Mínimos Quadrados - 2025

© Gustavo C. Buscaglia

gustavo.buscaglia@gmail.com

- Para esse tema recomendamos a leitura dos temas de mínimos quadrados (pp. 92-97) e sistemas sobredeterminados (pp. 141-143) do livro de texto (Quarteroni-Saleri).
- Seguimos tanto quanto possível a presentação de *Mathematics for Machine Learning*, capítulos 8 e 9,

  `https://mml-book.github.io/`
- O exemplo sobre preços de imóveis está tomado de

  `https://www.kaggle.com/datasets/prokshitha/home-value-insights`

---

1 Dados e modelos

- Assumimos que os dados originais estão sob a forma de uma tabela numérica.
- Cada linha da tabela representa uma **instância**, um **exemplo**, um **elemento amostral**.
- Cada coluna representa uma **variável**, das quais uma é a **variável resposta** y, ou **label**, ou **anotação** (suposta escalar) e as outras são as **covariáveis** originais.

#	Terr m^2	Constr m^2	Ano	suites	quartos	banh	Plantas	Piscina m	Vagas	Seg24h 1=sim	Preço kR$
1	202	140	1998	0	3	3	2	0	2	0	423
2	250	137	2011	1	2	4	1	0	3	1	611
3	156	102	2001	1	1	3	1	0	1	1	354
4	353	182	2004	3	0	4	1	12	3	0	712
5	198	145	1983	2	1	5	1	0	2	0	387
6	376	251	2007	3	1	5	2	14	3	1	971
7	242	165	2015	1	3	4	2	12	3	1	765
8	177	133	1976	0	3	3	1	0	1	1	313
9	298	223	1997	3	0	5	1	0	2	1	789
10	422	351	2004	3	2	7	2	18	3	1	1310
11	202	140	1998	0	3	3	2	0	2	0	423
12	250	137	2011	1	2	4	1	0	3	1	611
13	156	102	2001	1	1	3	1	0	1	1	354
14	353	182	2004	3	0	4	1	12	3	0	712
15	198	145	1983	2	1	5	1	0	2	0	387
16	376	251	2007	3	1	5	2	14	3	1	971
17	242	165	2015	1	3	4	2	12	3	1	765
18	177	133	1976	0	3	3	1	0	1	1	313
19	298	223	1997	3	0	5	1	0	2	1	789
20	422	351	2004	3	2	7	2	18	3	1	1310
21	202	140	1998	0	3	3	2	0	2	0	423
22	250	137	2011	1	2	4	1	0	3	1	611
23	156	102	2001	1	1	3	1	0	1	1	354
24	353	182	2004	3	0	4	1	12	3	0	712
25	198	145	1983	2	1	5	1	0	2	0	387
26	376	251	2007	3	1	5	2	14	3	1	971
27	242	165	2015	1	3	4	2	12	3	1	765
28	177	133	1976	0	3	3	1	0	1	1	313
29	298	223	1997	3	0	5	1	0	2	1	789
30	422	351	2004	3	2	7	2	18	3	1	1310
31	202	140	1998	0	3	3	2	0	2	0	423
32	250	137	2011	1	2	4	1	0	3	1	611
33	156	102	2001	1	1	3	1	0	1	1	354
34	353	182	2004	3	0	4	1	12	3	0	712
35	198	145	1983	2	1	5	1	0	2	0	387
36	376	251	2007	3	1	5	2	14	3	1	971

- O problema de mínimos quadrados lineares surge da necessidade de ajustar um **modelo matemático linear** a um conjunto de observações.
- Das variáveis disponíveis, seja

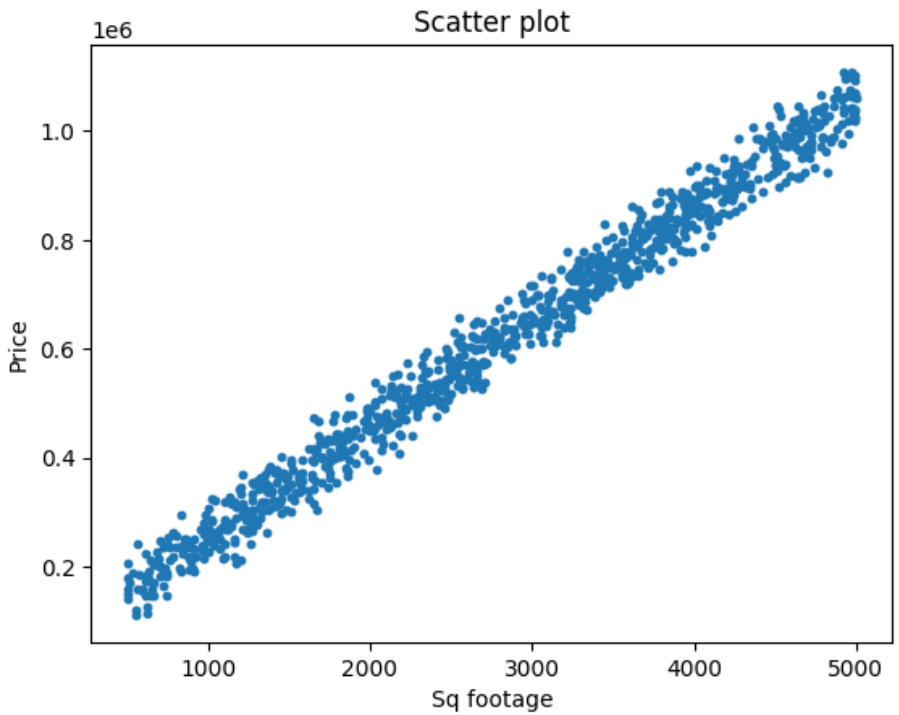
$$\mathbf{x} = (x_1, \dots, x_D)$$

  o vetor contendo aquelas selecionadas como **covariáveis** (ou **variáveis explicativas**, ou **atributos de interesse**, ou **features**).
- Cada observação (ou exemplo) $n = 0, \dots, N-1$ contém um vetor de covariáveis

$$\mathbf{x}_n = (X_{n1}, X_{n2}, \dots, X_{n,D})$$

  e um valor da variável resposta y_n.
- Um **conjunto de dados** é dado por N observações $\{(\mathbf{x}_0, y_0), (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N-1}, y_{N-1})\}$. Organizamos ele na forma de uma **matriz de exemplos X** (ou matriz de covariáveis, ou de features) e um **vetor resposta y**.

- Consideremos, por exemplo, que
  - a variável resposta y é o preço da casa,
  - a única variável explicativa selecionada x é a superfície.
- Grafiquemos y como função de x,



Preço vs. Superfície construída

- O interesse está na construção de um **modelo preditivo** do preço, que permita estimar o preço de venda de uma casa ainda não vendida.
- Seja $\mathbf{x}$ o vetor de atributos da casa cujo preçõ desejamos estimar. Procuramos uma função $f : \mathbb{R}^D \to \mathbb{R}$ tal que
$$\widehat{y} = f(\mathbf{x})$$
seja um bom estimador de y.
- Os modelos que consideramos são afins, i.e., da forma
$$f(\mathbf{x}) = \theta_0 + \theta_1 x_1 + ... + \theta_D x_D = \theta_0 + \boldsymbol{\theta}^T \mathbf{x}$$
Muitas vezes se escolhe uma variável unitária $x_0 = 1$ para adicionar um termo constante (igual a θ_0). Isto corresponde a adicionar uma coluna de 1's à matriz $\mathbf{X}$. Com essa convenção o modelo se escreve
$$f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$$
sendo, agora, $\mathbf{x} = (1, x_1, ..., x_D)$ e $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_D)$.
- No caso do modelo preço vs. superfície,
$$\widehat{y} = \text{preço estimado} = \theta_0 + \theta_1 x_1$$
onde x_1 é a superfície construída.
- Cada valor dos parâmetros $\boldsymbol{\theta}$ corresponde a um modelo diferente. Podemos fazer isto explícito escrevendo
$$\widehat{y} = f(\mathbf{x}, \boldsymbol{\theta}) \ .$$

- **Aprender é descobrir os parâmetros**. Ou **estimar** os parâmetros. Ou **treinar** o modelo. É o passo onde usamos os **dados de treinamento**.
- Seja $(\mathbf{X}, \mathbf{y})$ um conjunto de dados de treinamento. Procuramos valores $\boldsymbol{\theta}^*$ tais que o modelo $f(\cdot, \boldsymbol{\theta}^*)$
  - **ajuste** bem os dados de treinamento, i.e.,

$$y_n \simeq f\left(\mathbf{x}_n, \theta^*\right), \qquad n = 0, 2, \ldots, N-1 \ ,$$

  - **generalize** bem a valores de $\mathbf{x}$ não presentes em $\mathbf{X}$,

$$y \simeq f\left(\mathbf{x}, \theta^*\right), \qquad \mathbf{x} \notin \mathbf{X} \ .$$

- Notar que o primeiro item (ajuste) poderia ser satisfeito **de maneira exata** com a interpolada dos dados. Mas **a interpolada não generaliza bem**.

2 Ajuste dos dados: Minimização do risco empírico

- Queremos achar um modelo $f^*(\mathbf{x}) = f(\mathbf{x}, \theta^*)$ tal que

$$y \simeq f^*(\mathbf{x}) \ .$$

- Especificamos uma **função de perda** $\ell(y, \widehat{y})$ que represente o erro cometido por estimar y com $\widehat{y}$. No caso de **mínimos quadrados** a função de perda é

$$\ell(y, \widehat{y}) = (y - \widehat{y})^2$$

- O ideal seria minimizar o **risco verdadeiro**, ou **risco esperado**, isto é

$$R_{\text{true}}(f) = \mathbb{E}\left[\ell\left(y, f(x)\right)\right] \ .$$

- Como não temos acesso à distribuição verdadeira, substituímos o risco verdadeiro pelo **risco empírico** (simplesmente a média da perda nos dados de treino):

$$R_{\text{emp}}(f, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{n=0}^{n<N} \ell\left(y_n, \widehat{y}_n\right) \ ,$$

  onde $\widehat{y}_n = f(\mathbf{x}_n, \boldsymbol{\theta})$. Notar que R_{emp} pode ser visto também como função de $\boldsymbol{\theta}$, $\mathbf{X}$ e $\mathbf{y}$.

- A minimização do risco empírico é a estratégia de aprendizado que determina os parâmetros $\boldsymbol{\theta}^*$ através da minimização de R_{emp}.

3 O teorema de Gauss-Markov

- O risco empírico, no caso de mínimos quadrados, pode ser escrito da forma matricial:

$$R_{\text{emp}}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{n=0}^{n<N} (y_n - \theta_0 - \theta_1 X_{n1} - ... - \theta_D X_{nD})^2 \; = \frac{1}{N} \left\| \mathbf{y} - \mathbf{X}\, \boldsymbol{\theta} \right\|^2$$

  onde $\| \cdot \|$ é a norma euclidiana em $\mathbb{R}^N$. Para convencerse disto, notar que $\theta_0 + \theta_1 X_{n1} + ... + \theta_D X_{nD}$ é igual a $\mathbf{x}_n \cdot \boldsymbol{\theta}$, que a sua vez corresponde ao produto escalar da linha n de $\mathbf{X}$ com o vetor $\boldsymbol{\theta}$.

- **Exemplo:** No caso dos imóveis, com a superfície como única variável explicativa, resulta

$$R_{\text{emp}} = \frac{1}{N} \sum_{n=0}^{n<N} \left(P_n - \widehat{P}_n \right)^2 = \frac{1}{N} \sum_{n=0}^{n<N} (P_n - \theta_0 - \theta_1\, S_n)^2$$

  onde P_n é o preço (variável resposta $y = P$), $\widehat{P}_n$ é o preço estimado (estimador $\widehat{y}$) e S_n é a superfície, todos correspondendo ao datapoint n. O vetor $\mathbf{x}$ é, então, dado por

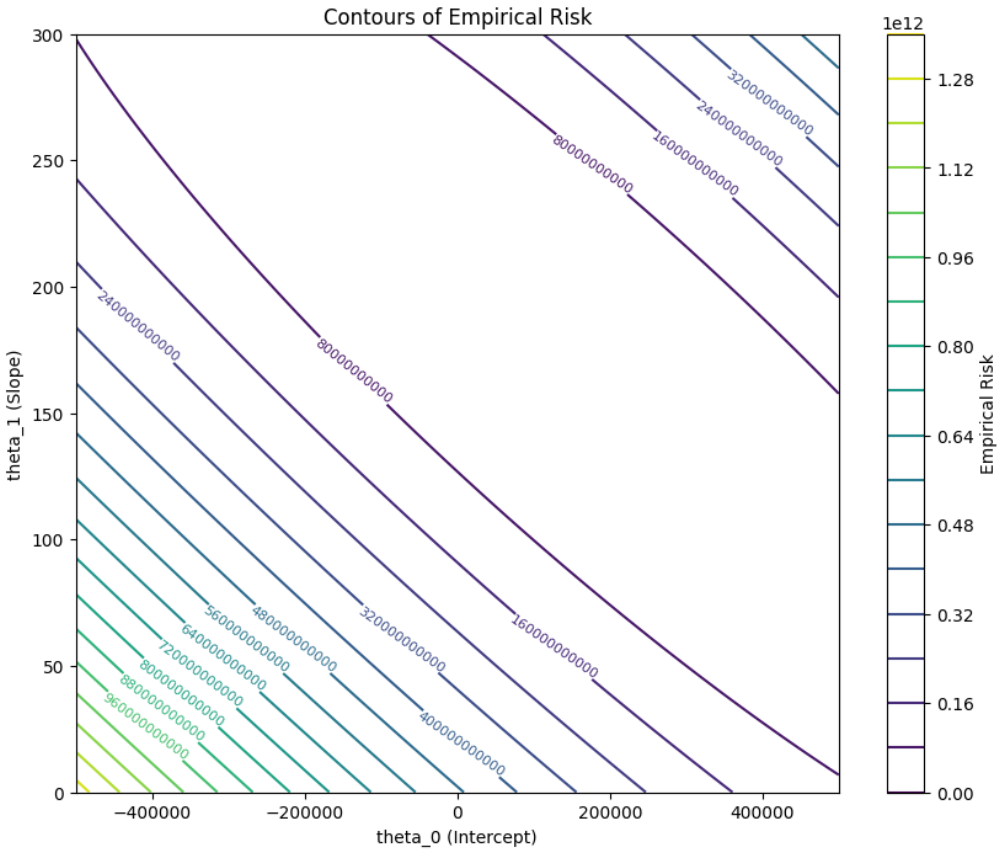
$$\mathbf{x} = (1, S)$$

  e o estimador correspondente a parâmetro $\boldsymbol{\theta}$ é

$$\widehat{y} = \widehat{P} = f\,(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x} \cdot \boldsymbol{\theta} = \theta_0 + \theta_1 S\ .$$

  Assim, com as definições anteriores,

$$\mathbf{X} = \begin{pmatrix} 1 & S_0 \\ 1 & S_1 \\ ... & ... \\ 1 & S_n \\ ... & ... \\ 1 & S_{N-1} \end{pmatrix}\ , \qquad \mathbf{y} = \begin{pmatrix} P_0 \\ P_1 \\ ... \\ P_n \\ ... \\ P_{N-1} \end{pmatrix}\ , \qquad \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$$



Contornos de nível do risco empírico.

Teorema[1]**:**

Consideremos uma relação linear entre as variáveis y e $\mathbf{x}$, da forma

$$y = \mathbf{x} \cdot \boldsymbol{\theta}_{\text{true}} + \epsilon$$

onde ϵ é um ruído aleatório de média zero e desvio padrão σ.

Seja $\mathbf{X}$ uma matriz de exemplos e $\mathbf{y}$ o vetor correspondente de respostas, a partir dos quais deseja-se estimar $\boldsymbol{\theta}_{\text{true}}$.

Então o **melhor estimador linear não-viesado (BLUE)** [2] de $\boldsymbol{\theta}_{\text{true}}$ é o **estimador de mínimos quadrados** $\boldsymbol{\theta}_{\text{MQ}}$, obtido minimizando a soma de quadrados $\|\mathbf{y} - \mathbf{X}\,\theta\|^2$:

$$\theta_{MQ} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{D+1}} \|\mathbf{y} - \mathbf{X}\,\boldsymbol{\theta}\|^2$$

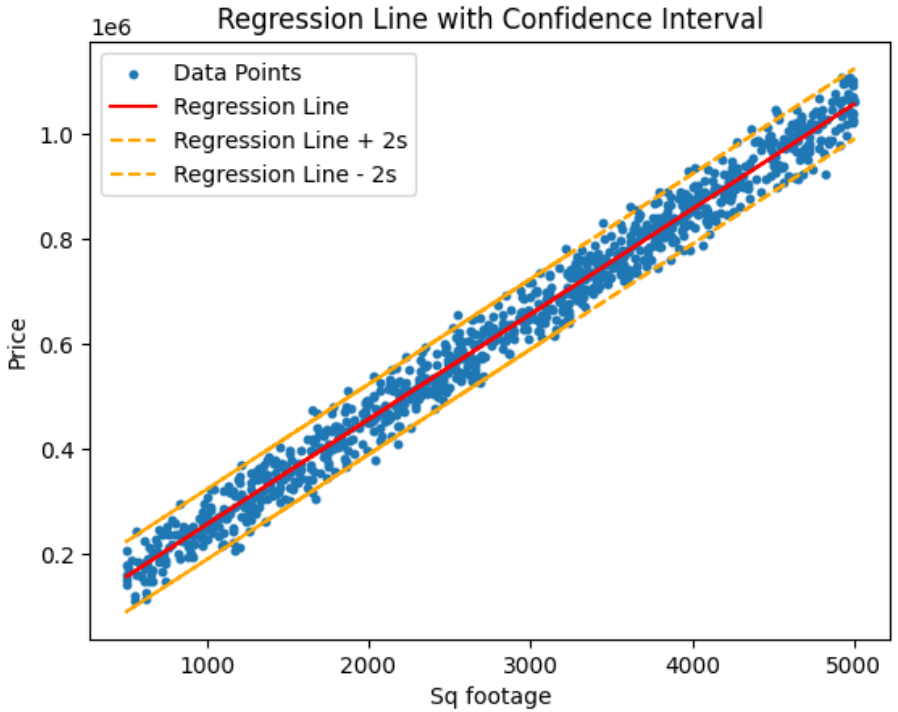
Ademais, o valor esperado da forma quadrática

$$s^2 = \frac{1}{N - (D+1)} \|\mathbf{y} - \mathbf{X}\,\boldsymbol{\theta}_{\text{MQ}}\|^2$$

é a variança do ruído: σ^2.

[1] Ver, por exemplo, A. Björk, *Numerical Methods for Least Squares Problems*, SIAM, 1996.

[2] Linear, no sentido de que $\boldsymbol{\theta}_{\text{MQ}}$ depende linearmente de $\mathbf{y}$, não-viesado, no sentido que $\mathbb{E}\left[\boldsymbol{\theta}_{\text{MQ}}\right] = \boldsymbol{\theta}_{\text{true}}$, e melhor, no sentido de ter variança mínima.



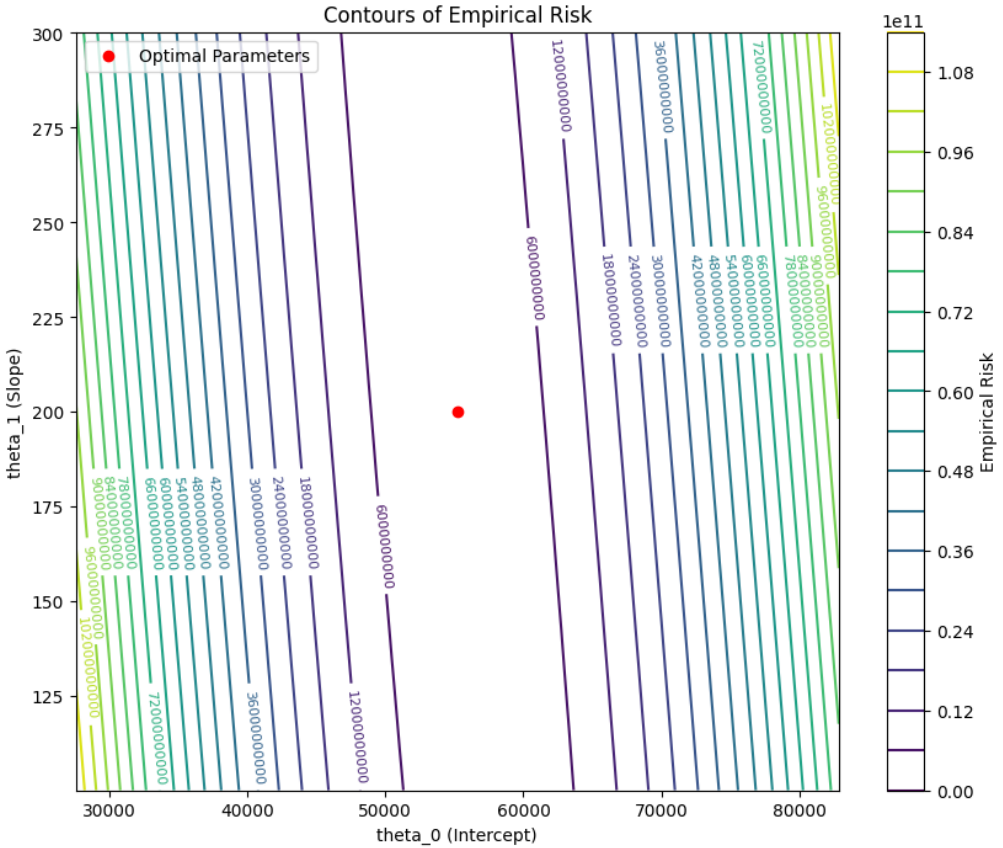
Preço vs. Superfície construída. A linha de regressão e o estimador s correspondem ao estimador de mínimos quadrados θ_{MQ}.

No exemplo, obtemos

$$\begin{aligned} \boldsymbol{\theta}_{\mathsf{MQ}} &= (\theta_0 = 55217.67\ ,\ \theta_1 = 200.20) \\ R_{\mathsf{emp}}\left(\boldsymbol{\theta}_{\mathsf{MQ}}\right) &= 1.11769 \times 10^9 \\ s &= 33465.38 \end{aligned}$$

Erro quadrático médio

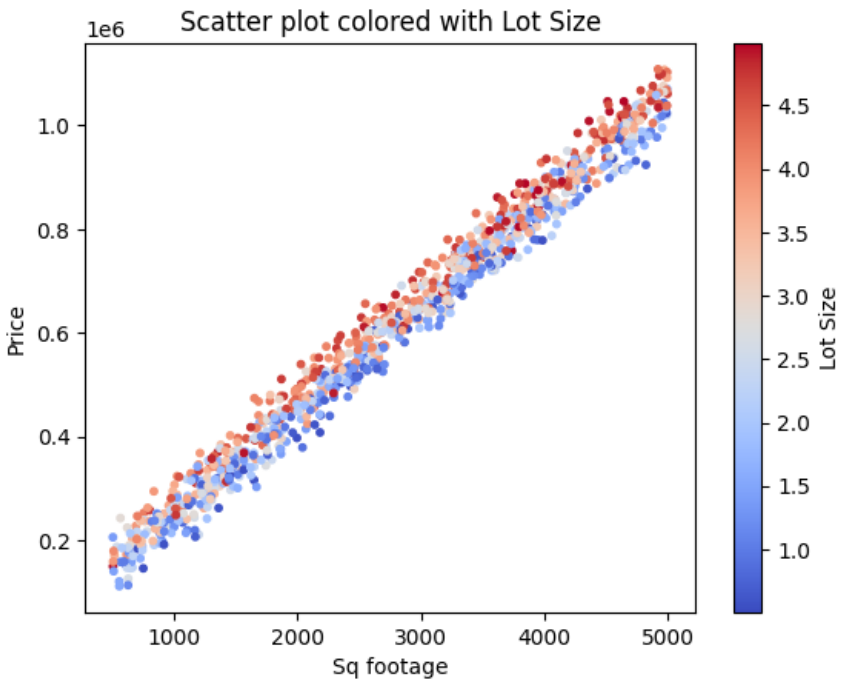
$$e = \sqrt{R_{\mathsf{emp}}} = \sqrt{\frac{1}{N}\sum_{n=0}^{n<N}\left(y_n - \widehat{y}_n\right)^2} = 33431.90$$



Contornos de nível do risco empírico na vizinhança de θ_{MQ}.

4 Duas variáveis explicativas

- Para melhorar a predição, podemos tentar explicar a variabilidade de y adicionando mais variáveis explicativas.
- A partir da visualização resulta claro que a variável $x_2 = T =$`Lot_size` é uma boa candidata.



Observa-se que a parte superior da faixa de variabilidade contem pontos cujo tamanho de terreno é maior que a média (para a mesma superfície construída).

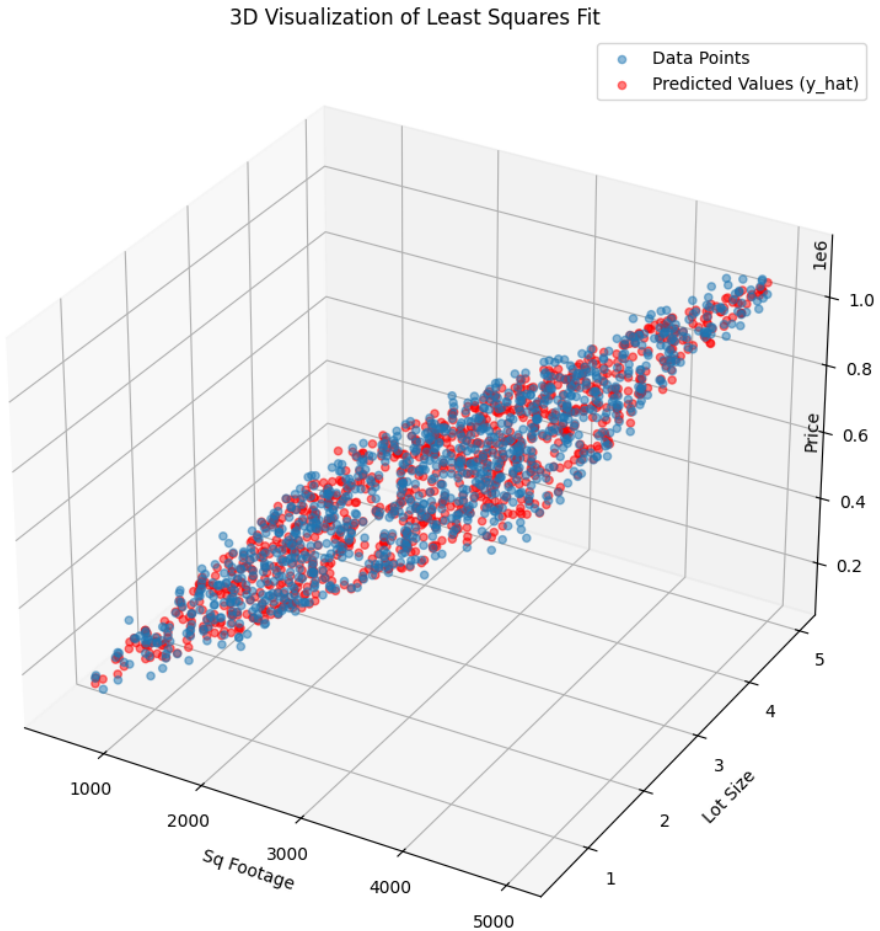
- Trocamos então para $\mathbf{x} = (x_0 = 1, x_1 = S, x_2 = T)$, utilizando então, na matriz $\mathbf{X}$, as colunas `Sq_Footage` e `Lot_size` dos dados de treinamento, além da primeira coluna ser de 1' s.
- O vetor $\mathbf{y}$ permanece o mesmo.
- Calculamos o novo estimador $\boldsymbol{\theta}_{\text{MQ}}$ minimizando $R_{\text{emp}} = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\,\boldsymbol{\theta}\|^2$, com o resultado

$$\boldsymbol{\theta}_{\text{MQ}} = (\theta_0 = 19658.57, \theta_1 = 198.89, \theta_2 = 14123.86)$$

Erro quadrático médio

$$e = \sqrt{R_{\text{emp}}} = \sqrt{\frac{1}{N}\sum_{n=0}^{n<N}(y_n - \widehat{y}_n)^2} = 28012.06$$

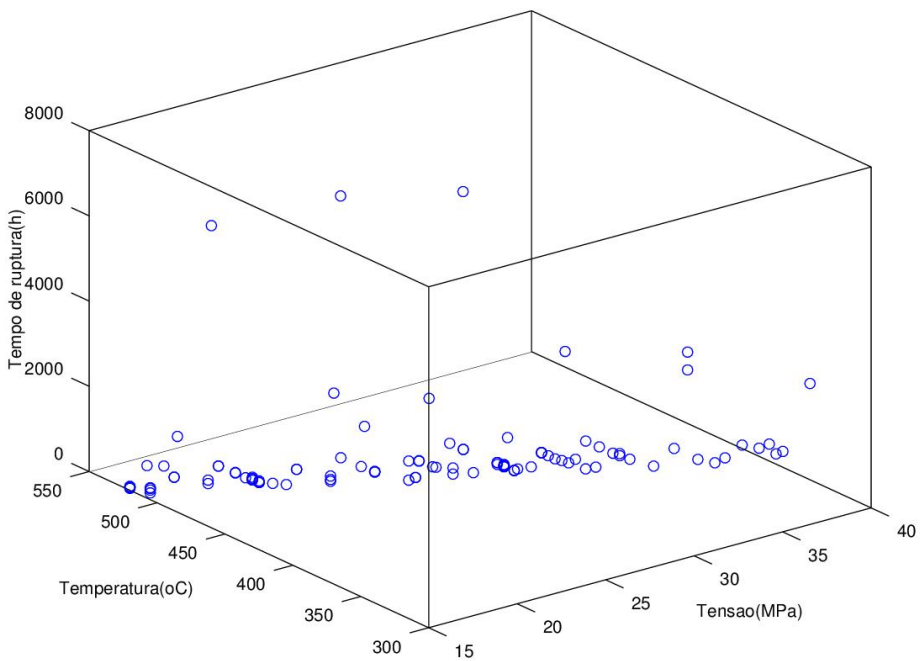
(o valor com apenas `Sq_Footage` como variável explicativa era de 33431.90).



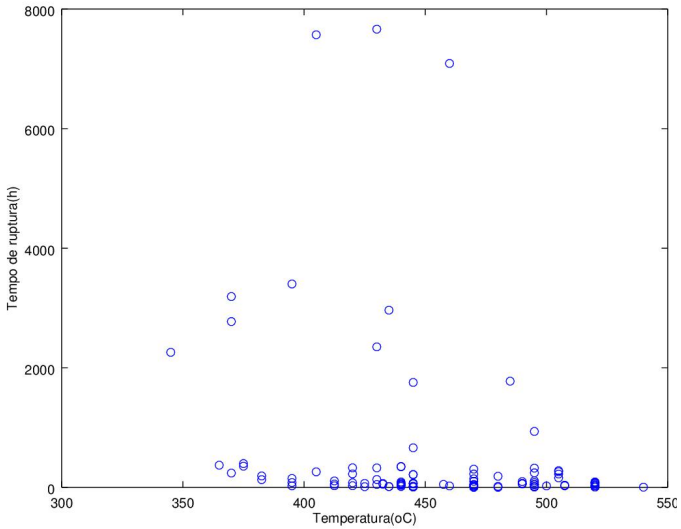
Visualização 3D do ajuste fornecido por θ_{MQ} no caso de duas variáveis explicativas.

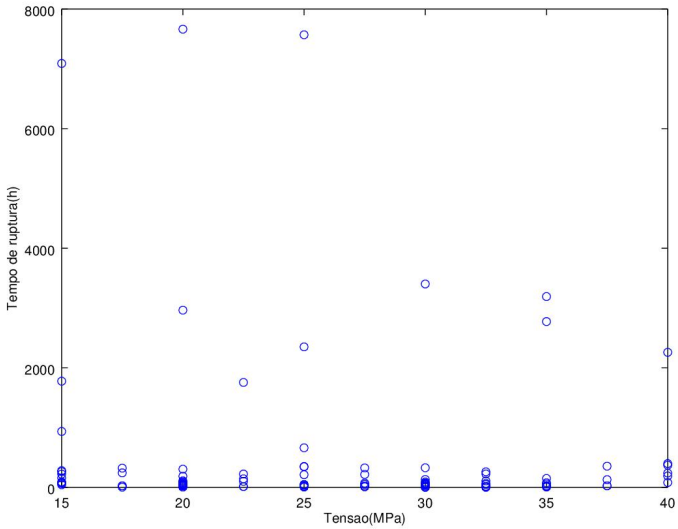
5 Atributos não lineares: Um exemplo da ciência de materiais

- Uma empresa que vende peças metálicas que operam a altas temperaturas e tensões precisa fazer testes de resistência do seu material antes de vendê-lo. Dessa forma considere que os dados sobre os experimentos realizados pela empresa foram fornecidos por um arquivo de texto, que contém a tensão σ (MPa) aplicada ao material, a temperatura T(ºC) do experimento e t_R, o tempo de ruptura (em horas).



- Procurando conseguir ajustar esses dados a alguma função, podemos verificar visualizando as variáveis duas a duas que temos sempre uma nuvem de dados.

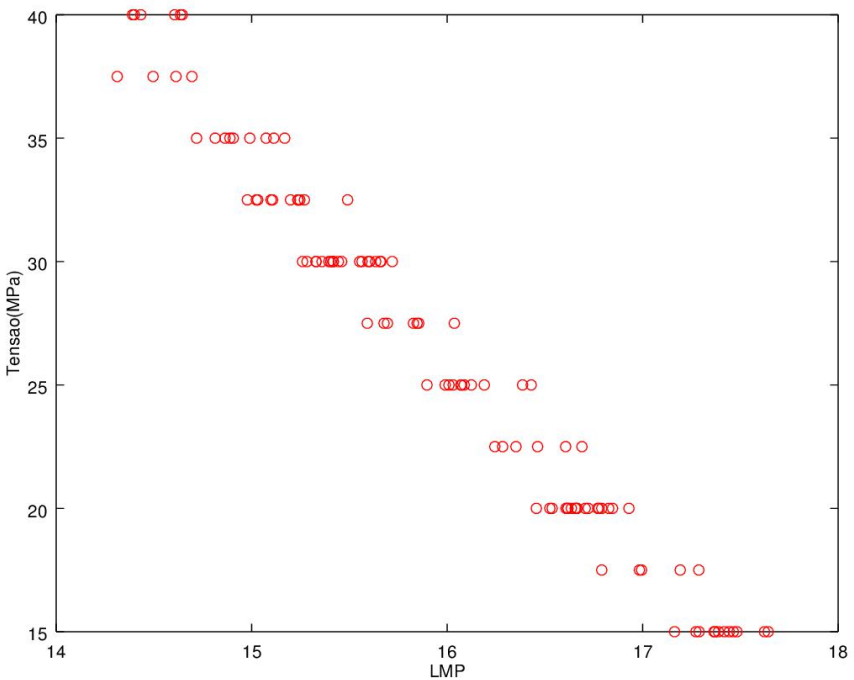




- Para relacionar melhor os parâmetros, fazendo uma redução de dimensão do problema, introduz-se o parâmetro de Larson-Miller (LMP) que é calculado segundo dados do experimento pela equação:

$$LMP = \frac{(273 + T) \cdot (20 + log(t_R))}{1000},$$

em que $T(C)$ é a temperatura aplicada e t_R é o tempo de ruptura (em horas).



- Dado um conjunto de valores referentes a testes realizados por essa empresa (`DadosFicticios.txt`), sabemos que estes devem ser ajustados com a equação de *Spera*:

$$LMP = \theta_0 + \theta_1 \log_{10}(\sigma) + \theta_2 \sigma + \theta_3 \sigma^2, \tag{1}$$

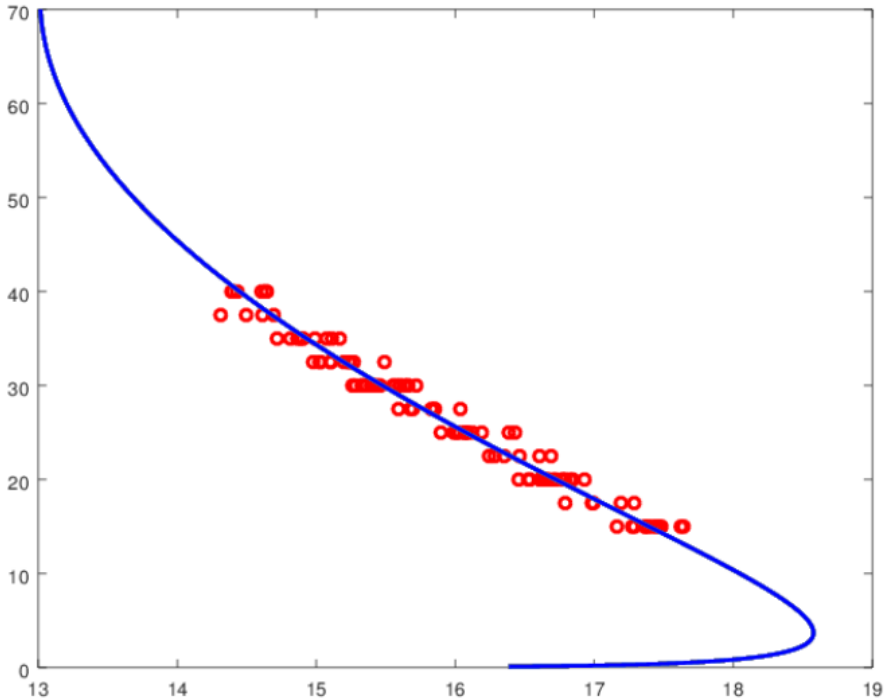
  em que os coeficientes $\boldsymbol{\theta}$ são desconhecidos. Como melhor determinar estes coeficientes a partir dos dados fornecidos?

- Seguimos a mesma metodologia:
  - Tomamos $y = LMP$ como variável resposta. Calculamos o vetor $\mathbf{y}$ a partir de

$$y_n = LMP_n = \frac{(273 + T_n) \cdot (20 + log(t_{Rn}))}{1000}.$$

  - Tomamos $\mathbf{x} = (x_0 = 1, x_1 = \log_{10} \sigma, x_2 = \sigma, x_3 = \sigma^2)$ e construimos a matriz de exemplos $\mathbf{X}$ cuja linha n é $\mathbf{x}_n$. Cada coluna proporciona um feature (covariável) linearmente independente.
  - Calculamos $\boldsymbol{\theta}_{\text{MQ}}$ minimizando o risco empírico de mínimos quadrados $R_{\text{exp}} = \frac{1}{N} \|\mathbf{y} - \mathbf{X}\,\boldsymbol{\theta}\|^2$.
  - O estimador finalmente é $\widehat{y}(\mathbf{x}) = \mathbf{x} \cdot \boldsymbol{\theta}_{\text{MQ}}$.

- Resultado:



LMP (em horizontal) vs. σ (em vertical), mostrando os datapoints e a curva $\widehat{y}(\sigma)$.

6 Algebra linear do estimador de mínimos quadrados

Seja **X** uma matriz de exemplos de N linhas e $M = D + 1$ colunas, e seja **y** um vetor coluna em $\mathbb{R}^N$. Vamos supor que $N > M$ e que o **posto** de X é M (suas colunas são l.i.).

Lembremos que um caso típico é quando são realizadas N medições/amostras de um fenômeno aleatório multivariado. Na medição número n, são coletadas as variáveis x_1, x_2, ..., x_D, e também a variável resposta y. Adicionando a variável $x_0 = 1$ e colocando na forma matricial resulta

$$X_{nj} = \text{ valor da variável } x_j \text{ na medição número } n \ ,$$

$$y_n = \text{ valor da variável resposta } y \text{ na medição número } n \ .$$

Procuramos valores de θ_0, θ_1, ..., θ_D (parâmetros) para ajustar a variável resposta como

$$y \simeq \widehat{y} = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_D x_D \ ,$$

com o **mínimo risco empírico** possível, definido como o **mínimo erro quadrático médio**,

$$R_{\text{emp}} = \frac{1}{N} \sum_{n=0}^{n<N} (\widehat{y}_n - y_n)^2 \ .$$

Teorema: Se X é matriz $N \times M$, de posto $M = D + 1$, com $N \geq M$, então

a) **Existe** um valor de $\boldsymbol{\theta}$ (que denotaremos $\boldsymbol{\theta}^*$ ou $\boldsymbol{\theta}_{\text{MQ}}$), a **solução de mínimos quadrados** do problema de ajuste) que minimiza, sobre o espaço de parâmetros $\mathbb{R}^M$, o resíduo quadrático médio

$$R(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=0}^{n<N} \left(\underbrace{\theta_0 X_{n0} + \theta_1 X_{n1} + ... + \theta_D X_{nD} - y_n}_{r_n} \right)^2 .$$

Isto é, $R(\boldsymbol{\theta}^*) \leq R(\boldsymbol{\theta})$ para todo $\boldsymbol{\theta} \in \mathbb{R}^M$. Notar que r_n é o **resíduo** do ajuste do datapoint número n.

b) O **gradiente** da função R, i.e., o vetor coluna $G(\boldsymbol{\theta}) = (\partial R/\partial\theta_0, ..., \partial R/\partial\theta_D)^T$, é dado por

$$G_k(\boldsymbol{\theta}) = \frac{\partial R}{\partial \theta_k} = \frac{2}{N} \sum_{n=0}^{n<N} \left(\underbrace{\theta_0 X_{n0} + \theta_1 X_{n1} + ... + \theta_D X_{nD} - y_n}_{r_n} \right) X_{nk}$$

e a **matriz Hessiana** de R, i.e., a matriz $H_{kj} = \partial G_k/\partial\theta_j = \partial^2 R/\partial\theta_k\partial\theta_j$ é dada por

$$H_{kj} = \frac{2}{N} \sum_{n=0}^{n<N} X_{nj} X_{nk} .$$

c) Em termos matriciais, se cumpre que

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{X}\,\boldsymbol{\theta} - \mathbf{y}, \qquad R(\boldsymbol{\theta}) = \frac{1}{N}\|\mathbf{r}(\boldsymbol{\theta})\|^2 = \frac{1}{N}\|\mathbf{X}\,\boldsymbol{\theta} - \mathbf{y}\|^2 = \frac{1}{N}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}),$$

onde $\|\cdot\|$ é a norma euclideana usual ($\|\mathbf{r}\|^2 = \mathbf{r}^T\mathbf{r}$).

d) Por tanto, o ajuste de mínimos quadrados **é também a solução de mínimos resíduos (em média quadrática) do sistema linear retangular**

$$\mathbf{X}\,\boldsymbol{\theta} = \mathbf{y}\ .$$

e) Notar também que

$$G(\boldsymbol{\theta}) = \frac{2}{N}\,\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}), \qquad H(\boldsymbol{\theta}) = \frac{2}{N}\,\mathbf{X}^T\mathbf{X}\ .$$

f) A solução de mínimos quadrados θ^* é solução do sistema linear (equações normais)

$$\mathbf{X}^T\mathbf{X}\ \boldsymbol{\theta}^* = \mathbf{X}^T\mathbf{y}\ .$$

Notar que $\boldsymbol{\theta}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$. O estimador é **linear** em $\mathbf{y}$.

A matriz $\mathbf{X}^T\mathbf{X}$ é simétrica, definida positiva (porque o posto de $\mathbf{X}$ é M), e consequentemente não singular ($\det(\mathbf{X}^T\mathbf{X}) > 0$). Por tanto, $\boldsymbol{\theta}^*$ é único.

g) O vetor $\widehat{\mathbf{y}} = \mathbf{X}\,\boldsymbol{\theta}^*$ é o vetor de **valores previstos** nas posições $\mathbf{X}$. Cumpre-se que $\widehat{\mathbf{y}}$ a **projeção ortogonal** de $\mathbf{y}$ sobre o **espaço de colunas de X** (também chamado de espaço **imagem** de $\mathbf{X}$).

h) A previsão para um $\mathbf{x}$ arbitrário é

$$\widehat{y}(\mathbf{x}) = \mathbf{x}^T\,\boldsymbol{\theta}^*\ .$$

Em termos dos dados $(\mathbf{X}, \mathbf{y})$ (notar linearidade em $\mathbf{y}$)

$$\widehat{y}(\mathbf{x}) = \mathbf{x}^T\ \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

7 Comentários e conclusões

- A minimização do erro quadrático médio (risco empírico) sobre os dados de treinamento é uma estratégia frequente em aprendizado supervisionado, i.e.,

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_{\text{MQ}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_n \left(f(\mathbf{x}_n, \boldsymbol{\theta}) - y_n \right)^2$$

  O estimador para y é

$$\widehat{y}(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}^*) \, .$$

- Quando o modelo $f(\mathbf{x}, \boldsymbol{\theta})$ é linear, da forma

$$f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta} = \theta_0 x_0 + \theta_1 x_1 + \ldots$$

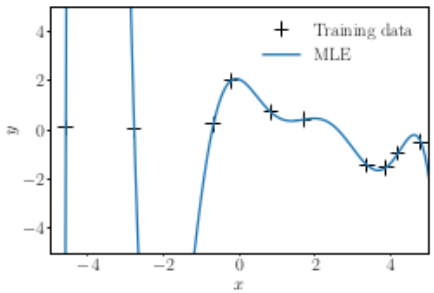
  o risco empírico é $R_{\text{emp}} = \frac{1}{N} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2$. Portanto, $\boldsymbol{\theta}^*$ é a **solução de mínimos quadrados** do sistema retangular $\mathbf{X}\boldsymbol{\theta} = \mathbf{y}$.

- No caso linear, a estratégia é justificada pelo teorema de Gauss-Markov, que prova que $\boldsymbol{\theta}^*$ é um BLUE (best linear unbiased estimator).

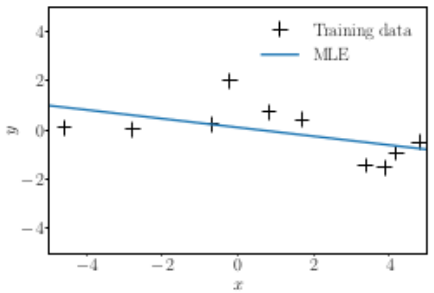
- Se as colunas de $\mathbf{X}$ são linearmente independentes

$$\boldsymbol{\theta}^* = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

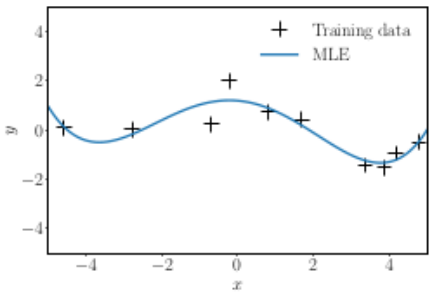
- Em termos probabilísticos, pode ser provado também que **o estimador θ_{MQ} é o estimador máxima verosimilhança** quando o ruído é Gaussiano (não correlacionado).



(a) Overfitting



(b) Underfitting.



(c) Fitting well.

- A estimação de máxima verosimilhança, assim como a estimação por minimização do risco empírico, podem sofrer de **overfitting**. A solução dessa dificuldade é usar regularização, estimação MAP (maximum a posteriori) ou outros métodos fazem um compromisso entre erro empírico e complexidade.

- A restrição de linearidade de $f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T\boldsymbol{\theta}$ pode ser aliviada usando **features** $\phi(\mathbf{x})$ não lineares. Essa metodologia é totalmente equivalente à usada na interpolação,

  $$\widehat{y}(\mathbf{x}) = \theta_0\phi_0(\mathbf{x}) + \theta_1\phi_1(\mathbf{x}) + ...$$

  vendo-se que os features cumprem o papel que, no cálculo da interpolada, era cumprido pelas funções de base. A matriz de exemplos **X** (às vezes chamada de **Φ** nesse caso), coincide com a matriz **M**.

- Quando o número M de features é igual ao número de datapoints N o estimador $\widehat{y}(\mathbf{x})$ coincide com a função interpolante e o risco empírico é zero.

- Porém, continúa sendo uma restrição forte. Por exemplo, o modelo aparentemente simples

  $$f(x, \theta) = \begin{cases} 1 & \text{se } x \leq \theta \\ 0 & \text{se não} \end{cases}$$

  não é linear em θ.