ quando $x_i = k$; e
$\widehat{\beta_0}$ é a média das diferenças entre os dois tipos

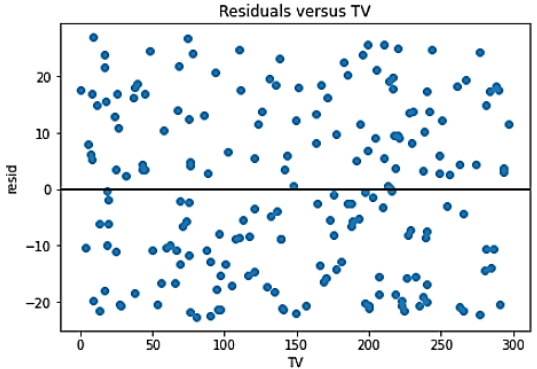
Regressão linear

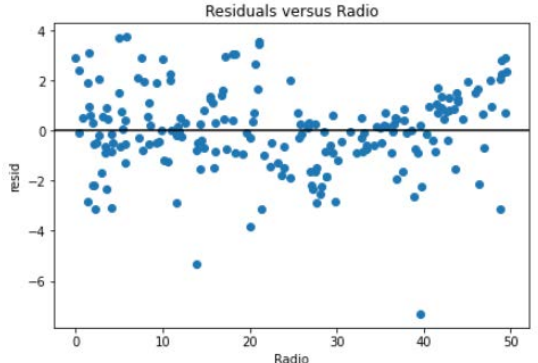
Possíveis problemas na modelagem linear:

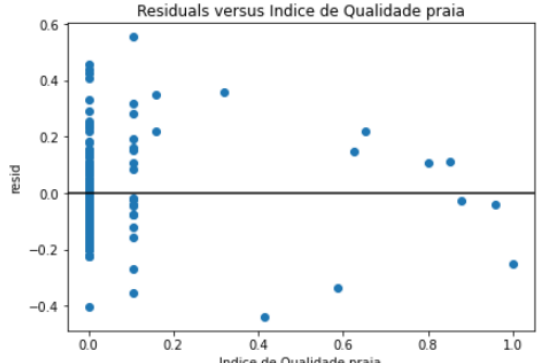
- não linearidade dos dados, isto é, entre variáveis preditoras e variável de resposta, ou heterocedasticidade dos erros

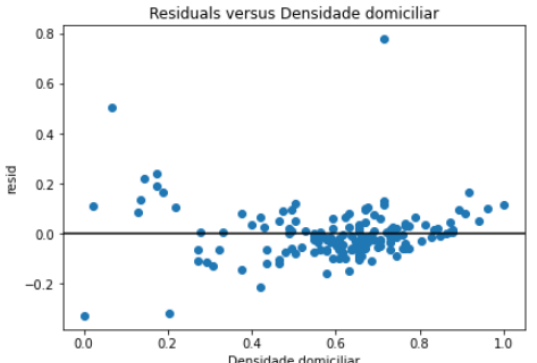
  Um modelo linear dificilmente capturará a forma real da relação entre variáveis preditoras e predita. Como consequências prováveis, os resíduos terão um padrão de comportamento, o que viola a suposição de homocedasticidade, e o ajuste será de má qualidade. O uso do gráfico de dispersão de resíduos (saídas preditas x resíduos) ajuda a identificar o problema: se houver alguma tendência de comportamento dos resíduos, há indícios de heterocedasticidade.

  Possível solução: transformar os preditores com funções não lineares tais como $\exp x$, $\log x$, $\sqrt{x}$ e x^2, usar pesos nos dados de modo a reduzir a influência de quem mais varia, dentre outros.









Modelos lineares

Regressão linear

Possíveis problemas na modelagem linear:

- presença de *outliers*

  Os *outliers* afetam as medidas de dispersão e de acurácia do modelo.

  Possível solução: excluí-los, pois provavelmente nada significativo em relação ao ajuste será observado.

- colinearidade de preditores

  Além de prejudicar a identificação do efeito de cada preditora envolvida, traz incertezas às estimativas dos coeficientes e pode causar a exclusão de alguma variável indevidamente.

  Possível solução: excluir variáveis a partir da matriz de correlações.

- multicolinearidade de preditores

  É possível que variáveis não correlacionadas apresentem correlações em grupo.

  Possível solução: calcular o fator de inflação de variância (VIF) para identificar variáveis candidatas à exclusão; centralizar variáveis; usar PCA para combinar as variáveis envolvidas; usar regularização.

Regressão linear

Possíveis problemas na modelagem linear:

- multicolinearidade de preditores

Se uma variável x_j está fortemente correlacionada com outras variáveis do modelo, haverá dificuldade em determinar o efeito dela na predição. Como consequência, a variância estimada do parâmetro $\widehat{\beta_j}$ aumenta (infla). Assim, o VIF mede o quanto a variância estimada de um parâmetro $\widehat{\beta_j}$ aumenta devido à colinearidade com os demais preditores do modelo.

$$VIF(\widehat{\beta_j}) = \frac{1}{1 - R^2_{j,\sim j}}$$

em que $R^2_{j,\sim j}$ é o coeficiente de determinação da regressão de x_j a partir de todas as outras variáveis regressoras do problema.

$VIF = 1$: sem multicolinearidade
$1 \leq VIF \leq 5$: multicolinearidade moderada
$5 < VIF \leq 10$: multicolinearidade alta
$VIF > 10$: multicolinearidade altíssima

Regressão linear

Possíveis problemas na modelagem linear:

- correlação entre erros

  A suposição da modelagem linear de que os erros (ou resíduos) são independentes entre si é violada quando se tem a autocorrelação dos erros: o erro de uma observação é influenciado pelo erro de outra.

  Se os resíduos dos registros são correlacionados, então estão comprometidas as decisões em que foram usados os erros padrão dos parâmetros, que tenderão a subestimar os erros reais e a superestimar a confiança no modelo.

  Analisar autocorrelação apenas faz sentido em dados sequenciais, sejam eles temporais, espaciais ou outra sequência lógica.

  O uso do gráfico de dispersão de resíduos ajuda a identificar o problema. Pode-se aplicar o teste de Durbin-Watson, para autocorrelações sequenciais diretas, ou o teste de Breusch-Godfrey, mais geral e capaz de detectar autocorrelação com defasagens mais altas.

  Possível solução: não utilizar o método de mínimos quadrados tradicional para encontrar os parâmetros e escolher outro capaz de lidar com dados sequenciais, tais como as modelagens ARIMA.

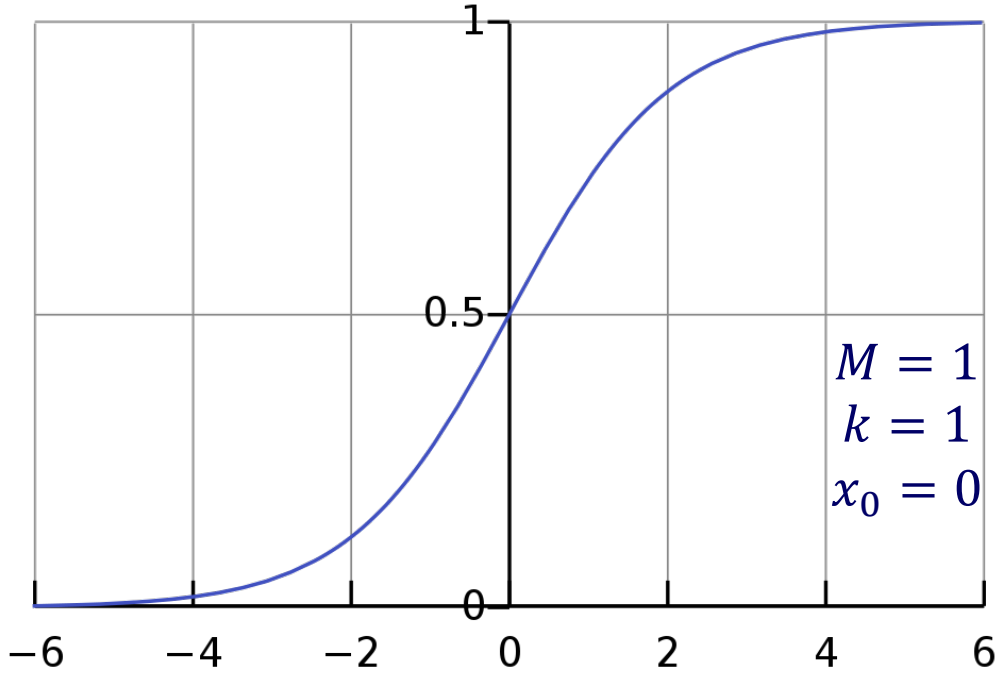
Regressão logística

Modelagem estatística que utiliza uma função logística para estimar a probabilidade de pertinência de pontos em uma classe.

$$f(x) = \frac{M}{1 + e^{-k(x-x_0)}}$$

M: valor máximo
k: inclinação
x_0: centro da sigmoide

$$p(y = 1|x) = f(x)$$



$$f(x) = \frac{1}{1 + e^{-x}}$$

O valor máximo da sigmoide é 1 (código da classe alvo). Os demais parâmetros da função são estimados por máxima verossimilhança a partir dos dados de treinamento.

Modelos lineares

Regressão logística

$$f(x) = \frac{1}{1 + e^{-k(x-x_0)}} = \frac{e^{k(x-x_0)}}{e^{k(x-x_0)} + 1} \quad \Rightarrow \quad \left(e^{k(x-x_0)} + 1\right) \cdot f(x) = e^{k(x-x_0)}$$

$$\Rightarrow \quad f(x) = e^{k(x-x_0)} - e^{k(x-x_0)} \cdot f(x) \quad \Rightarrow \quad f(x) = e^{k(x-x_0)} \cdot \left(1 - f(x)\right)$$

$$\Rightarrow \quad \frac{f(x)}{1 - f(x)} = e^{k(x-x_0)} \quad \Rightarrow \quad \ln \frac{f(x)}{1 - f(x)} = k(x - x_0)$$

log *odds*
logit

chance/*odds* $\in [0, +\infty[$

função linear

$$\ln \frac{f(x)}{1 - f(x)} = k(x - x_0) = \beta_0 + \beta_1 x$$

Assim, estimar os parâmetros β_0 e β_1 da função linear permitirá encontrar a melhor função logística de acordo com os dados. Os melhores parâmetros são os que maximizam a função de verossimilhança dos dados de treinamento.

Modelos lineares

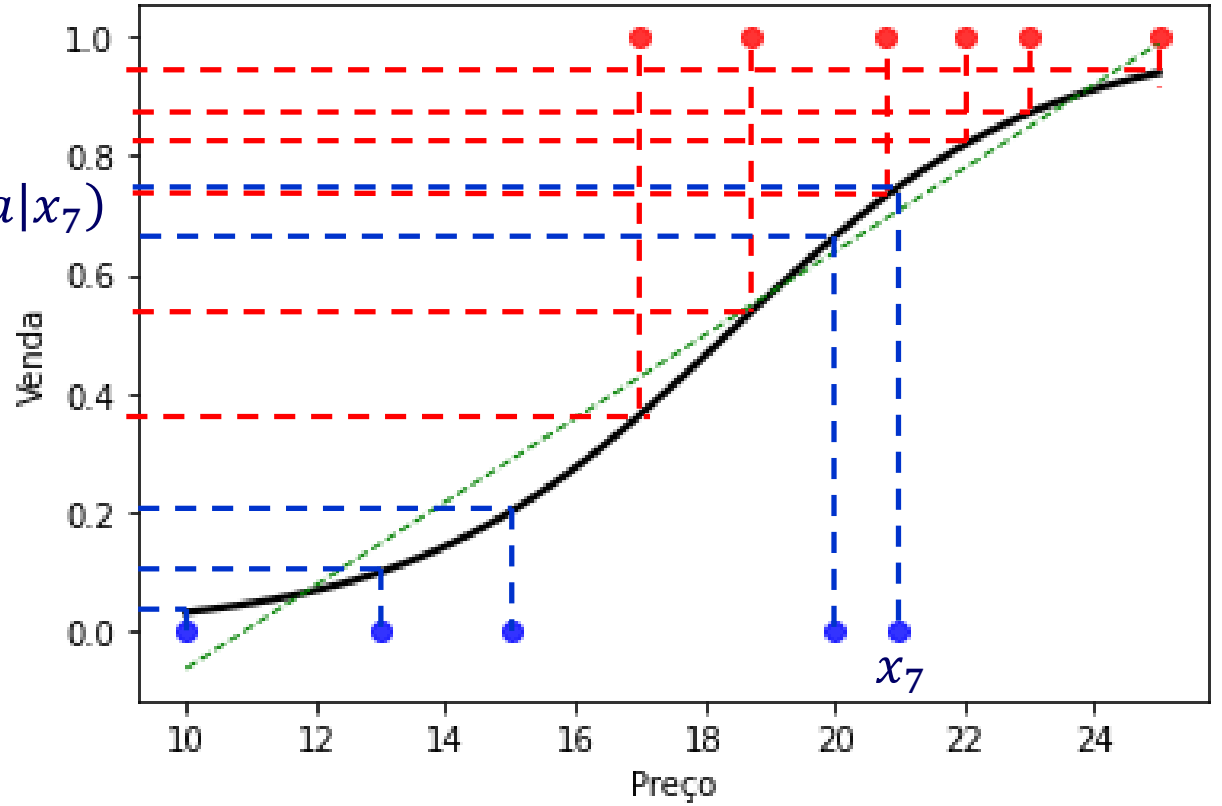
Regressão logística

Os melhores parâmetros são os que maximizam a verossimilhança dos dados de treinamento:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{x,y=1} p(x) \cdot \prod_{x,y=0} \big(1 - p(x)\big)$$

x é registro da classe alvo

x não é registro da classe alvo



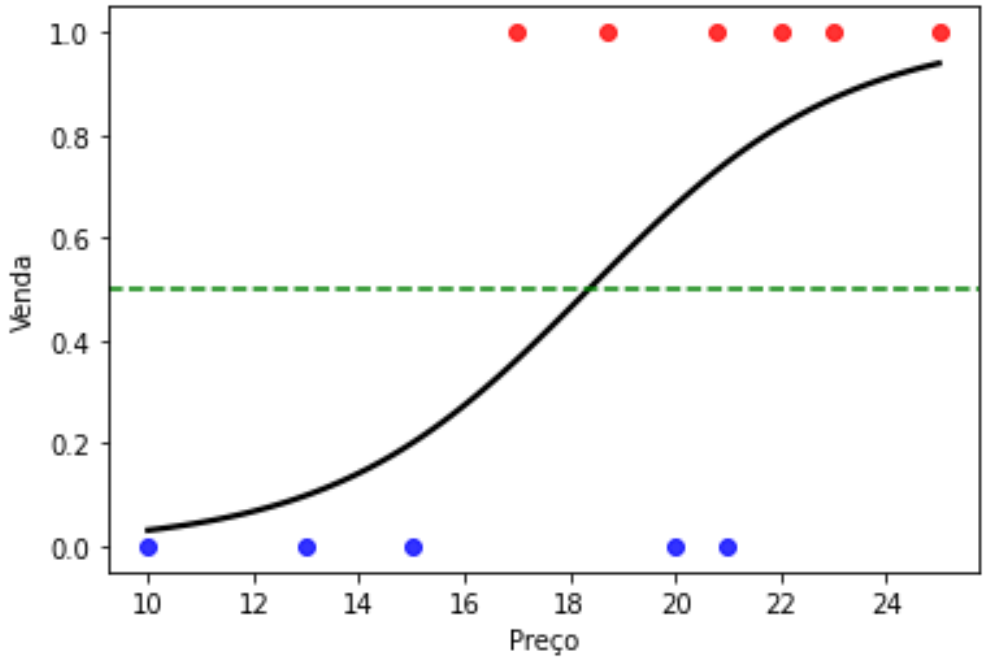
Como os produtórios acima resultam em valores muito pequenos, na prática usa-se o logaritmo da função de verossimilhança:

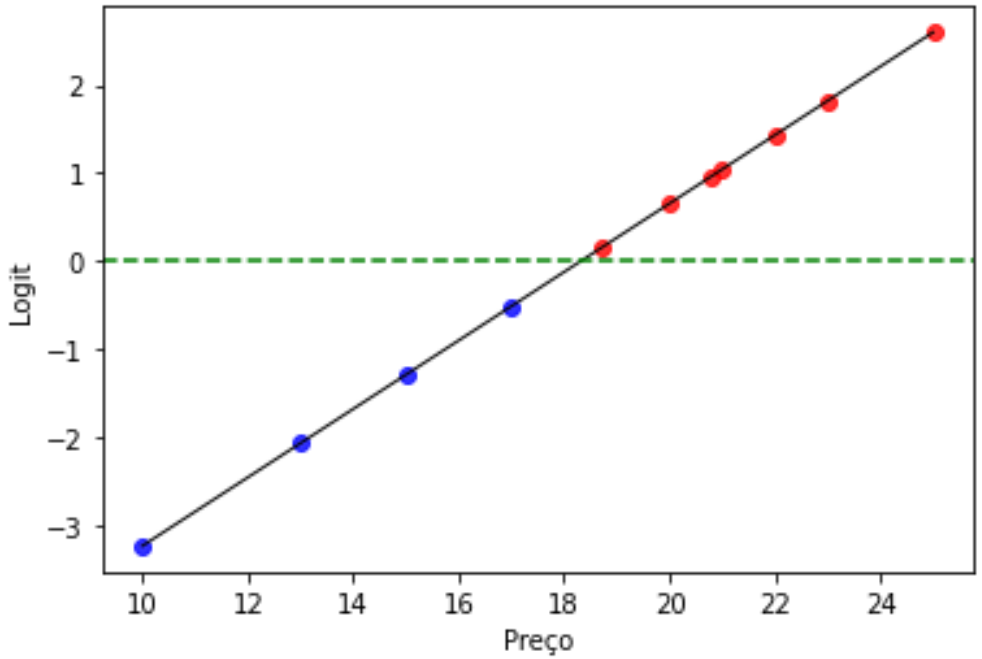
$$\ln \mathcal{L}(\beta_0, \beta_1) = \sum_{x,y=1} \ln p(x) + \sum_{x,y=0} \ln\big(1 - p(x)\big)$$

$$Prob.Vendas(x) = \frac{1}{1 + e^{7{,}132 - 0{,}389x}}$$

Regressão logística

Durante o aprendizado, o problema é levado do espaço original, em que o domínio da variável resposta (probabilidade de pertinência à classe alvo) é $[0,1]$, para outro espaço em que a variável resposta (logit) tem domínio $[-\infty, +\infty]$.





Assim, os parâmetros da função logística do espaço original serão dados pelo hiperplano produzido no espaço de busca do problema de otimização $\max \ln \mathcal{L}(\beta_0, \beta_1)$.

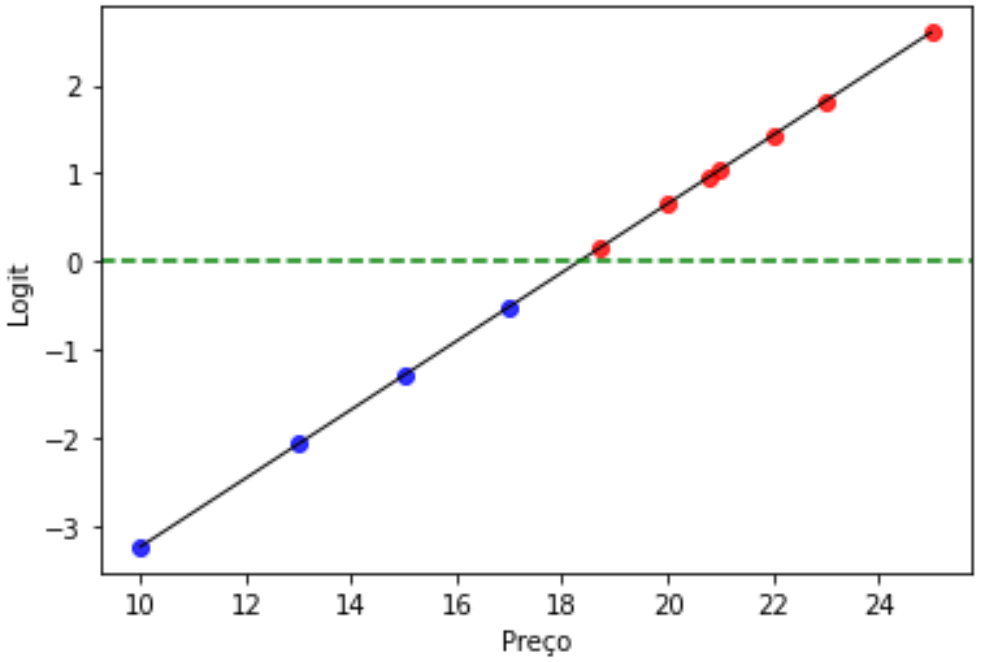
Modelos lineares

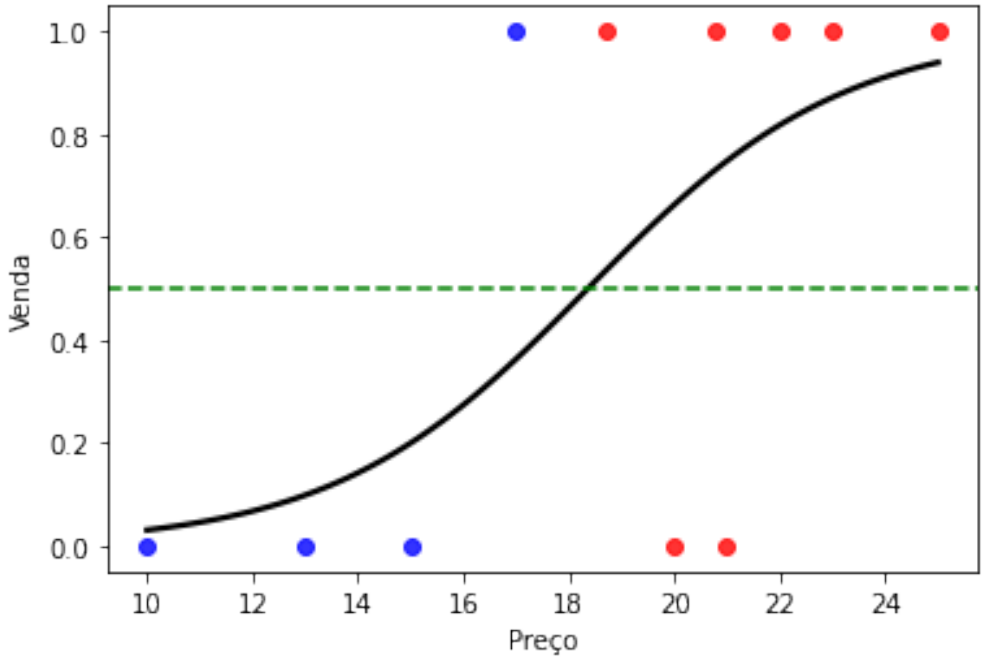
Regressão logística

Uso do modelo obtido:

$\beta_0 + \beta_1 x \geq 0$: é da classe alvo
$\beta_0 + \beta_1 x < 0$: não é da classe alvo

Exemplo:





$-7{,}132 + 0{,}389x \geq 0$: é da classe alvo
$-7{,}132 + 0{,}389x < 0$: não é da classe alvo

$$Prob.Vendas(x) = \frac{1}{1 + e^{7{,}132 - 0{,}389x}}$$

Regressão logística

Passos:

1. dados $\widehat{\beta_0}, \dots, \widehat{\beta_n}$ candidatos, obtém-se um modelo linear múltiplo $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_n} x_n$;
2. transformam-se as respostas deste modelo para os dados de treino em suas probabilidades/ verossimilhanças: $\mathrm{p}(x_{i1}, \dots, x_{in}) = e^{\widehat{y_i}} / \left(e^{\widehat{y_i}} + 1\right)$
3. calcula-se o logaritmo da verossimilhança $\mathcal{L}\left(\widehat{\beta_0}, \dots, \widehat{\beta_n}\right)$ da curva;
4. ajustam-se os parâmetros $\widehat{\beta_0}, \dots, \widehat{\beta_n}$ segundo alguma técnica de otimização numérica a partir das (estimativas das) derivadas parciais de $\ln \mathcal{L}\left(\widehat{\beta_0}, \dots, \widehat{\beta_n}\right)$ ou de outras técnicas;
5. os passos acima são repetidos até que $\mathcal{L}\left(\widehat{\beta_0}, \dots, \widehat{\beta_n}\right)$ atinja seu máximo valor ou algum critério de parada.

Ao final, tem-se o modelo linear mais ajustado no espaço transformado que corresponderá à melhor função logística no espaço original. Como não é possível calcular o coeficiente de determinação sem resíduos, pode-se calcular um falso R^2 para ser interpretado como um indicador de desempenho relativo para a seleção de modelos, assim como AUC e tantos outros.

$$R^2{}_{McFadden} = 1 - \frac{\ln \mathcal{L}(\beta_0, \dots, \beta_n)}{\ln \mathcal{L}(\beta_0)}$$

sklearn.linear_model.LogisticRegression