MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 6

17 de abril de 2023

Sumário

Um exemplo

- Consideremos um exemplo [proposto em Bishop (2006)] cujo objetivo é ajustar um modelo de regressão polinomial a um conjunto de 10 pontos gerados por meio da expressão $y_i = sen(2\pi x_i) + e_i$ em que e_i segue um distribuição Normal com média nula e variância σ^2.
- Os dados estão representados na Figura 1 por pontos em azul. A curva verde corresponde a $y_i = sen(2\pi x_i)$; em vermelho estão representados os ajustes baseados em regressões polinomiais de graus, 0, 1, 3 e 9.
- Claramente, a curva baseada no polinômio do terceiro grau consegue reproduzir o padrão da curva geradora dos dados sem, no entanto, predizer os dados com total precisão. A curva baseada no polinômio de grau 9, por outro lado, tem um ajuste perfeito, mas não reproduz o padrão da curva utilizada para gerar os dados, Esse fenômeno é conhecido como sobreajuste.

Um exemplo

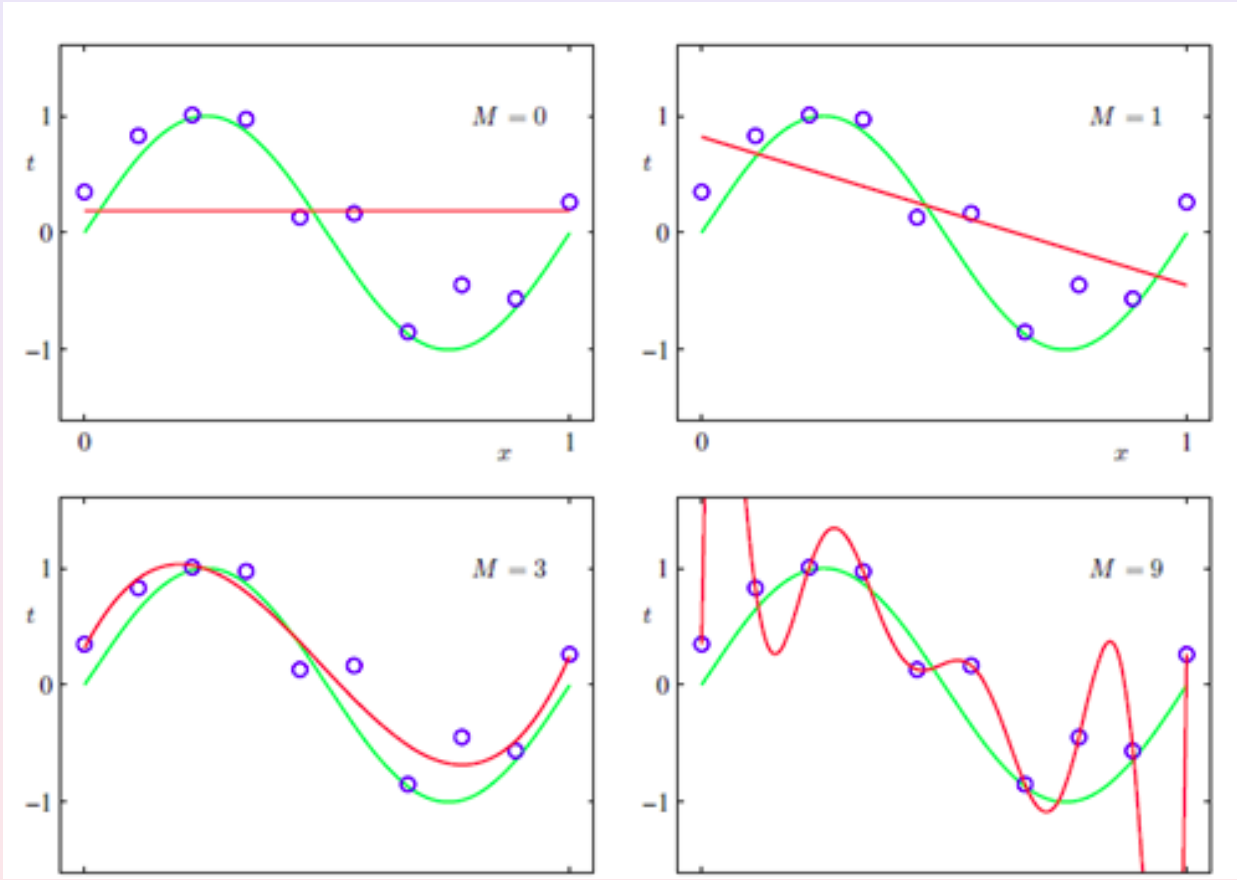

Figura 1: Ajuste de modelos polinomiais a um conjunto de dados hipotéticos.

Quando usar EMQ?

- Se a relação entre resposta e preditores for aproximadamente linear, então um modelo de regressão linear múltipla (RLM) pode ser adequado e estimadores de mínimos quadrados (EMQ) tenderão a ter baixo viés e conduzir a previsões boas.
- Se $n >> p$, EMQ terão também baixa variância.
- Se n não for muito maior do que p, EMQ apresentarão muita variabilidade (sobreajuste) e previsões ruins.
- Se $p > n$, não existirão EMQ univocamente determinados. Aumentando-se o número de variáveis, $R^2 \to 1$, o EQM de treinamento tenderá zero e o EQM de teste crescerá. Não use MQ!

Possíveis abordagens

Possíveis alternativas para remover variáveis irrelevantes de um modelo de RLM, de modo a obter maior interpretabilidade:

- Seleção de subconjuntos de variáveis (subset selection) ; vários procedimentos podem ser usados (stepwise (forward e backward), uso de critérios de informação (AIC, BIC), R^2 ajustado, validação cruzada).
- Encolhimento (shrinkage): usa todos os preditores mas os coeficentes são encolhidos para zero; pode funcionar para selecionar variáveis. Reduz variância. Também chamada regularização.
- Redução da dimensão: projetar os preditores sobre um subespaço de dimensão menor $q < p$, que consiste em obter combinações lineares (ou projeções) dos preditores. Essas q projeções são usadas como novos preditores no ajuste de MQ. ACP, AF, ICA.

Regularização

- O termo regularização refere-se a um conjunto de técnicas utilizadas para especificar modelos que se ajustem a um conjunto de dados evitando o sobreajuste (*overfitting*).
- Essencialmente, essas técnicas servem para ajustar modelos de regressão em que a função de perda contém um termo de penalização cuja finalidade é reduzir a influência de coeficientes responsáveis por flutuações excessivas.
- Embora haja várias técnicas de regularização, consideraremos apenas a regularização L_2, ou Ridge, a regularização L_1 ou Lasso (*least absolute shrinkage and selection operator*) e uma mistura dessas duas, chamada de Elastic net.

Regularização

- O termo de regularização da técnica Lasso usa uma soma de valores absolutos dos parâmetros e um coeficiente de penalização que os encolhe para zero. Essa técnica serve para seleção de modelos, porque associa pesos nulos a parâmetros não significativos.
- Isso implica uma solução esparsa (Dizemos que um modelo é esparso se a maioria dos elementos do correspondente vetor de parâmetros é nula ou desprezável).
- Na regularização L_2, por outro lado, o termo de regularização usa uma soma de quadrados dos parâmetros e um coeficiente de penalização que força alguns pesos a serem pequenos, mas não os anula e consequentemente não conduz a soluções esparsas. Essa técnica de regularização não é robusta com relação a valores atípicos, pois pode conduzir a valores muito grandes do termo de penalização.

O contexto

- Consideremos o modelo de regressão

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

ou

$$y_i = \beta_0 + \boldsymbol{\beta}_i^\top \mathbf{x}_i + e_i, \tag{2}$$

com as p variáveis preditoras reunidas no vetor $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^\top$, y_i representando a variável resposta, e_i indicando as inovações de média zero e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denotando o vetor de parâmetros a serem estimados.

- Vamos considerar $\mathbf{x}_i = (\mathbf{x}_i(1)^\top, \mathbf{x}_i(2)^\top)^\top$, com $\mathbf{x}_i(1) \in \mathbb{R}^s$ o vetor de variáveis relevantes e $\mathbf{x}_i(2) \in \mathbb{R}^{p-s}$ o vetor de variáveis irrelevantes, $\boldsymbol{\beta} = (\boldsymbol{\beta}(1)^\top, \boldsymbol{\beta}(2)^\top)^\top$.
- Objetivos:
  1. Selecione o conjunto de variáveis correto: $\hat{\boldsymbol{\beta}}(1) \neq 0$ e $\hat{\boldsymbol{\beta}}(2) = 0$ (seleção do modelo);
  2. Estime $\boldsymbol{\beta}(1)$ como se o conjunto de variáveis correto fosse conhecido.

Regularização Ridge

- Supomos adicionalmente que $\beta_0 = 0$ e consideremos estimadores de mínimos quadrados (EMQ) penalizados da forma

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \tag{3}$$

  em que λ é o coeficiente de regularização, que controla o número de parâmetros do modelo. Se $\lambda = \infty$, não há variáveis a serem incluídas no modelo e se $\lambda = 0$, obtemos os EMQ usuais.
- Dizemos que $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)$ é o estimador Ridge.

Ridge - Propriedades

Pode-se mostrar que

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{4}$$

em que $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$ é a matriz de especificação do modelo e $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

Alguns resultados sobre as propriedades dessa classe de estimadores são:

1) Em geral, o estimador *Ridge* não é consistente. Sua consistência assintótica vale quando $\lambda = v\lambda_n \to \infty$, $\lambda_n/n \to 0$ e $p < n$.
2) O estimador *Ridge* é enviesado para os parâmetros não nulos.
3) A técnica de regularização *Ridge* não serve para a seleção de modelos.
4) A escolha do coeficiente de regularização λ pode ser feita via validação cruzada ou por meio de algum critério de informação.
5) A técnica de regressão *Ridge* foi introduzida por Hoerl e Kennard (1970) para tratar do problema da multicolinearidade.

Ridge - Propriedades

Obter o mínimo em (3) é equivalente a minimizar a soma de quadrados não regularizada sujeita à restrição

$$\sum_{j=1}^{p} \beta_j^2 \leq m, \tag{5}$$

para algum valor apropriado m, ou seja, é um problema de otimização com multiplicadores de Lagrange.

Na Figura 2 (a) apresentamos um esquema com o valor ótimo do vetor $\boldsymbol{\beta}$, a região circular correspondente à restrição (5) e os círculos representando as curvas de nível da função erro não regularizada.

Ridge - Propriedades

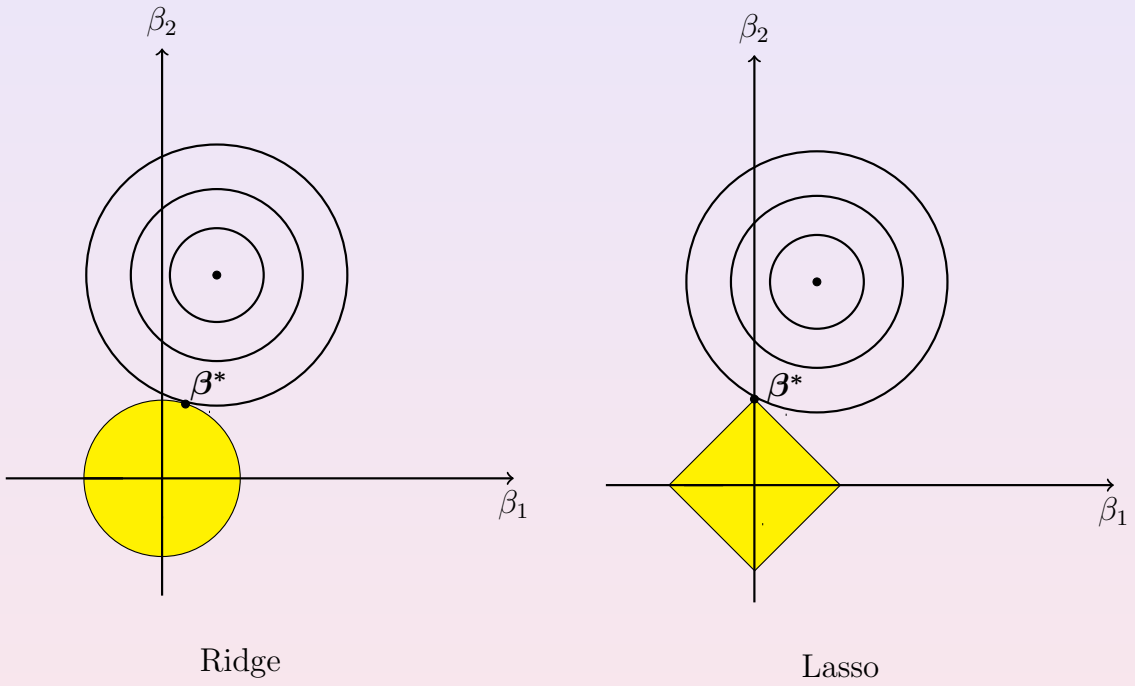

Figura 2: Esparsidade do modelo: (a) Ridge; (b) Lasso.

Regularização Lasso

- Consideremos, agora, o **estimador Lasso**, obtido de

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{6}$$

- Neste caso, a restrição (5) é substituída por

$$\sum_{j=1}^{p} |\beta_j| \leq m, \tag{7}$$

- No painel (b) da Figura 2 (b), podemos observar que a regularização Lasso pode gerar uma solução esparsa, ou seja com $\beta_1^* = 0$.

Regularização Lasso

- Existe uma correspondência $1-1$ entre as formulações (6) e (7): para cada valor de m para a qual (7) vale, existe um valor de λ que fornece a mesma solução para (6). Reciprocamente, a solução $\hat{\boldsymbol{\beta}}(\lambda)$ de (6) resolve o problema restrito, com $m = ||\hat{\boldsymbol{\beta}}(\lambda)||_1$.
- Tanto no caso Ridge, como no Lasso, a constante $1/2n$ pode ser substituída por $1/2$ ou mesmo 1. Esa padronização torna os valores de λ comparáveis para diferentes tamanhos amostrais (por exemplo ao usar CV).
- Em análise convexa, a condição necessária e suficiente para a solução de (6) é

$$-\frac{1}{n}\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\boldsymbol{\beta}\rangle + \lambda s_j = 0, \quad j = 1, \ldots, p, \tag{8}$$

  onde s_j é uma quantidade desconhecida, igual a sinal(β_j), se $\beta_j \neq 0$ e algum valor no intervalo $[-1, 1]$, se $\beta_j = 0$ (sub-gradiente para a função valor absoluto).
- O sistema (8) é uma forma das chamadas condições de Karush-Kuhn-Tucker (KKT) para o problema (6).

Lasso - Propriedades

Algumas propriedades estatísticas do estimador Lasso:

1) O estimador Lasso encolhe para zero os parâmetros que correspondem a preditores redundantes.
2) O estimador é enviesado para parâmetros não nulos.
3) Sob certas condições, o estimador Lasso seleciona as variáveis relevantes do modelo atribuindo pesos nulos aos respectivos coeficientes.
4) O estimador não é consistente em geral.
5) Quando $p = n$, ou seja, quando o número de variáveis preditoras é igual ao número de observações, a técnica Lasso corresponde à aplicação de um limiar suave (*soft threshold*) a $Z_j = \mathbf{x}_j^\top \mathbf{y}/n$, ou seja,

$$\widehat{\beta}_j(\lambda) = \text{sinal}(Z_j)\left(|Z_j| - \lambda/2\right)_+, \tag{9}$$

em que $(x)_+ = \max\{x, 0\}$.

Threshold dado por (9)

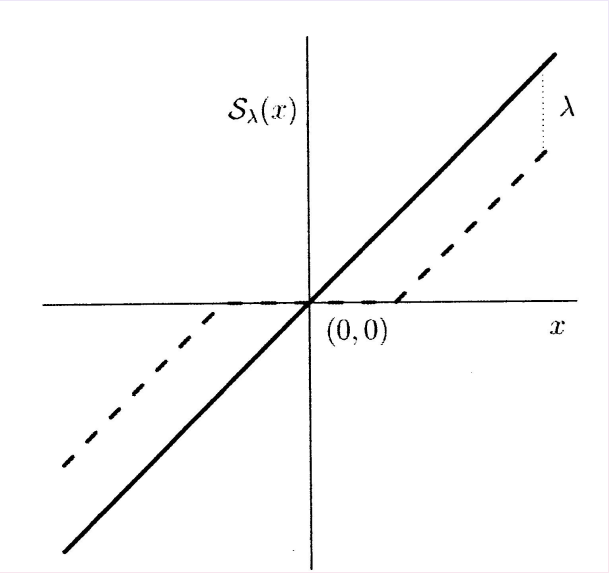

Figura 3: Threshold: MQ (linha cheia) e Lasso (linha tracejada), para o caso $p = n$.

Elastic net

- O estimador *Elastic net* é

$$\widehat{\boldsymbol{\beta}}_{\mathsf{EN}}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} \frac{1}{2n}(y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda_2 \sum_{i=1}^{p} \beta_i^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i|. \quad (10)$$

- Na Figura 4 apresentamos esquematicamente uma região delimitada pela restrição $J(\boldsymbol{\beta}) \leq m$, em que $J(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|$, para algum m, com $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, além daquelas delimitadas pelas restrições *Ridge* e Lasso.
- Pode-se mostrar que sob determinadas condições, o estimador *Elastic Net* é consistente.

Elastic net

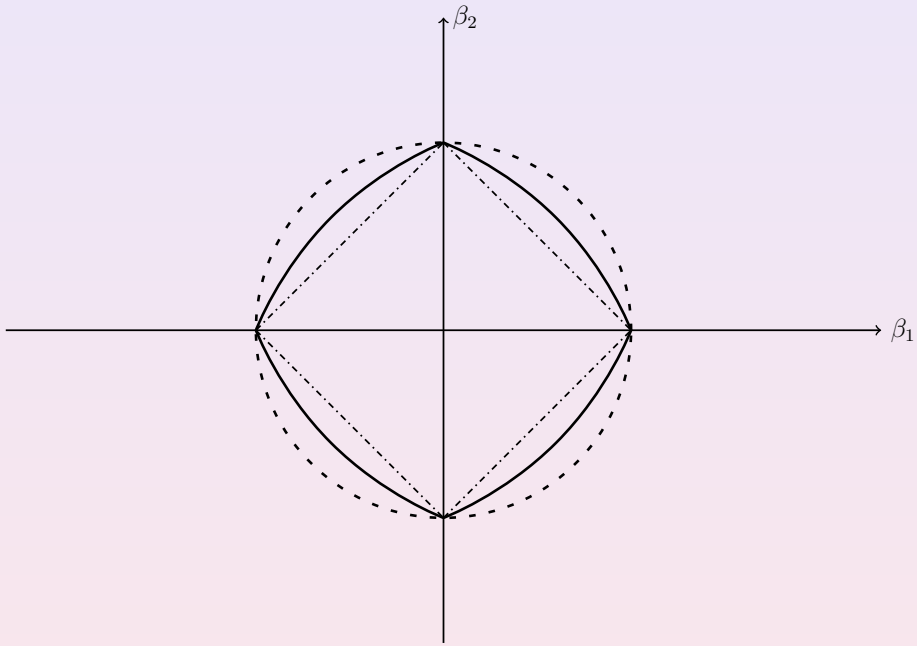

Figura 4: Geometria das restrições *Elastic Net* (curva contínua), *Ridge* (curva tracejada) e Lasso (curva pontilhada).

Lasso adaptativo

- O estimador Lasso adaptativo (adaLASSO) é dado por

$$\widehat{\boldsymbol{\beta}}_{AL}(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + \lambda \sum_{i=1}^{p} w_i |\beta_i|, \qquad (11)$$

  em que $w_1, \ldots, w_p$ são pesos não negativos pré-definidos.
- Usualmente, toma-se $w_j = |\tilde{\beta}_j|^{-\tau}$, para $0 < \tau \leq 1$ e $\tilde{\beta}_j$ é um estimador inicial (por exemplo o estimador Lasso).
- O estimador **Lasso adaptativo** é consistente sob condições não muito fortes.
- A função `adalasso` do pacote `parcor` do `R` pode ser usada para calcular esse estimador.
- O pacote `glmnet` do `R` pode ser usado para obter estimadores Lasso e *Elastic net* sob modelos de regressão linear, regressão logística e multinomial, regressão Poisson além de modelos de Cox. Para detalhes, veja Friedman et al. (2010).

Comparação entre os métodos

- Tanto Ridge como o Lasso encolhem os coeficiente para zero. No caso do Lasso, a penalidade L_1 tem a finalidade de tornar alguns dos coeficientes serem efetivamente nulos. Logo, o Lasso realiza seleção de modelos.
- Lasso resulta em modelos esparsos, ou seja, mais fáceis de interpretar.
- Tanto no Ridge, quanto no Lasso, a variância decresce e o viés cresce à medida que λ cresce.
- Ridge tem desepenho melhor que o Lasso nos caso que um número grande de preditores tem relação com a variável resposta. Em caso contrário, o Lasso tem desempenho melhor (em termos de EQM).
- Em geral, nenhum método domina os outros em todas as situações.

Viés da regularização Ridge

Supondo $p < n$, fazendo $\mathbf{R} = \mathbf{X}^\top \mathbf{X}$, e usando a expressão do estimador *ridge* (3), obtemos

$$\begin{aligned}
\widehat{\beta}_{\text{Ridge}}(\lambda) &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R} [\mathbf{R}^{-1} \mathbf{X}^\top \mathbf{y}) \\
&= [\mathbf{R}(\mathbf{I} + \lambda \mathbf{R}^{-1})]^{-1} \mathbf{R}((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}] \\
&= (\mathbf{I} + \lambda \mathbf{R}^{-1})^{-1} \mathbf{R}^{-1} \mathbf{R} \hat{\beta}_{\text{MQ}} \\
&= (\mathbf{I} + \lambda \mathbf{R}^{-1}) \widehat{\beta}_{\text{MQ}},
\end{aligned}$$

em que $\widehat{\beta}_{\text{MQ}}$ denota o estimador de mínimos quadrados ordinários. Tomando a esperança condicional da expressão anterior, dada $\mathbf{X}$, obtemos

$$\begin{aligned}
E[\widehat{\beta}_{\text{Ridge}}(\lambda) | \mathbf{X}] &= E[(\mathbf{I} + \lambda \mathbf{R}^{-1}) \hat{\beta}_{\text{MQ}}] \\
&= (\mathbf{I} + \lambda \mathbf{R}^{-1}) \beta,
\end{aligned}$$

de onde segue

$$E[\widehat{\beta}_{\text{Ridge}}(\lambda)] = (\mathbf{I} + \lambda \mathbf{R}^{-1}) \beta \neq \beta.$$

Ridge: um resultado

Pode-se provar que

$$\widehat{\beta}_{\text{Ridge}}(\lambda) = \mathbf{V}\text{diag}\left(\frac{d_1}{d_1^2+\lambda}, \frac{d_2}{d_2^2+\lambda}, \ldots, \frac{d_p}{d_p^2+\lambda}\right)\mathbf{U}^\top \mathbf{y},$$

em que $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ é a decomposição em valores singulares de $\mathbf{X}$, com $\mathbf{U}$ denotando uma matriz ortogonal de dimensão $n \times p$, $\mathbf{V}$ uma matriz ortogonal de dimensão $p \times p$ e $\mathbf{D}$ uma matriz diagonal com dimensão $p \times p$, contendo os correspondentes valores singulares $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ (raízes quadradas dos autovalores de $\mathbf{X}^\top\mathbf{X}$).

Ridge quando $\mathbf{X}$ é ortogonal

Quando $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$, pode-se provar que:

1. Ridge e EMQ:

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \frac{1}{1+\lambda}\hat{\beta}_{\text{MQ}}. \tag{12}$$

2. A escolha ótima de λ minimizando o erro de previsão esperado é

$$\lambda^* = \frac{p\sigma^2}{\sum_{i=1}^{p}\beta_j^2}. \tag{13}$$

Ridge e Lasso: escolha do parâmetro λ

- A escolha do parâmetro de regularização λ pode ser baseada em validação cruzada ou em algum critério de informação.
- Seja $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ uma grade de valores para λ. No segundo caso,

$$\widehat{\lambda} = \arg\min_{\lambda \in \Lambda} [-\log \text{verossimilhança} + \text{penalização}],$$

como

$$\begin{aligned} AIC &= \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{2}{n}, \\ BIC &= \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{\log n}{n}, \\ HQ &= \log[\widehat{\sigma}^2(\lambda)] + \text{gl}(\lambda)\frac{\log\log n}{n}, \end{aligned}$$

em que $\text{gl}(\lambda)$ é o número de graus de liberdade associado a λ, nomeadamente

$$\text{gl}(\lambda) = \text{tr}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\right] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

e

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n - \text{gl}(\lambda)} \sum_{i=1}^{n} [y_i - \widehat{\beta}_{\text{Ridge}}(\lambda)^\top \mathbf{x}_i]^2.$$

Ridge e Lasso: escolha do parâmetro λ

No caso de validação cruzada (VC):

- Calcule o erro da validação cruzada, como descrito abaixo, para cada valor de λ nessa grade.
- Escolha λ para o qual o erro da VC seja mínimo.
- O modelo é re-ajustao usando todas as observações disponíveis e o valor selecionado de λ.
- Pode-se usar o método LOOCV ou KFCV.

Ridge - Consistência

- Quando $\lambda = \lambda_n$ e $p < n$.
- O estimador Ridge pode ser escrito na foma

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda_n) &= (\mathbf{X}^\top\mathbf{X} + \lambda_n\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \boldsymbol{\beta} - \lambda_n(n^{-1}\mathbf{X}^\top\mathbf{X} + \lambda_n\mathbf{I})^{-1}\boldsymbol{\beta} \\ &\quad + (n^{-1}\mathbf{X}^\top\mathbf{X} + \lambda_n\mathbf{I})^{-1}n^{-1}\mathbf{X}\mathbf{e}.\end{aligned}$$

- Quando $\lambda_n \to 0$, para $n \to \infty$, pelo Lema de Slutsky,

$$\begin{aligned}\lambda_n(n^{-1}\mathbf{X}^\top\mathbf{X} + \lambda_n\mathbf{I})^{-1}\boldsymbol{\beta} &\to 0, \\ (n^{-1}\mathbf{X}^\top\mathbf{X} + \lambda_n\mathbf{I})^{-1}n^{-1}\mathbf{X}\mathbf{e} &\to 0.\end{aligned}$$

- Pode ainda ser provado que, quando $\sqrt{n}\lambda_n \to \lambda_0 \geq 0$, quando $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{Ridge}} - \boldsymbol{\beta}) + \lambda_0\mathbf{Q}^{-1}\boldsymbol{\beta} \to N(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}^{-1}),$$

onde $\mathbf{Q} = p\lim n^{-1}\mathbf{X}^\top\mathbf{X}$ e $\mathbf{V}$ é a matriz de variância assintótica de $n^{-1/2}\mathbf{X}\mathbf{e}$.

Lasso: teoria

Considere o caso de $\mathbf{X}$ ortogonal e $p < n$. Então, pode-se provar que

$$\begin{aligned}
\hat{\beta}_{\text{Lasso}}(\lambda) &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_1 \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(-\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}||\boldsymbol{\beta}||_2^2\right) + \lambda||\boldsymbol{\beta}||_1 \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{p} \left(-\hat{\beta}_{MQ,i}\beta_i + \frac{1}{2}\beta_i^2 + \lambda|\beta_i|\right).
\end{aligned}$$

- Como o problema é um programa de otimização quadrática com restrição convexa, a função objetivo torna-se uma soma de funções objetivos.
- Para cada i, minimiza-se $Q_i = -\hat{\beta}_{MQ,i}\beta_i + \frac{1}{2}\beta_i^2 + \lambda|\beta_i|$.

Lasso: teoria

- No modelo RLM, suponha $\mathbf{X}$ fixa ou iid e o erro normal, com média $\mathbf{0}$ e variância $\sigma^2\mathbf{I}$.
- O número de variáveis $p = p_n$ e $p >> n$.
- No caso de $\mathbf{X}$ fixa, resultados de consistência dependem da condição

$$||\boldsymbol{\beta}^0||_1 = ||\boldsymbol{\beta}_n^0||_1 = o\left(\sqrt{\frac{n}{\log p}}\right).$$

- No caso de $\mathbf{X}$ aleatória, sob condições sobre o erro e

$$||\boldsymbol{\beta}^0||_1 = o\left(\left(\frac{n}{\log n}\right)^{1/4}\right),$$

e para λ da ordem de $\sqrt{\log p/n}$,o estimador Lasso é consistente:

$$\left[\hat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) - \boldsymbol{\beta}^0\right] \Sigma \left[\hat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) - \boldsymbol{\beta}^0\right]^\top = o_p(1), \quad n \to \infty,$$

com $\Sigma = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ no caso fixo ou Σ sendo a matriz de covariância de $\mathbf{X}$, no caso aleatório.

Exemplo

Consideremos os dados do arquivo `antracose`, extraídos de um estudo cuja finalidade era avaliar o efeito da idade (`idade`), tempo vivendo em São Paulo (`tmunic`), horas diárias em trânsito (`htransp`), carga tabágica (`cargatabag`), classificação sócio-econômica (`ses`), densidade de tráfego na região onde habitou (`densid`) e distância mínima entre a residência a vias com alta intensidade de tráfego (`distmin`) num índice de antracose (`antracose`) que é uma medida de fuligem (*black carbon*) depositada no pulmão. Como esse índice varia entre 0 e 1, consideramos

$$\textit{logrc} = \log[\text{índice de antracose}/(1 - \text{índice de antracose})]$$

Exemplo - MQ

Os estimadores de mínimos quadrados para um modelo linear podem ser obtidos por meio dos comandos

```
pulmao_lm <- lm(logrc ~ idade + tmunic + htransp + cargatabag +
                        ses + densid + distmin, data=pulmao)
summary(pulmao_lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.977e+00  2.459e-01 -16.169  < 2e-16 ***
idade        2.554e-02  2.979e-03   8.574  < 2e-16 ***
tmunic       2.436e-04  2.191e-03   0.111 0.911485
htransp      7.505e-02  1.634e-02   4.592 5.35e-06 ***
cargatabag   6.464e-03  1.055e-03   6.128 1.61e-09 ***
ses         -4.120e-01  1.238e-01  -3.329 0.000926 ***
densid       7.570e+00  6.349e+00   1.192 0.233582
distmin      3.014e-05  2.396e-04   0.126 0.899950
Residual standard error: 1.014 on 598 degrees of freedom
Multiple R-squared:  0.1965,Adjusted R-squared:  0.1871
F-statistic: 20.89 on 7 and 598 DF,  p-value: < 2.2e-16
```

Exemplo - Ridge

O ajuste dos modelos de regressão *Ridge*, *Lasso* ou *Elastic net* pode ser obtido com o pacote glmnet.

Utilizando esse pacote, ajustamos o modelo de regressão *Ridge* por meio de validação cruzada e obtivemos o gráfico da Figura 2 em que o erro quadrático médio (*MSE*) é expresso em função do logaritmo do coeficiente de regularização λ.

```
regridgecv = cv.glmnet(X, y, alpha = 0)
plot(regridgecv)
```

Exemplo - Ridge

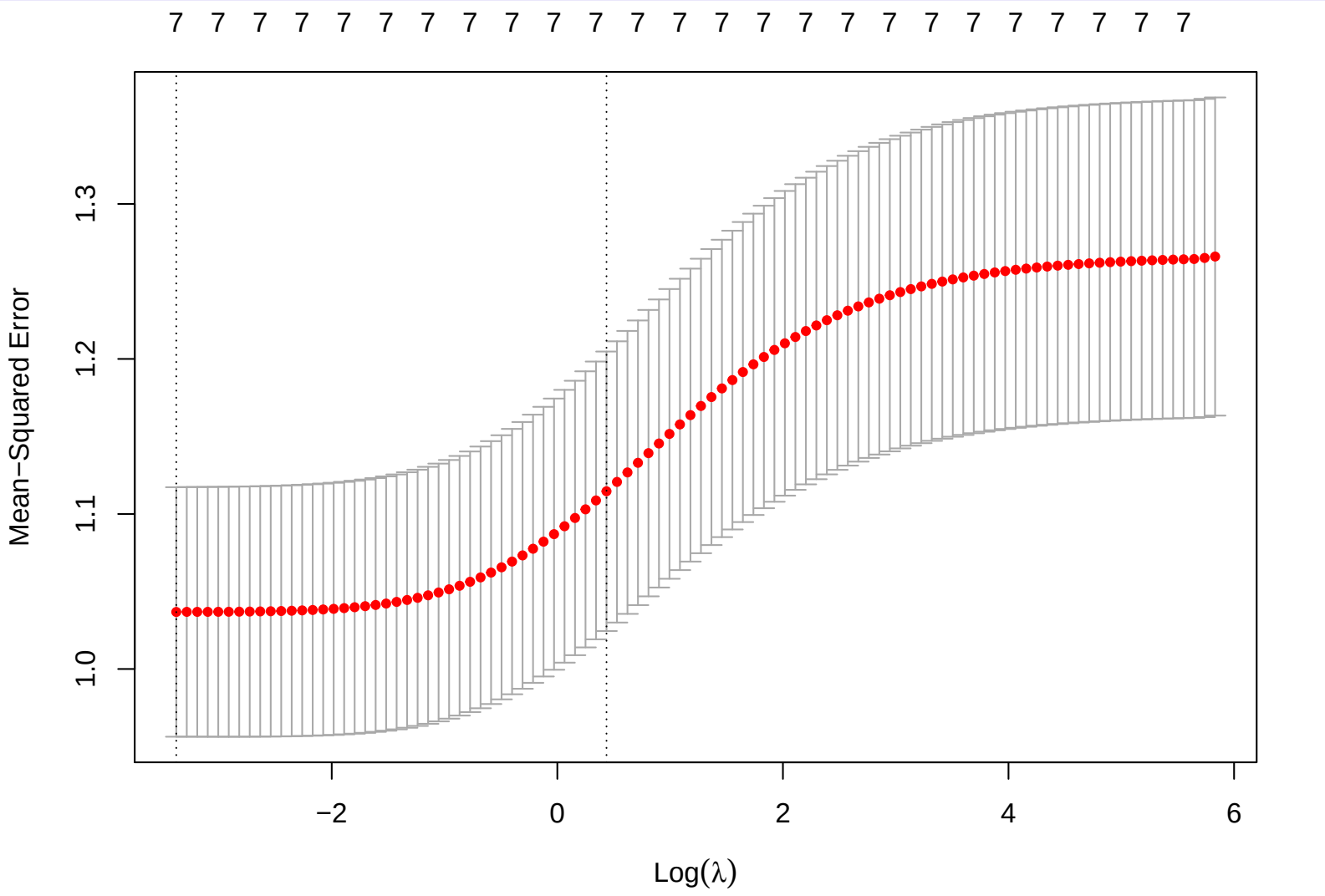

Figure 2: Gráfico para avaliação do efeito do coeficiente de regularização (*Ridge*).

Exemplo - Ridge

Os coeficientes do ajuste correspondentes ao valor mínimo do coeficiente λ, juntamente com esse valor, são obtidos com os comandos

```
coef(regridgecv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                          1
(Intercept) -3.905299e+00
idade        2.456715e-02
tmunic       4.905597e-04
htransp      7.251095e-02
cargatabag   6.265919e-03
ses         -3.953787e-01
densid       7.368120e+00
distmin      3.401372e-05
> regridgecv$lambda.min
[1] 0.03410028
```

Exemplo -Ridge

Com exceção das estimativas dos coeficientes das variáveis `tmunic` e `distmin` as demais foram encolhidas em direção a zero relativamente àquelas obtidas por mínimos quadrados. Os valores preditos e a correspondente raiz quadrada do *MSE* (usualmente denotada *RMSE*) são obtidos por meio de

```
predict(regridgecv, X, s = "lambda.min")
sqrt(regridgecv$cvm[regridgecv$lambda == regridgecv$lambda.min])
[1] 1.050218
```

Exemplo - Lasso

O ajuste do modelo de regressão Lasso juntamente com o gráfico para a escolha do coeficiente λ, disposto na Figura 3, são obtidos com

```
reglassocv = cv.glmnet(X, y, alpha = 1)
plot(reglassocv)
```

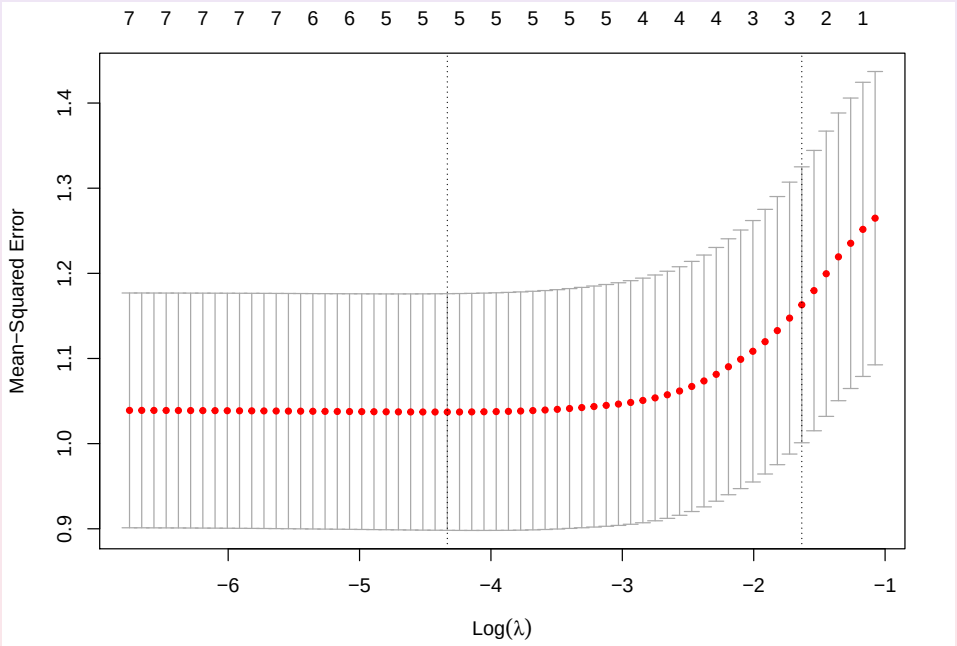

Figura 3: Gráfico para avaliação do efeito do coeficiente de regularização (*Lasso*).

Exemplo - Lasso

Os coeficientes correspondentes à regularização *Lasso*, o valor mínimo do coeficiente λ e o *RMSE* são gerados por intermédio dos comandos

```
coef(reglassocv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -3.820975473
idade        0.024549358
tmunic       .
htransp      0.069750435
cargatabag   0.006177662
ses         -0.365713282
densid       5.166969594
distmin      .
reglassocv$lambda.min
[1] 0.01314064
sqrt(reglassocv$cvm[reglassocv$lambda == reglassocv$lambda.min])
[1] 1.018408
```

Neste caso, todos os coeficientes foram encolhidos em direção ao zero, e aqueles correspondentes às variáveis `tmunic` e `distmin` foram anulados.

Exemplo - Elastic Net

Para o modelo Elastic Net com $\alpha = 0,5$ os resultados são

```
regelncv = cv.glmnet(X, y, alpha = 0.5)
regelncv$lambda.min
[1] 0.02884367
coef(regelncv, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) -3.776354935
idade        0.024089256
tmunic       .
htransp      0.068289153
cargatabag   0.006070319
ses         -0.354080190
densid       4.889074555
distmin      .           .
sqrt(regelncv$cvm[regelncv$lambda == regelncv$lambda.min])
[1] 1.0183
```

Vemos que os 3 procedimentos resultam em EQM similares, com pequena vantagem para Elastic Net.

MAG

- Modelos lineares têm um papel muito importante na análise de dados, tanto pela facilidade de ajuste quanto de interpretação. De uma forma geral, os modelos lineares podem ser expressos como

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) + e_i \tag{14}$$

$i = 1, \ldots, n$ em que as funções f_i são conhecidas. No modelo de regressão polinomial de segundo grau, por exemplo, $f_1(x_{i1}) = x_{i1}$ e $f_2(x_{i2}) = x_{ij}^2$. Em casos mais gerais, poderíamos ter $f_1(x_{i1}) = x_{ij}$ e $f_2(x_{i2}) = \exp(x_{i2})$. Em muitos problemas reais, no entanto, nem sempre é fácil especificar a forma das funções f_i e uma alternativa proposta por Hastie e Tibshirani (1996) são os chamados Modelos Aditivos Generalizados (*Generalized Additive Modelos* - GAM) que são expressos como (14) sem a especificação da forma das funções f_i.

- Quando a distribuição da variável resposta y_i pertence à **família exponencial**, o modelo pode ser considerado como uma extensão dos **Modelos Lineares Generalizados** (*Generalized Linear Models* - GLM) e é expresso como

$$g(\mu_i) = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) \tag{15}$$

em que g é uma **função de ligação** e $\mu_i = E(y_i)$ (ver Nota de Capítulo 3).

MAG

- Existem diversas propostas para a representação das funções f_i que incluem o uso de splines naturais, splines suavizadas e regressões locais.
- A suavidade dessas funções é controlada por parâmetros de suavização, que devem ser determinados *a priori*. Curvas muito suaves podem ser muito restritivas, enquanto curvas muito rugosas podem causar sobreajuste.
- O procedimento de ajuste dos modelos aditivos generalizados depende da forma escolhida para as funções f_i. A utilização de *splines* naturais, por exemplo, permite a aplicação direta do método de mínimos quadrados, graças à sua construção a partir de funções base.
- Para *splines* penalizadas, o processo de estimação envolve algoritmos um pouco mais complexos, como aqueles conhecidos sob a denominação de retroajustamento (*backfitting*). Para detalhes sobre o ajuste dos modelos aditivos generalizados, consulte Hastie e Tibshirani (1990) e Hastie et al. (2008).

Splines

- Para entender o conceito de *splines*, consideremos o seguinte modelo linear com apenas uma variável explicativa

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n. \tag{16}$$

- A ideia subjacente aos modelos aditivos generalizados é a utilização de funções base e consiste na substituição do termo $\beta_1 x_i$ em (16) por um conjunto de transformações conhecidas $b_1(x_i), \ldots, b_t(x_i)$, gerando o modelo

$$y_i = \alpha_0 + \alpha_1 b_1(x_i) + \ldots + \alpha_t b_t(x_i) + e_i, \quad i = 1, \ldots, n. \tag{17}$$

  O modelo de regressão polinomial de grau t é um caso particular de (17) com $b_j(x_i) = x_i^j$, $j = 1, \ldots, t$.

Splines

- Uma proposta para aumentar a flexibilidade da curva ajustada consiste em segmentar o domínio da variável preditora e ajustar diferentes polinômios de grau d aos dados de cada um dos intervalos gerados pela segmentação. Cada ponto de segmentação é chamado de nó e uma segmentação com k nós gera $k+1$ polinômios. Na Figura 43, apresentamos um exemplo com polinômios de terceiro grau e 4 nós.
- Nesse exemplo, a expressão (17) tem a forma

$$y_i = \begin{cases} \alpha_{01} + \alpha_{11}x_i + \alpha_{21}x_i^2 + \alpha_{31}x_i^3 + e_i, & \text{se } x_i \leq -0.5, \\ \alpha_{02} + \alpha_{12}x_i + \alpha_{22}x_i^2 + \alpha_{32}x_i^3 + e_i, & \text{se } -0.5 < x_i \leq 0, \\ \alpha_{02} + \alpha_{13}x_i + \alpha_{23}x_i^2 + \alpha_{33}x_i^3 + e_i, & \text{se } 0 < x_i \leq 0.5, \\ \alpha_{02} + \alpha_{14}x_i + \alpha_{24}x_i^2 + \alpha_{34}x_i^3 + e_i, & \text{se } 0.5 < x_i \leq 1, \\ \alpha_{05} + \alpha_{15}x_i + \alpha_{25}x_i^2 + \alpha_{35}x_i^3 + e_i, & \text{se } x_i > 1, \end{cases} \quad (18)$$

sendo que nesse caso, as funções base $b_1(X), b_2(X), ..., b_k(X)$ são construídas com a ajuda de funções indicadoras. Esse modelo é conhecido como modelo polinomial cúbico segmentado.

Splines

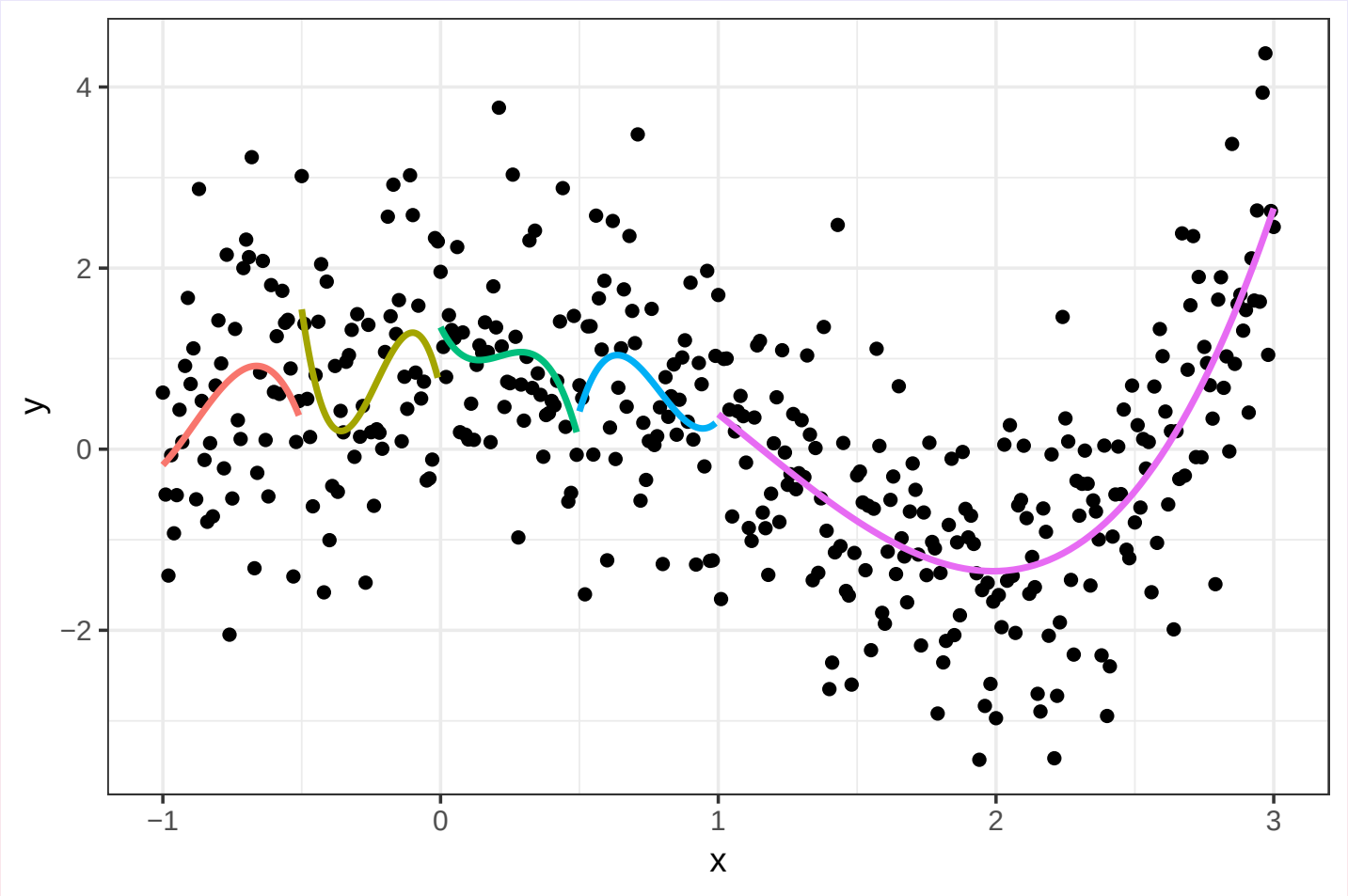

Figura 4: Polinômios de terceiro grau ajustados aos dados de cada região segmentada da variável X. Os nós são os pontos $x = -0.5$, $x = 0$, $x = 0.5$ e $x = 1$.

Splines

- A curva formada pela junção de cada um dos polinômios na Figura 43 não é contínua, apresentando "saltos" nos nós.
  Essa característica não é desejável, pois essas descontinuidades não são interpretáveis. Para contornar esse problema, podemos definir um *spline* de grau d como um polinômio segmentado de grau d com as $d-1$ primeiras derivadas contínuas em cada nó. Essa restrição garante a continuidade e suavidade (ausência de vértices) da curva obtida.
- Utilizando a representação por bases (17), um *spline* cúbico com k nós pode ser expresso como

$$y_i = \alpha_0 + \alpha_1 b_1(x_i) + \alpha_2 b_2(x_i) + \ldots + \alpha_{k+3} b_{k+3}(x_i) + e_i, \quad i = 1, ..., n, \quad (19)$$

  com as funções $b_1(x), b_2(x), \ldots, b_{k+3}(x)$ escolhidas apropriadamente.
- Usualmente, essas funções envolvem três termos polinomiais, a saber, x, x^2 e x^3 e k termos $h(x, c_1), \ldots, h(x, c_k)$ da forma

$$h(x, c_j) = (x - c_j)_+^3 = \begin{cases} (x - c_j)^3, & \text{se } x < c_j, \\ 0, & \text{em caso contrário,} \end{cases}$$

  com $c_1, \ldots, c_k$ indicando os nós.

Splines

- Com a inclusão do termo α_0, o ajuste de um *spline* cúbico com k nós envolve a estimação de $k + 4$ parâmetros e, portanto, utiliza $k + 4$ graus de liberdade. Mais detalhes sobre a construção desses modelos podem ser encontrados em Hastie (2008) e James et al. (2017).
- Além das restrições sobre as derivadas, podemos adicionar restrições de fronteira, exigindo que a função seja linear na região de x abaixo do menor nó e acima do maior nó. Essas restrições diminuem a variância dos valores extremos gerados pelo preditor, produzindo estimativas mais estáveis. Um *spline* cúbico com restrições de fronteira é chamado de spline natural.
- No ajuste de *splines* cúbicos ou naturais, o número de nós determina o grau de suavidade da curva e a sua escolha pode ser feita por validação cruzada conforme indicado em James et al. (2017). De uma forma geral, a maior parte dos nós é posicionada nas regiões do preditor com mais informação, isto é, com mais observações. Por pragmatismo, para modelos com mais de uma variável explicativa, costuma-se adotar o mesmo número de nós para todos os preditores.

Splines

- Os *splines* suavizados constituem uma classe de funções suavizadoras que não utilizam a abordagem por funções bases. De maneira resumida, um *spline* suavizado é uma função f que minimiza

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int f''(u)^2 du \tag{20}$$

  em que f'' corresponde à segunda derivada da função f e indica sua curvatura; quanto maior for a curvatura maior a penalização.
- O primeiro termo dessa expressão garante que f se ajustará bem aos dados, enquanto o segundo penaliza a sua variabilidade, isto é, controla a suavidade de f, que é regulada pelo parâmetro λ. A função f se torna mais suave conforme λ aumenta. A escolha desse parâmetro é geralmente feita por validação cruzada.
- Outro método bastante utilizado no ajuste funções não lineares para a relação entre a variável preditora X e a variável resposta Y é conhecido como regressão local. Esse método consiste em ajustar modelos de regressão simples em regiões em torno de cada observação x_0 da variável preditora X.

Splines

- Essas regiões são formadas pelos k pontos mais próximos de x_0, sendo que o parâmetro $s = k/n$ determina o quão suave ou rugosa será a curva ajustada. O ajuste é feito por meio de mínimos quadrados ponderados, com pesos inversamente proporcionais à distância entre cada ponto da região centrada em x_0 e x_0. Aos pontos dessas regiões mais afastados de x_0 são atribuídos pesos menores.
- Lowess: ajusta retas. Veja a Nota de Capítulo 5.2.
- Para uma excelente exposição sobre *splines* e penalização o leitor pode consultar Eilers e Marx (1996) e Eilers e Marx (2021).
- Modelos aditivos generalizados podem ser ajustados utilizando-se a função gam() do pacote mgcv. Essa função permite a utilização de *splines* como funções suavizadoras. Para regressão local, é necessário usar a função `gam()` do pacote `gam`. Também é possível utilizar o pacote `caret`, a partir da função `train()` e `method = "gam"`.

MAG: exemplo

Consideremos os dados do arquivo `esforco` com o objetivo de prever os valores da variável `vo2fcpico` (VO2/FC no pico do exercício) a partir das variáveis `NYHA`, `idade`, `altura`, `peso`, `fcrep` (frequência cardíca em repouso) e `vo2rep` (VO2 em repouso). Um modelo inicial de regressão linear múltipla também pode ser ajustado por meio dos seguintes comandos da função `gam`

```
mod0 <- gam(vo2fcpico ~ NYHA + idade + altura + peso + fcrep
                        + vo2rep, data=esforco)
```

Como não especificamos nem a distribuição da resposta, nem a função de ligação, a função `gam` utiliza a distribuição normal com função de ligação logarítmica, conforme indica o resultado.

MAG: exemplo

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + idade + altura + peso + fcrep + vo2rep
Parametric coefficients:
            Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -4.80229    4.43061  -1.084  0.280642
NYHA1       -0.45757    0.50032  -0.915  0.362303
NYHA2       -1.78625    0.52629  -3.394  0.000941 ***
NYHA3       -2.64609    0.56128  -4.714  6.75e-06 ***
NYHA4       -2.43352    0.70532  -3.450  0.000780 ***
idade       -0.05670    0.01515  -3.742  0.000284 ***
altura       0.09794    0.02654   3.690  0.000342 ***
peso         0.08614    0.01739   4.953  2.48e-06 ***
fcrep       -0.07096    0.01318  -5.382  3.84e-07 ***
vo2rep       0.35564    0.24606   1.445  0.151033
---
R-sq.(adj) =  0.607   Deviance explained = 63.5%
GCV =  4.075  Scale est. = 3.7542    n = 127
```

MAG: exemplo

Para avaliar a qualidade do ajuste, produzimos gráficos de dispersão entre os resíduos do ajuste e cada uma das variáveis preditoras. Esses gráficos estão dispostos na Figura 5 e sugerem relações possivelmente não lineares pelo menos em alguns casos.

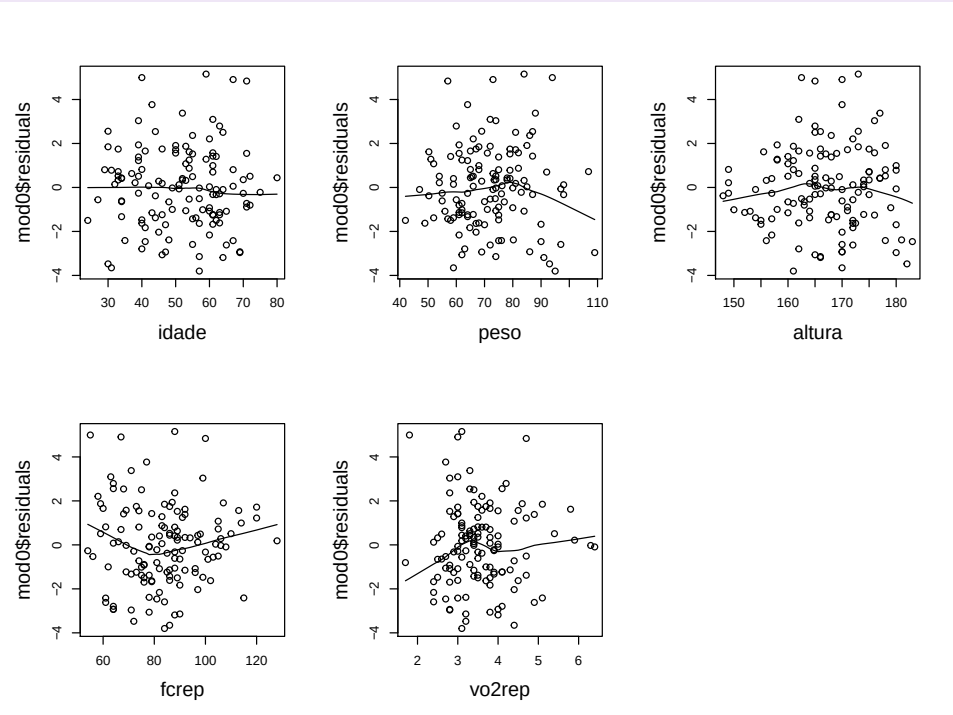

Figura 5: Gráficos de dispersão (com curva *lowess*) entre `vo2fcpico` e cada variável preditora contínua considerada no Exemplo.

MAG: exemplo

Uma alternativa é considerar modelos GAM do tipo (14) em que as funções f_i são expressas em termos de *splines*. Em particular, um modelo GAM com *splines* cúbicos para todas as variáveis explicativas contínuas pode ser ajustado por meio do comando

```
mod1 <- gam(vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                        s(fcrep) + s(vo2rep), data=esforco)
```

gerando os seguintes resultados:

MAG: exemplo

```
Family: gaussian
Link function: identity
Formula:
vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) + s(fcrep) +
    s(vo2rep)
Parametric coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  10.2101     0.3207   31.841   < 2e-16 ***
NYHA1        -0.5498     0.4987   -1.103   0.272614
NYHA2        -1.8513     0.5181   -3.573   0.000522 ***
NYHA3        -2.8420     0.5664   -5.018   1.99e-06 ***
NYHA4        -2.5616     0.7031   -3.643   0.000410 ***
---
Approximate significance of smooth terms:
            edf Ref.df    F    p-value
s(idade)  1.000  1.000 15.860  0.00012 ***
s(altura) 5.362  6.476  3.751  0.00142 **
s(peso)   1.000  1.000 22.364  6.32e-06 ***
s(fcrep)  1.742  2.185 16.236  3.95e-07 ***
s(vo2rep) 1.344  1.615  0.906  0.47319
---
R-sq.(adj) =   0.64   Deviance explained = 68.2%
GCV = 3.9107  Scale est. = 3.435     n = 127
```

MAG: exemplo

- O painel superior contém estimativas dos componentes paramétricos do modelo e o painel inferior, os resultados referentes aos termos suavizados. Neste caso apenas a variável categorizada `NYHA` não foi suavizada, dada sua natureza não paramétrica.
- A coluna rotulada `edf` contém os graus de liberdade efetivos associados a cada variável preditora. Para cada variável preditora contínua não suavizada, perde-se um grau de liberdade; para as variáveis suavizadas a atribuição de graus de liberdade é mais complexa em virtude do número de funções base e do número de nós utilizados no processo de suavização. A linha rotulada `GCV` (*Generalized Cross Validation*) está associada com a escolha (por validação cruzada) do parâmetro de suavização.
- A suavização é irrelevante apenas para a variável `vo2rep` e dado que ela também não apresentou contribuição significativa no modelo de regressão linear múltipla, pode-se considerar um novo modelo GAM obtido com a sua eliminação. Os resultados correspondentes, apresentados a seguir, sugerem que todas as variáveis preditoras contribuem significativamente para explicar sua relação com a variável resposta,

MAG: exemplo

```
Formula:
vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) + s(fcrep)
Parametric coefficients:
            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  10.2301     0.3202  31.948   < 2e-16 ***
NYHA1        -0.5818     0.4985  -1.167   0.245650
NYHA2        -1.8385     0.5161  -3.563   0.000539 ***
NYHA3        -2.9669     0.5512  -5.382   4.04e-07 ***
NYHA4        -2.4823     0.6980  -3.556   0.000551 ***
---
Approximate significance of smooth terms:
            edf  Ref.df     F     p-value
s(idade)  1.000  1.000   16.322  9.59e-05 ***
s(altura) 5.311  6.426    3.857  0.00115 **
s(peso)   1.000  1.000   22.257  6.56e-06 ***
s(fcrep)  1.856  2.337   14.865  8.39e-07 ***
---
R-sq.(adj) =   0.64   Deviance explained = 67.8%
GCV = 3.8663  Scale est. = 3.435     n = 127
```

MAG: exemplo

- Como o número efetivo de graus de liberdade para `idade` e `peso` é igual a 1, elas se comportam de forma linear no modelo. Os gráficos dispostos na Figura 6, produzidos por meio do comando `plot(mod2, se=TRUE)` evidenciam esse fato; além disso mostram a natureza "mais não linear" da variável `altura` (com `edf`$= 5.311$).
- Uma avaliação da qualidade do ajuste pode ser realizada por meio de uma análise de resíduos e de comparação dos valores observados e preditos. Para essa finalidade, o comando `gam.check(mod2)` gera os gráficos apresentados na Figura 7 que não evidenciam problemas no ajuste.
- Além disso, é possível comparar os modelos por meio de uma **análise de desviância**, que pode ser obtida com o comando `anova(mod0, mod1, mod2, test= "F")`.

MAG: exemplo

```
Analysis of Deviance Table
Model 1: vo2fcpico ~ NYHA + idade + altura + peso + fcrep +
                     vo2rep
Model 2: vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                     s(fcrep) +  s(vo2rep)
Model 3: vo2fcpico ~ NYHA + s(idade) + s(altura) + s(peso) +
                     s(fcrep)
     Resid.    Df  Resid.   Dev    Df Deviance      F         Pr(>F)
1    117.00     439.24
2    109.72     383.18    7.2766     56.052       2.2425     0.03404 *
3    111.24     387.58   -1.5129     -4.399       0.8465     0.40336
```

MAG: exemplo

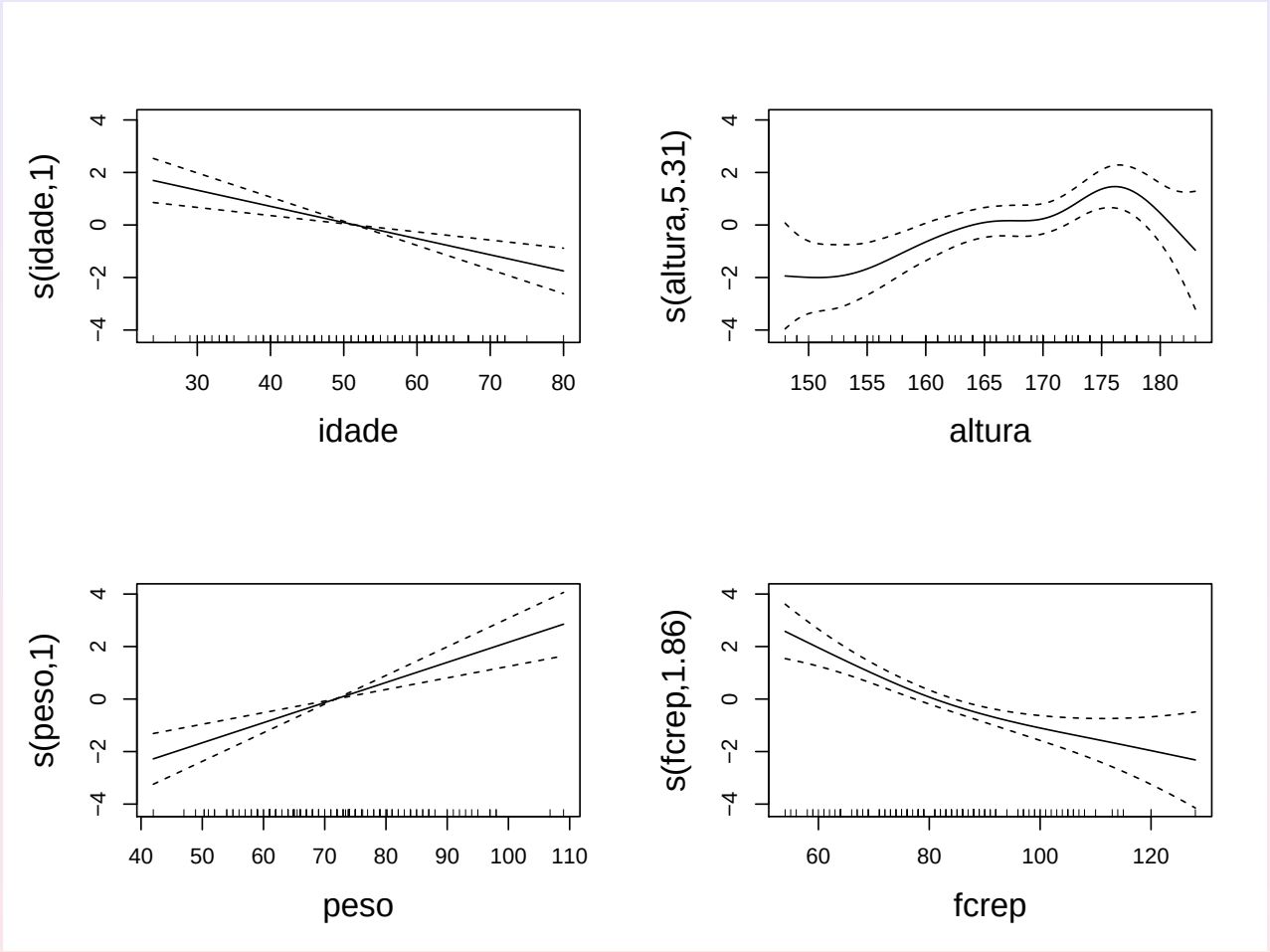

Figura 6: Funções suavizadas (com bandas de confiança) obtidas por meio do modelo GAM para os dados do Exemplo.

MAG: exemplo

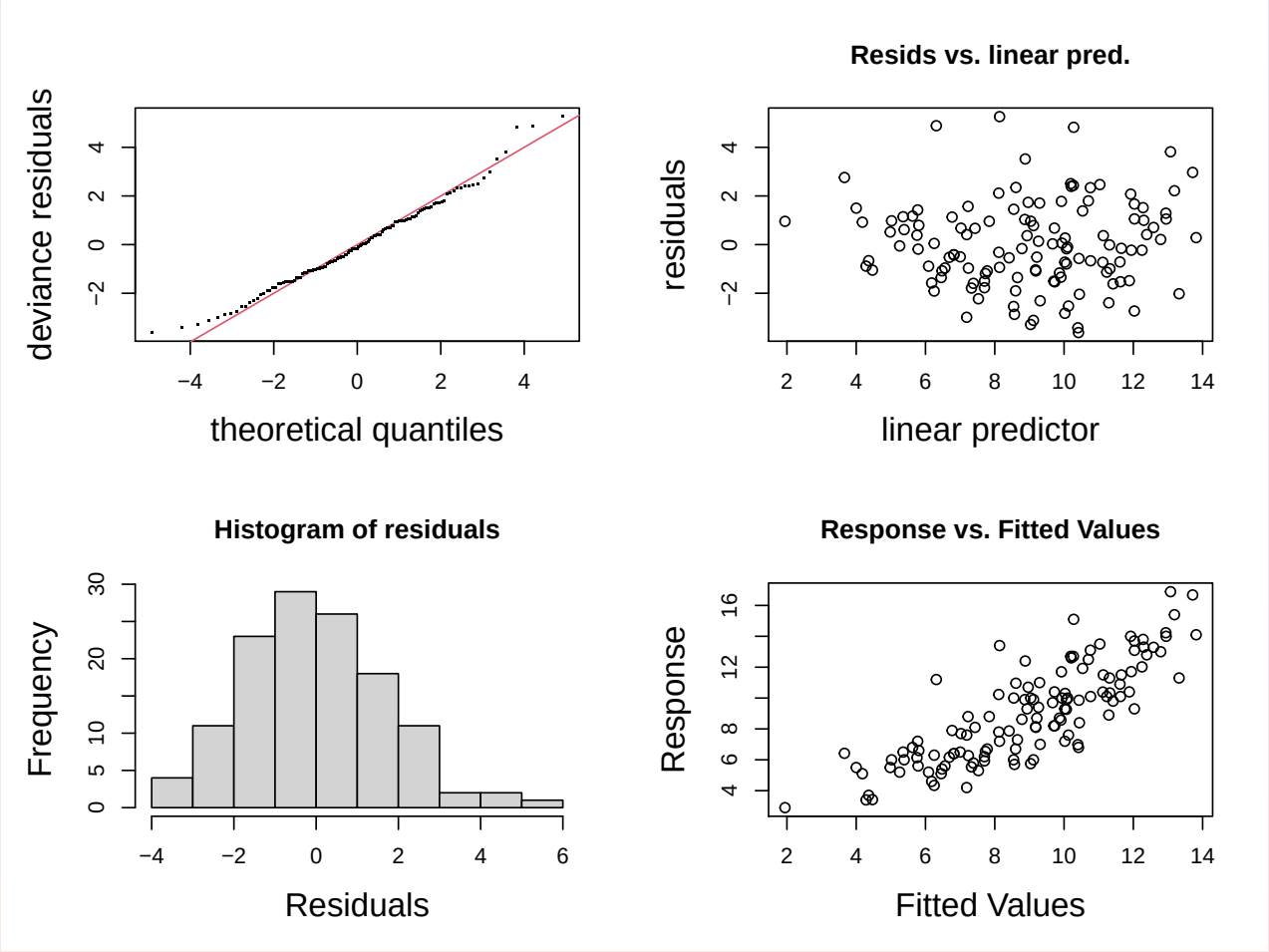

Figura 7: Gráficos diagnósticos para o ajuste do modelo GAM aos dados do Exemplo.

MAG: exemplo

- Esses resultados mostram que ambos os modelos GAM são essencialmente equivalentes ($p = 0.403$) mas significativamente mais adequados ($p = 0.034$) que o modelo de regressão linear múltipla.
- A previsão para um novo conjunto dados em que apenas os valores das variáveis preditoras estão disponíveis pode ser obtida por meio do comando `predict(mod2, newdata=esforcoprev, se=TRUE, type="response")`. Consideremos, por exemplo, o seguinte conjunto com dados de 5 novos pacientes

idade	altura	peso	NYHA	fcrep	vo2rep
66	159	50	2	86	3,4
70	171	77	4	108	4,8
64	167	56	2	91	2,5
42	150	67	2	70	3,0
54	175	89	2	91	2,9

MAG: exemplo

O resultado da previsão com o modelo adotado é

```
$fit
     1          2          3          4          5
 4.632615   5.945157   5.928703   7.577097 10.273719
$se.fit
     1          2          3          4          5
0.6747203 0.7155702 0.6255449 0.7731991 0.5660150
```

Referências

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Berlin: Springer.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparcity*. Chapman and Hall.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2021). *Estatística e Ciência de Dados*. Texto Preliminar, IME-USP.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B (methodological), **58**, 267–288.