2 *Fundamentos de Regressão Linear*

Dada a importância do tema para a apresentação das propostas desenvolvidas nesta tese, neste capítulo são apresentados conceitos básicos sobre o problema de regressão linear simples e múltipla, bemo como sobre o problema de estimação de parâmetros através do método dos mínimos quadrados.

2.1 O Problema de Regressão Linear

Em muitos problemas práticos, há duas ou mais variáveis numéricas que parecem estar intrisicamente relacionadas, sendo necessário analisar a natureza matemática dessa relação de maneira mais formal a fim de entender melhor o problema. A análise de regressão é uma técnica estatística cujo objetivo principal reside justamente na investigação das relações entre duas ou mais variáveis e na consequente modelagem matemática do problema. A análise de regressão pode ser usada, por exemplo, na construção de um modelo que expresse o resultado de uma variável como função de uma ou mais variáveis. Esse modelo pode, então, ser usado para predizer o resultado de uma variável em função da outra (HINES et al., 2006).

Na descrição que se segue, assume-se que exista uma única variável dependente, ou de resposta, $y \in \mathbb{R}$, relacionada com p variáveis independentes, ou *regressoras*, $x_1, x_2, \ldots, x_p$, $x_j \in \mathbb{R}$. A variável de resposta y é uma variável aleatória, enquanto que as variáveis regressoras $x_1, x_2, \cdots, x_p$ são medidas com erro desprezível e são frequentemente controladas pelo experimentador (usuário). Isto posto, a relação entre y e as p variáveis regressoras é comumente escrita da seguinte forma:

$$y = f(x_1, x_2, \ldots, x_p|\beta) + \varepsilon, \tag{2.1}$$

$$= f(\mathbf{x}|\beta) + \varepsilon, \tag{2.2}$$

em que $f(\cdot|\cdot)$ é denominada de função de regressão, $\mathbf{x} \in \mathbb{R}^p$ é o vetor de variáveis regressoras, $\beta \in \mathbb{R}^p$ é o vetor de parâmetros da função regressora, e ε denota o erro (ruído) aleatório, de

média zero e variância σ_ε^2, presente na medição de y. Assume-se também que ε é uma variável aleatória independente, ou seja, amostras de ε são independentes entre si.

Note que o modelo descrito na Equação (2.1) é um modelo teórico, uma vez que, em geral, nem a função de regressão $f(\cdot|\cdot)$, nem a componente aleatória ε são conhecidas. A escolha da forma funcional da equação de regressão $f(\cdot)$ é feita com base em informação *a priori*, fruto de conhecimento prévio acerca do problema; ou então, através de experimentação com diferentes formas funcionais. A escolha da forma funcional mais adequada a um dado problema é feita (ou pelo menos deveria ser) através de rigoroso processo de análise dos resultados das predições para cada forma funcional escolhida.

Qualquer que seja a forma funcional da equação de regressão, o seu vetor de parâmetros β deve ser estimado. Para isso, faz-se necessário medir um conjunto de n valores de y e de suas variáveis regressoras $\{x_1, x_2, \ldots, x_p\}$:

$$(y_i, x_{i1}, x_{i2}, \cdots, x_{ip}), \qquad i = 1, \ldots, n, \tag{2.3}$$

ou, em forma condensada, faz-se necessário coletar n pares entrada-saída $(y_i, \mathbf{x}_i)$, $i = 1, \ldots n$.

A estimativa do vetor β é simbolizada como $\hat{\beta}$, sendo ela utilizada na seguinte equação para predizer novos valores da variável de resposta:

$$\begin{aligned} \hat{y} &= \hat{f}(x_1, x_2, \ldots, x_p|\hat{\beta}), & (2.4)\\ &= \hat{f}(\mathbf{x}|\hat{\beta}), & (2.5) \end{aligned}$$

em que $\hat{f}(\cdot|\cdot)$ denota uma aproximação da função de regressão do modelo teórico.

A análise de regressão é dita linear quando se assume que a relação matemática entre as variáveis de interesse é uma função linear de seus parâmetros. Neste caso, o modelo de regressão teórico passa a ser chamado modelo de *regressão linear múltipla*, sendo escrito como

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon & (2.6)\\ &= \beta^T \mathbf{x} + \varepsilon & (2.7) \end{aligned}$$

em que o vetor de parâmetros $\beta \in \mathbb{R}^{p+1}$ contém $p+1$ componentes, $\beta_j, j = 0, 1, \cdots, p$, chamadas genericamente de coeficientes de regressão. Como conseqüência, a primeira componente do vetor $\mathbf{x} \in \mathbf{R}^{p+1}$ é igual a 1, sendo as restantes as próprias variáveis regressoras $\{x_1, x_2, \ldots, x_p\}$.

Os modelos de regressão linear múltipla são usados, em geral, como funções aproximadoras, e a equação de regressão é ajustada ao conjunto de pares entrada saída $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. Lembrando que a verdadeira relação funcional entre y e $x_1, x_2, \cdots, x_p$ é geralmente descon-

hecida, mas em muitos casos práticos o modelo de regressão linear apresenta-se como uma aproximação adequada (HINES et al., 2006). Nestes casos, a equação de predição passa então a ser escrita como

$$\begin{aligned} \hat{y} = \quad E[y|\mathbf{x},\hat{\beta}] &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p, \\ &= \hat{\beta}^T \mathbf{x}, \end{aligned} \tag{2.8}$$

em que $E[y|\mathbf{x},\hat{\beta}]$ denota o valor esperado da variável de resposta y condicionado ao vetor de variáveis regressoras $\mathbf{x}$ e à estimativa do vetor de parâmetros $\hat{\beta}$. A Equação (2.8) define um hiperplano no espaço p-dimensional das variáveis regressoras x_j.

Além de sua simplicidade, uma vantagem do modelo linear reside na interpretação direta que pode ser dada ao parâmetro β_j, como representando a mudança esperada na resposta y por unidade de mudança em x_j quando todas as demais variáveis independentes $x_i (i \neq j)$ são mantidas constantes, ou seja, $\beta_j = \frac{dy}{dx_j}$. Isto permite identificar diretamente quais variáveis regressoras são mais relevantes para a variável de saída ou, dito de outra forma, quais variáveis regressoras influenciam mais a variável de resposta.

Quando o problema de regressão lienar envolve apenas uma única variável regressora, x, tem-se uma regressão linear simples. Neste caso, a relação matemática entre uma única variável de entrada x e uma variável de saída y é definida por uma reta, ou seja,

$$y = \beta_0 + \beta_1 x_1 + \varepsilon, \tag{2.9}$$

em que β_0 é o intercepto e β_1, a inclinação da reta. Consequentemente, a equação de predição para o problema de regressão linear simples é dada por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1, \tag{2.10}$$

em que $\hat{\beta}_0$ e $\hat{\beta}_1$ são, respectivamente, as estimativas do intercepto e do coeficiente angular da reta de regressão.

Como o modelo de regressão linear múltipla é mais geral que o de regressão simples, todo o desenvolvimento teórico que se segue será feito com base nesta formulação do problema de regressão.

2.2 Estimador de Mínimos Quadrados Ordinário

Como já mencionado na seção anterior, os dados a serem usados para estimar o vetor de parâmetros β consistem de n observações do par entrada-saída $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. Em palavras,

a i-ésima observação inclui uma resposta escalar y_i e o vetor de variáveis regressoras correspondente $\mathbf{x}_i = [x_{i1} \;\; x_{i2} \;\; \cdots \;\; x_{ip}]^T$, em que x_{ij} denota a i-ésima observação da j-ésima variável regressora.

Assim, para o modelo de regressão linear múltipla, a variável resposta é uma função linear das variáveis regressoras:

$$\begin{aligned} y_i &= \beta^T \mathbf{x}_i + \varepsilon_i, && (2.11) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i && (2.12) \end{aligned}$$

em que ε_i corresponde à i-ésima observação do erro aleatório. Para a formulação que se segue, assume-se que as seguintes suposições são verdadeiras:

1. Existem (muito) mais observações que incógnitas (i.e. $n \gg p$).
2. O erro ou ruído no modelo (ε) tem média 0 e variância σ_ε^2.
3. As observações $\{\varepsilon_i\}$ são não-correlacionadas.

Em forma expandida, o modelo de regressão mostrado na Equação (2.11) corresponde a um sistema de equações com n equações e $p+1$ incógnitas, ou seja

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{aligned} \qquad (2.13)$$

O sistema de equações mostrado na Equação (2.13) pode ser escrito em notação matricial como

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad (2.14)$$

em que os vetores $\mathbf{y} \in \mathbb{R}^n$ e $\varepsilon \in \mathbb{R}^n$, assim como a matriz de regressão $\mathbf{X} \in \mathbf{R}^n \times \mathbf{R}^{p+1}$, são definidos como

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)},$$

com o vetor de parâmetros $\beta \in \mathbb{R}^{p+1}$ e o vetor aleatório $\varepsilon \in \mathbb{R}^n$ sendo definidos por

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} \quad \text{e} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}.$$

Usar a técnica de mínimos quadrados para encontrar estimativas para os coeficientes de regressão β_j, $j = 1, \ldots, n$, corresponde a minimizar a seguinte função-custo:

$$J(\beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2. \tag{2.15}$$

Dessa forma, minimizar a função-custo $J(\beta_1, \beta_2, \ldots, \beta_p)$ equivale a fazer com que a soma dos quadrados dos desvios ε_i entre os valores observados de y_i e o hiperplano de regressão seja mínima. Em forma vetorial, a função-custo $J(\beta_1, \beta_2, \ldots, \beta_p)$ pode ser escrita como

$$J(\beta) = \|\varepsilon\|^2 = \varepsilon^T \varepsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta), \tag{2.16}$$

em que $\|\cdot\|$ denota a norma euclidiana de um vetor

A forma vetorial permite também adicionar uma outra interpretação ao problema de minimizar a função-custo $J(\beta)$: minimizar esta função-custo corresponde a encontrar uma estimativa do vetor de parâmetros β que produza o vetor de erros aleatórios com menor norma quadrática.

As expressões para cálculo das estimativas de β, denotadas como $\hat{\beta}$, podem ser obtidas a partir da minização da função-custo dos erros quadráticos na forma escalar (Equação 2.15) ou na forma vetorial (Equação 2.16). Ambos os casos são apresentados a seguir.

2.2.1 Equações Normais dos Mínimos Quadrados (Forma Escalar)

A função $J(\beta_1, \beta_2, \ldots, \beta_p)$ deve ser minimizada individualmente em relação a cada um dos parâmetros $\beta_0, \beta_1, \cdots, \beta_p$, ou seja, parâmetro a parâmetro. Para isso, a derivada parcial de $J(\beta_1, \beta_2, \ldots, \beta_p)$ deve ser tomada em relação a cada parâmetro β_j, $j = 1, \ldots, p$ e igualada a zero, ou seja

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right) = 0 \tag{2.17}$$

e

$$\frac{\partial J}{\partial \beta_j} = -2\sum_{i=1}^{n} x_{ij}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right) = 0, \qquad j = 1,2,\ldots,p. \tag{2.18}$$

Resolvendo as Equações (2.17) e (2.18), obtemos um sistema de equações conhecido como *equações normais de mínimos quadrados*, em sua forma escalar:

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} x_{i1} + \hat{\beta}_2\sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_p\sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} y_i, \tag{2.19}$$

$$n\hat{\beta}_0\sum_{i=1}^{n} x_{i1} + \hat{\beta}_1\sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_p\sum_{i=1}^{n} x_{i1}x_{ip} = \sum_{i=1}^{n} x_{i1}y_i, \tag{2.20}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$n\hat{\beta}_0\sum_{i=1}^{n} x_{ip} + \hat{\beta}_1\sum_{i=1}^{n} x_{ip}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n} x_{ip}x_{i2} + \ldots + \hat{\beta}_p\sum_{i=1}^{n} x_{ip}^2 = \sum_{i=1}^{n} x_{ip}y_i. \tag{2.21}$$

Note que existem $p+1$ equações normais, uma para cada coeficiente de regressão; logo, o sistema acima é quadrado. A solução das equações normais produz as estimativas de mínimos quadrados dos coeficientes de regressão $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ (HINES et al., 2006).

2.2.2 Equações Normais dos Mínimos Quadrados (Forma Vetorial)

Para o caso vetorial, a função-custo $J(\beta)$ mostrada na Equação (2.16) precisa primeiro ser decomposta da seguinte forma:

$$\begin{aligned} J(\beta) &= = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta), && (2.22)\\ &= \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta, && (2.23)\\ &= \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta, && (2.24) \end{aligned}$$

em que se fez uso do fato de o termo $\beta^T\mathbf{X}^T\mathbf{y}$ ser um escalar e, portanto, seu transposto $\beta^T\mathbf{X}^T\mathbf{y}^T = \mathbf{y}^T\mathbf{X}\beta$ resulta no mesmo escalar.

Portanto, para minimizar o funcional $J(\beta)$, deve-se tomar a sua derivada parcial em relação ao vetor de parâmetros β e igualá-la ao vetor nulo $\mathbf{0} \in \mathbb{R}^{p+1}$, ou seja

$$\frac{\partial J}{\partial \beta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{0}, \tag{2.25}$$

que, simplificando, resulta em

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}, \tag{2.26}$$

que corresponde à versão vetorial das equações normais dos mínimos quadrados mostradas nas

Equações (2.19) a (2.21).

Para encontrar a solução das equações normais, multiplica-se ambos os lados da Equação 2.26 pela inversa de $\mathbf{X}^T\mathbf{X}$. Assim, a estimativa de mínimos quadrados ordinário (MQO) do vetor de parâmetros β é dada por

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.27}$$

que também é conhecida como solução pelo método da pseudoinversa de Moore-Penrose (GOLUB; VAN LOAN, 1996).

O modelo de predição linear baseado na estimativa MQO do vetor de parâmetros é dado por

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}, \tag{2.28}$$

em que $\hat{\mathbf{y}} \in \mathbb{R}^n$ é o vetor de predições da variável resposta. O vetor de erros de predição é então dado por $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, sendo que a sua norma quadrática é a menor possível $\|\mathbf{e}\|^2$, segundo o critério dos mínimos quadrados.

Na notação escalar, a equação de predição é dada por

$$\hat{y}_i = \hat{\beta}^T\mathbf{x}_i = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j x_{ij}, \tag{2.29}$$

para $i = 1, 2, \ldots, n$, e o erro correspondente é dado por $e_i = y_i - \hat{y}_i$. A soma dos erros quadráticos (SEQ), $\sum_{i=1}^{n} e_i^2$, que nada mais é do que a norma quadrática $\|\mathbf{e}\|^2$, é comumente usada como critério de avaliação da qualidade da regressão.

2.3 Estimador de Mínimos Quadrados Regularizado

Muitas vezes, a matriz $\mathbf{X}^T\mathbf{X}$ é singular (i.e. não é de posto completo) ou muito próxima da singularidade. Neste caso, tem-se que

$$det(\mathbf{X}^T\mathbf{X}) \approx 0, \tag{2.30}$$

fato este que pode comprometer toda a validade do processo de inferência da regressão linear múltipla, pois é fonte de instabilidades numéricas durante o cálculo da estimativa MQO do vetor de parâmetros mostrada na Equação (2.27). Isto ocorre geralmente quando as variáveis de entrada são intercorrelacionadas. Quando essa intercorrelação é grande, dizemos que existe *multicolinearidade*, ou seja, as linhas da matriz $\mathbf{X}^T\mathbf{X}$ não são linearmente independentes.

A fim de evitar problemas numéricos, faz-se necessário utilizar estratégias que permitam estimar β de modo confiável. Para este fim, um dos métodos mais conhecidos é o *método de regularização de Thikonov* (HOERL; KENNARD, 1970). Utilizando regularização de Thikonov, o estimador de mínimos quadrados de β é dado por

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.31}$$

em que

- $0 \leq \lambda \ll 1$ é uma constante de valor bem pequeno, e
- $\mathbf{I}$ é uma matriz identidade de dimensão $(p+1) \times (p+1)$.

A regressão que utiliza esse tipo de regularização é chamada de regressão de cumeeira (*ridge regression*) e a função-custo associada é dada por

$$J(\beta) = \|\varepsilon\|^2 + \lambda\|\beta\|^2, \tag{2.32}$$

que permite interpretar a Equação (2.31) como aquela que produz uma estimativa do vetor de parâmetros $\hat{\beta}$ que tenta satisfazer dois critérios de otimalidade:

- Um que procura minimizar a norma do erro quadrático, ou seja, $\|\mathbf{e}\|^2$,
- E um outro que procura minimizar a norma do vetor de parâmetros, ou seja, $\|\beta\|^2$,

tal que a importância do segundo critério frente ao primeiro é regulada pelo valor de λ.

2.4 Regressão Linear no Matlab

A versão vetorial da solução dos mínimos quadrados (Equação (2.27)), assim como a sua versão regularizada mostrada na Equação (2.31), são úteis não apenas pela notação matemática compacta, mas principalmente porque sua implementação em ambientes de computação científica, tais como Matlab©, Octave e Scilab, é possível sem maior esforço de programação.

Como estes ambientes são amplamente utilizados em Engenharia e Ciências, existem diversos comandos e funções que geram, em princípio, os mesmos resultados numéricos para um dado problema de regressão linear simples ou múltipla. Contudo, alguns comandos são mais eficientes, seja porque consomem menos tempo, seja porque são menos susceptíveis a erros numéricos. Isto posto, vale a pena fazer alguns comentários sobre este tema e sugerir

formas eficientes de implementação das equações de estimação de parâmetros pelo método dos mínimos quadrados, ordinário ou regularizado.

Assumindo que o vetor de observações da variável de resposta $\mathbf{y}$ e a matriz de variáveis regresssoras $\mathbf{X}$ são denotadas como `y` e `X` no Matlab, então a Equação (2.27) pode ser implementada da forma que se lê, ou seja,

```
» B = inv(X'*X)*X'y;
```

em que `B` denota a estimativa MQO do vetor de parâmetros β. Contudo, esta forma de se estimar β não é a mais recomendada para problemas de maior escala por duas razões: (*i*) Possui elevado custo computacional, e (*ii*) é muito susceptível a erros numéricos.

Para problemas maiores recomenda-se usar o operador *barra invertida* (\), ou seja

```
» B = X\y;
```

Os cálculos realizados pelo uso do operador *barra invertida* (\) baseiam-se em grande parte no método de ortogonalização conhecido como fatoração QR.

Uma segunda maneira de estimar o vetor β é através do comando `PINV`:

```
» B = pinv(X)*y;
```

Embora a obtenção da solução via comando `pinv` seja computacionalmente mais custosa que a obtida pelo uso do operador *barra invertida*, visto que é baseada na técnica conhecida como SVD (*singular value decomposition*), ela é preferível em problemas em que a matriz de variáveis regressoras $\mathbf{X}$ possui deficiência de posto (i.e. tem posto incompleto).

Por fim, uma terceira maneira de se estimar o vetor β no Matlab é por meio do comando `REGRESS`:

```
» B = regress(y,X);
```

Para a versão regularizada do estimador MQO, também é possível estimar $\hat{\beta}$ através da escrita direta da Equação (2.31) no prompt do Matlab:

```
» l = 0.01;
» I=ones(size(X'*X));
» B = inv(X'*X + l*I)*X'y;
```

Porém, pelas mesmas razões apontadas anteriormente, recomenda-se também usar o operador *barra invertida* (\). Neste caso, a seqüência de comandos passa ser a seguinte:

```
» l = 0.01;
» I=eye(size(X'*X));
» A=X'*X + l*I;
» r=X'*y;
» B = A\r;
```

Um exemplo de aplicação dos comandos acima em um problema real será apresentado logo a seguir. Este exemplo, além de ser útil do ponto de vista didático, será retomado no próximo capítulo, quando abordaremos o problema de regressão robusta.

2.4.1 Exemplo Numérico: Regressão Linear Simples

A título de ilustração, vamos aplicar os conceitos de regressão linear apresentados neste capítulo ao conjunto de dados mostrado na Tabela 2.1. Este conjunto está disponível em Freedman et al. (2007) e envolve duas variáveis: uma é variável regressora $x \in \mathbb{R}$ (consumo per capita de cigarros em um dado país em 1930), enquanto a variável resposta $y \in \mathbb{R}$ corresponde ao número de mortes (por milhão de pessoas) por câncer de pulmão naquele país em 1950.

Índice	País	Cigarro per capita	Mortes por milhão de pessoas
1	Austrália	480	180
2	Canadá	500	150
3	Dinamarca	380	170
4	Finlândia	1100	350
5	Grã Bretanha	1100	460
6	Islândia	230	60
7	Holanda	490	240
8	Noruega	250	90
9	Suécia	300	110
10	Suíça	510	250

Tabela 2.1: Consumo per capita de cigarros em vários países em 1930 e as taxas de morte por câncer de pulmão em 1950.

Como o conjunto de dados envolve apenas um variável regressora é possível visualizar os pares (x_i, y_i), $i = 1, \ldots, 10$ em um diagrama de dispersão (*scatterplot*), marcando um círculo em cada coordenada, conforme ilustrado na Figura 2.1.

Pode-se perceber pelo diagrama de dispersão que há uma tendência linear nos dados, ou seja, há uma tendência de os pontos se organizarem ao longo de uma reta hipotética. Esta reta é, na verdade, a reta de regressão, cujos parâmetros podem ser calculados facilmente usando os

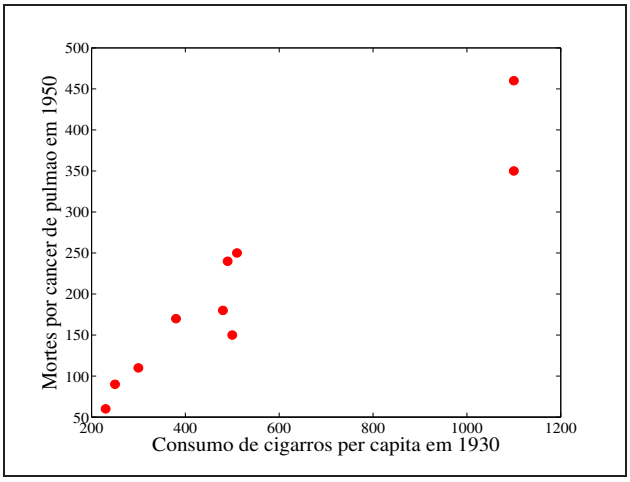

Figura 2.1: Diagrama de dispersão para o conjunto de dados da Tabela 2.1.

comandos do Matlab discutidos na seção anterior. Para isso, podemos usar a seguinte seqüência de comandos:

```
» x=[480; 500; 380; 1100; 1100; 230; 490; 250; 300; 510];
» y=[180; 150; 170; 350; 460; 60; 240; 90; 110; 250];
» n=length(x);
» X=[ones(n,1) x];
» B=X\y
B=
9.1393
0.3687
```

Assim, tem-se que $\hat{\beta}_0 = 9{,}14$ e $\hat{\beta}_1 = 0{,}37$, com a equação da reta de regressão sendo dada por $\hat{y}_i = 9,14 + 0,37x$. O gráfico da reta de regressão superposta aos pontos do conjunto de dados é mostrado na Figura 2.2.

No próximo capítulo, este mesmo conjunto de dados, adicionado de mais um par de pontos (x_i, y_i) relativo ao consumo de cigarros nos Estados Unidos, será usado para introduzir conceitos de regressão robusta, ou seja, na presença de pontos discrepantes (*outliers*).

2.5 Resumo do Capítulo

Esse capítulo apresentou os conceitos fundamentais sobre o problema de regressão linear, principalmente o de regressão linear múltipla. A análise de regressão trata da modelagem e investigação das relações entre duas ou mais variáveis, ou seja, da relação entre uma variável

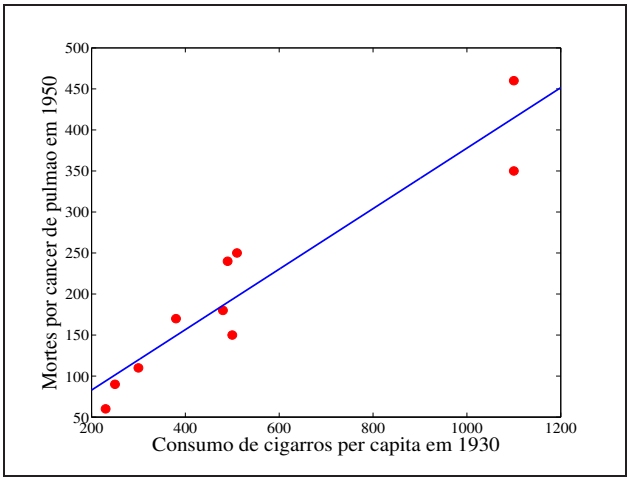

Figura 2.2: Gráfico da reta de regressão ajustada aos dados da Tabela 2.1.

resposta (ou de saída), e uma ou mais variáveis regressoras (independentes). A regressão linear simples se dá quando existe apenas uma variável regressora, e é definida por uma reta, enquanto que a regressão linear múltipla envolve mais de uma variável regressora, sendo definida por um plano.

Para estimar os parâmetros do modelo de regressão linear apresentado neste capítulo foi utilizado o método dos mínimos quadrados ordinário, bem como uma variante regularizada deste método. Além disso, foram apresentadas e discutidas diferentes meios para se implementar o método dos mínimos quadrados, ordinário e regularizado, em ambientes de computação científica, tais como Matlab©e Octave, tendo em vista questões relacionadas a problemas de natureza numérica.

Por fim, um conjunto de dados real foi utilizado com o propósito de ilustrar os conceitos introduzidos neste capítulo, principalmente para mostrar que, graças à formulação vetorial do estimador dos mínimos quadrados, a implementação deste método no Matlab é bem simples e direta.

No próximo capítulo este mesmo conjunto de dados, adicionado de mais um par de pontos (x_i, y_i) relativo ao consumo de cigarros nos Estados Unidos, será usado para introduzir conceitos de regressão robusta, ou seja, na presença de pontos discrepantes (*outliers*).

Referências Bibliográficas

ANDERSON, J. A simple neural network generating an interactive memory. *Mathematical Biosciences*, v. 14, n. 3-4, p. 197–220, 1972.

ANDREWS, D. F. A robust method for multiple linear regression. *Technometrics*, v. 16, n. 4, p. 523–531, 1974.

AUGUSTEIJN, M. F.; FOLKERT, B. A. Neural network classification and novelty detection. *International Journal of Remote Sensing*, v. 23, n. 14, p. 2891–2902, 2002.

BADDELEY, R. et al. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. In: *Proc. R. Soc. Lond.* [S.l.: s.n.], 2005. p. 1775–1783.

BAEK, D.; OH, S.-Y. Improving optimal linear associative memory using data partitioning. In: *Proceedings of the 2006 IEEE International Conference on Systems, Man, and Cybernetics (SMC'06)*. [S.l.: s.n.], 2006. v. 3, p. 2251–2256.

BAI, Z. D.; WU, Y. General M-estimation. *Journal of Multivariate Analysis*, v. 63, p. 119–135, 1997.

BARNETT, V.; LEWIS, T. *Outliers in Statistical Data*. 3a.. ed. [S.l.]: John Wiley & Sons, 1994.

BARRETO, G. A. *Redes Neurais Não-Supervisionadas para Processamento de Seqüências Temporais*. Dissertação (Mestrado) — Departamento de Engenharia Elétrica, Universidade de São Paulo, São Carlos, SP, 1998.

BARRETO, G. A.; FROTA, R. A. A unifying methodology for the evaluation of neural network models on novelty detection tasks. *Pattern Analysis and Applications*, v. 16, n. 1, p. 83–972, 2013.

BELLHUMER, P. N.; HESPANHA, J.; KRIEGMAN, D. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Special Issue on Face Recognition*, v. 17, n. 7, p. 711–720, 1997.

BEN-GAL, I. Outlier detection. In: MAIMON, O.; ROKACH, L. (Ed.). *Data Mining and Knowledge Discovery Handbook*. [S.l.]: Springer, 2005. p. 131–146.

BLATNÁ, D. Outliers in regression. In: *Proceedings of the 9th International Scientific Conference on Applications on Mathematics and Statistics in Economy (AMSE'2006)*. [S.l.: s.n.], 2006. p. 1–6.

BOCCATO, L. *Novas Propostas e Aplicaçõs de Redes Neurais com Estados de Eco*. Tese (Doutorado) — Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade de Campinas, São Paulo, 2013.

BOULESTEIX, A.-L. PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, v. 3, n. 1, p. 1–32, 2004.

BRANHAM, R. L. Alternatives to least squares. *The Astronomical Journal*, v. 87, n. 6, p. 928–937, 1982.

BRATTON, D.; KENNEDY, J. Defining a standard for particle swarm optimization. In: *Proceedings of the IEEE Swarm Intelligence Symposium*. Honolulu, Hawaii: [s.n.], 2007. p. 120–127.

CHATTERJEE, S.; MÄCHLER, M. Robust regression: a weighted least squares approach. *Communications in Statistics - Theory and Methods*, v. 26, n. 6, p. 1381–1394, 1997.

CHERKASSKY, V.; FASSETT, K.; VASSILAS, N. Linear algebra approach to neural associative memories and noise performance of neural classifiers. *IEEE Transactions on Computers*, v. 40, n. 12, p. 1429–1435, 1991.

CUDMORE, R. H.; DESAI, N. S. Intrinsic plasticity. *Scholarpedia*, v. 3, n. 2, p. 1363, 2008.

DENG, W.; ZHENG, Q.; CHEN, L. Regularized extreme learning machine. In: *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09)*. [S.l.: s.n.], 2009. p. 389–395.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2nd. ed. [S.l.]: John Wiley & Sons, 2006.

EICHMANN, G.; KASPARIS, T. Pattern classification using a linear associative memory. *Pattern Recognition*, v. 22, n. 6, p. 733–740, 1989.

EMMERICH, C.; REINHART, F.; STEIL, J. Recurrence enhances the spatial encoding of static inputs in reservoir networks. In: *Proceedings of the 20th International Conference on Artificial Neural Networks*. [S.l.]: Springer, 2010. LNCS 6353, p. 148–153.

FAWZYA, A.; MOKHTAR, H. M.; HEGAZY, O. Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal*, 2013.

FOX, J. *Applied Regression Analysis, Linear Models, and Related Methods*. [S.l.]: Sage Publications, 1997.

FOX, J. *Robust Regression: Appendix to An R and S-PLUS Companion to Applied Regression*. [S.l.], 2002.

FRANK, A.; ASUNCION, A. *UCI Machine Learning Repository*. 2010. Disponível em: <http://archive.ics.uci.edu/ml>.

FREEDMAN, D.; PISANI, R.; PURVES, R. *Statistics*. 4a.. ed. [S.l.]: W. W. Norton & Company, 2007.

GAUSS, C. F. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. [S.l.]: Perthes et I. H. Besser, Hamburgi, 1809.

GLAVIN, F. G.; MADDEN, M. G. Analysis of the effect of unexpected outliers in the classification of spectroscopy data. In: *Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'09)*. [S.l.: s.n.], 2010. p. 124–133.

GOLUB, G. H.; VAN LOAN, C. F. *Matrix Computations*. 3a.. ed. [S.l.]: The Johns Hopkins University Press, 1996.

HAMPEL, F. Robust statistics: a brief introduction and overview. In: *Robust Statistics and Fuzzy Techniques in Geodesy and GIS*. [S.l.: s.n.], 2001. p. 1–6.

HEBB, D. *The Organization of Behavior*. [S.l.]: New York: Wiley, 1949.

HILL, R. W.; HOLLAND, P. W. Two robust alternatives to least-squares regression. *Journal of the American Statistical Association*, v. 72, n. 360, p. 828–833, 1977.

HINES, W. W. et al. *Probabilidade e Estatística na Engenharia*. Quarta. [S.l.]: LTC, 2006.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies. *Artificial Intelligence Review*, v. 22, n. 2, p. 85–126, 2004.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 42, n. 1, p. 80–86, 1970.

HORATA, P.; CHIEWCHANWATTANA, S.; SUNAT, K. Robust extreme learning machine. *Neurocomputing*, v. 102, p. 31–44, 2012.

HOWELL, A. J. *Automatic Face Recognition using Radial Basis Function Networks*. Tese (Doutorado) — University of Sussex, Brighton, UK, September 1997.

HUANG, G.-B.; WANG, D. H.; LAN, Y. Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, v. 2, p. 107–122, 2011.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing*, v. 70, p. 489–501, 2006.

HUBER, P. J. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, v. 35, n. 1, p. 73–101, 1964.

HUNT, B. et al. Synthesis of a nonrecurrent associative memory model based on a nonlinear transformation in the spectral domain. *IEEE Transactions on Neural Networks*, v. 4, n. 5, p. 873–878, 1993.

JOHNSON, W.; GEISSER, S. A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association*, v. 78, p. 137–144, 1983.

KENNEDY, J.; EBERHART, R. C. Particle swarm optimization. In: *Proceedings of the IEEE International Conference on Neural Networks*. Piscataway, NJ, USA: [s.n.], 1995. v. 4, p. 1942–1948.

KIM, H.-C.; GHAHRAMANI, Z. Outlier robust gaussian process classification. In: *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition (SSPR)'08*. [S.l.: s.n.], 2008. p. 896–905.

KOHONEN, T. Correlation matrix memory. *IEEE Transactions on Computers*, C-21, n. 4, p. 353–359, 1972.

KOHONEN, T.; OJA, E. Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biological Cybernetics*, v. 25, p. 85–95, 1976.

KOHONEN, T.; RUOHONEN, M. Representation of associated data by matrix operators. *IEEE Transactions on Computers*, v. 22, n. 7, p. 701–702, 1973.

LEE, C.-C. et al. Noisy time series prediction using *m*-estimator based robust radial basis function neural networks with growing and pruning techniques. *Expert Systems and Applications*, v. 36, n. 3, p. 4717–4724, 2009.

LEE, C.-C. et al. Robust radial basis function neural networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, v. 29, n. 6, p. 674–685, 1999.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, v. 401, p. 788–791, 1999.

LEE, D. D.; SEUNG, H. S. Algorithms for non-negative matrix factorization. In: PRESS, M. (Ed.). *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*. [S.l.: s.n.], 2000. p. 556–562.

LEGENDRE, A. M. *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. [S.l.]: Courcier, Paris, 1805.

LI, C. A model of neuronal intrinsic plasticity. *IEEE Transactions on Autonomous Mental Development*, v. 3, n. 4, p. 277–284, 2011.

LI, D.; HAN, M.; WANG, J. Chaotic time series prediction based on a novel robust echo state network. *IEEE Transactions on Neural Networks and Learning Systems*, v. 23, n. 5, p. 787–799, 2012.

LIU, N.; WANG, H. Ensemble based extreme learning machine. *IEEE Signal Processing Letters*, v. 17, n. 8, p. 754–757, 2010.

MARONNA, R. A.; MARTIN, D. R.; YOHAI, V. J. *Robust Statistics: Theory and Methods*. 1a.. ed. [S.l.]: John Wiley & Sons, 2006.

MESQUITA, M. E. R. V. Introdução às memórias associativas lineares, morfológicas e fuzzy. In: *Anais do 2o. Colóquio de Matemática da Região Sul (ColMatSul'2012*. [S.l.: s.n.], 2012.

MICHE, Y. et al. OP-ELM: Optimally pruned extreme learning machine. *IEEE Transactions on Neural Networks*, v. 21, n. 1, p. 158–162, 2010.

MICHE, Y. et al. TROP-ELM: a double-regularized ELM using LARS and Tikhonov regularization. *Neurocomputing*, v. 74, n. 16, p. 2413–2421, 2011.

MOHAMMED, A. et al. Human face recognition based on multidimensional PCA and extreme learning machine. *Pattern Recognition*, v. 44, n. 10–11, p. 2588–2597, 2011.

NAKANO, K. Associatron: A model of associative memory. *IEEE Transactions on on Systems, Man, Cybernetics*, SMC-2, n. 3, p. 380–388, 1972.

NETO, A. R. R.; BARRETO, G. A. On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: A comparative analysis. *IEEE Latin America Transactions*, v. 7, n. 4, p. 487–496, 2009.

NEUMANN, K.; STEIL, J. Optimizing extreme learning machines via ridge regression and batch intrinsic plasticity. *Neurocomputing*, v. 102, p. 23–30, 2013.

OES, R. S. B.; LIMA, V. M. C. Comparação de estimadores de regressão. In: *Anais do XIX Simpósio Nacional de Probabilidade e Estatística (SINAPE'2010)*. [S.l.: s.n.], 2010. p. 1–6.

PEDERSEN, M. E. H.; CHIPPERFIELD, A. J. Simplifying particle swarm optimization. *Applied Soft Computing*, v. 10, n. 2, p. 618–628, 2010.

POGGIO, T.; GIROSI, F. Networks for approximation and learning. *Proceedings of the IEEE*, v. 78, n. 9, p. 1481–1497, 1990.

PRASAD, B. et al. A study on associative neural memories. *International Journal of Advanced Computer Science and Applications*, v. 1, n. 6, p. 124–133, 2010.

RAMALHO, G. L. B.; MEDEIROS, F. N. S. Using boosting to improve oil spill detection in sar images. In: *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'2006)*. [S.l.: s.n.], 2006. v. 2, p. 1066–1069.

RAO, C. R.; TOUTENBURG, H. *Linear Models: Least Squares and Alternatives*. 2nd. ed. [S.l.]: Springer, 1999.

RIPLEY, B. D. *Robust Statistics*. [S.l.], 2004.

RITTER, G.; GALLEGOS, M. T. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, v. 18, n. 6, p. 525–539, 1997.

RONCHETTI, E. The historical development of robust statistics. In: *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7)*. [S.l.: s.n.], 2006. p. 1–4.

ROUSSEEUW, P. J.; LEROY, A. M. *Robust Regression and Outlier Detection*. 1a.. ed. [S.l.]: John Wiley & Sons, 1987.

SCHONHOFF, T.; GIORDANO, A. *Detection and Estimation Theory*. 1a.. ed. [S.l.]: Prentice Hall, 2006.

SINGH, S.; MARKOU, M. An approach to novelty detection applied to the classification of image regions. *IEEE Transactions on Knowledge and Data Engineering*, v. 16, n. 4, p. 396–407, 2004.

SMITH, M. R.; MARTINEZ, T. Improving classification accuracy by identifying and removing instances that should be misclassified. In: *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN'2011)*. [S.l.: s.n.], 2011. p. 2690–2697.

STANIMIROVA, I.; WALCZAK, B. Classification of data with missing elements and outliers. *Talanta*, v. 76, n. 3, p. 602–609, 2008.

STEVENS, J. P. Outliers and influential data points in regression analysis. *Psychological Bulletin*, v. 95, n. 2, p. 334–344, 1984.

STILES, G.; DENQ, D.-L. A quantitative comparison of the performance of three discrete distributed associative memory models. *IEEE Transactions on Computers*, v. 36, n. 3, p. 257–263, 1987.

STILES, G. S.; DENQ, D. On the effect of noise on the Moore-Penrose generalized inverse associative memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 7, n. 3, p. 358–360, 1985.

TONG, T. T. *Diagnostics for Outliers and Influential Points*. [S.l.], 2010.

TRIESCH, J. A gradient rule for the plasticity of a neuron's intrinsic excitability. In: *Proc. of the Int. Conf. on Artificial Neural Networks (ICANN)*. [S.l.: s.n.], 2005.

TUKEY, J. W. The future of data analysis. *Annals of Mathematical Statistics*, v. 33, n. 1, p. 1–67, 1964.

VASCONCELOS, G. C.; FAIRHURST, M. C.; BISSET, D. L. Investigating feedforward neural networks with respect to the rejection of spurious patterns. *Pattern Recognition Letters*, v. 16, p. 207–212, 1995.

WEBB, A. *Statistical Pattern Recognition*. 2. ed. [S.l.]: John Wiley & Sons, LTD, 2002.

ZONG, W.; HUANG, G.-B. Face recognition based on extreme learning machine. *Neurocomputing*, v. 74, n. 16, p. 2541–2551, 2011.