O uso de Splines em Regressão Não Paramétrica

Ronaldo Dias

Universidade Estadual de Campinas *

1 Motivação

Suponha que observações são coletadas de uma variável contínua Y em n valores da variável independente t. Sejam (t_j, y_j), $j = 1, \ldots, n$, tal que o seguinte modelo de regressão pode ser proposto:

$$y_j = f(t_j) + \varepsilon_j, \quad j = 1, \ldots, n$$

onde as variáveis aleatórias ε_j tem média zero, são não correlacionadas, têm variância comum σ^2. Mas ainda, $f(t_j)$ são valores obtidos de alguma função f, desconhecida, calculada nos pontos $t_1, \ldots, t_n$. A função f é geralmente chamada de *função de regressão* ou *curva de regressão*.

Para motivar o conceito de suavização por splines, vamos assumir que: $f \in W_2^m[a,b] = \{f : f' \text{ é abs. cont. e } \int_a^b (f^{(m)})^2 < \infty\}$. Desejamos uma estimativa de f que se ajuste bem aos dados e que ao mesmo tempo tenha um certo grau de suavidade. Uma medida de bondade de ajuste aos dados é dada por $\frac{1}{n}\sum_{i=1}^n (y_i - f(t_i))^2$ enquanto que uma medida de suavidade para $f \in W_2^m[a,b]$ é $\int_a^b (f^{(m)})^2$. Assim, combinando os dois critérios temos:

$$A_\alpha(f) = (1-\alpha)\sum_{i=1}^n (y_i - f(t_i))^2 + \alpha \int_a^b (f^{(m)})^2$$

*Departamento de Estatística, IMECC, Cidade Universitária Zeferino Vaz, Caixa Postal 6065, 13.081-970 - Campinas, SP - BRAZIL.

com $\alpha \in [0,1]$

Note que podemos então encontrar um estimador ótimo minimizando o funcional $A_\alpha(f)$ com $f \in W_2^m[a,b]$. Porém, pondo $\lambda = \alpha/(1-\alpha)$, temos um problema equivalente de otimização onde obteremos uma solução f_α que minimiza

$$A_\lambda(f) = \sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_a^b (f^{(m)})^2$$

para $\lambda > 0$

A solução f_λ é estimador de suavização por splines da função de regressão. O parâmetro λ é o parâmetro de suavização, pois é ele quem governa o balanço entre bondade de ajuste e suavização. Quando λ é grande maior peso é dado ao critério de suavização, enquanto que para λ pequeno, próximo de zero, maior ênfase é dada à bondade de ajuste.

Suavização por splines teve sua origem com Whittaker (1923) cujos métodos motivaram Shoemberg(1964) a obter um estimador de suavização por splines e mais Reinch (1967) obteve uma derivação com a qual foi criado um algorítmo baseado em B-splines. Aparentemente, muito do que é conhecido sobre splines foi devido a analistas númericos até que um extenso trabalho desenvolvido por Grace Wahba mostrou as propriedades estatísticas deste procedimento e sua utilidade nos modelos não paramétricos de análise de regressão.

2 "Splines"

Um spline de ordem m com knots em $\xi_1, \ldots, \xi_k$ é qualquer função da forma

$$s(t) = \sum_{i=0}^{m-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i (t - \xi_i)_+^{m-1},$$

onde $(x)_+ = \max(0,x)$ e os coeficientes $\theta_0, \ldots, \theta_{m-1}$, $\delta_1, \ldots, \delta_k$ são números reais.

Da definição acima notamos que um spline satisfaz:

1. s é um polinômio por partes de ordem m em qualquer subintervalo $[\xi_i, \xi_{i+1})$.

2. s tem $m-2$ derivadas contínuas.

3. s tem a $(m-1)$-ésima derivada e é uma função escada com saltos em $\xi_1, \ldots, \xi_k$.

Assim, um spline é um polinômio por partes cujos diferentes segmentos são unidos nos "knots" $\xi_1, \ldots, \xi_k$ no sentido de garantir continuidade.

Seja $S^m(\xi_1, \ldots, \xi_k)$ o conjunto formado por todos os splines como definidos anteriormente. Então $S^m(\xi_1, \ldots, \xi_k)$ é um espaço vetorial cuja dimensão é $m+k$, desde que as funções $1, t, \ldots, t^{m-1}, (t-\xi_1)_+^{m-1}, \ldots, (t-\xi_k)_+^{m-1}$ sejam linearmente independentes.

De particular importância temos os splines naturais de ordem $m = 2p$ e $k = n$, isto é, um spline natural é um spline de ordem $2p$ com knots em $a \leq t_1, \ldots, t_n \leq b$. O nome spline natural vem do fato das chamadas condições naturais de fronteira, $s^{(j)}(a) = s^{(j)}(b) = 0$, $j = m, \ldots, 2m-1$. Note que temos mais uma condição que um spline devem satisfazer: 4. s é um polinômio de ordem p fora do intervalo $[t_1, t_n]$.

Seja $NS^{2p}(t_1, \ldots, t_n)$ o conjunto de todos os splines naturais de ordem $2p$ com knots $a \leq t_1, \ldots, t_n \leq b$. Então este conjunto de funções forma um subespaço do espaço vetorial $S^{2p}(t_1, \ldots, t_n)$ obtido por inserir $2p$ restrições lineares que provêm da propriedade 4. Assim, para que um spline seja um spline natural devemos ter $\theta_p = \ldots, \theta_{2p-1} = 0$ na definição de spline, desde que s seja um polinômio de ordem p para $t \notin [t_1, t_n]$. Pode-se verificar que a dimensão de $NS^{2p}(t_1, \ldots, t_n)$ é n.

3 Splines e Regressão Polinomial

Considere o modelo de regressão:

$$y_j = f(t_j) + \varepsilon_j, \quad j = 1, \ldots, n,$$

onde $a \leq t_1, \ldots, \leq t_n \leq b$ e o vetor ε do erros não correlacionados têm média zero e variância comum σ^2. Para motivar o conceito de suavização por splines, suponha que $f \in W_2^m[a, b]$. Queremos encontrar um estimador para a curva de regressão f que ajuste bem os dados e que ao mesmo tempo tenha um certo grau de suavidade.

Regressão polinomial está bastante relacionada com o procedimento descrito acima. Para se ver isto, assuma que o estimador de f dado por f_λ seja bem definido. Desde que $f \in W_2^m[a, b]$, segue que existem constantes, $\theta_0, \ldots, \theta_{m-1}$ tais que

$$f(t) = \sum_{i=0}^{m-1} \theta_i t^i + \text{resto}(t),$$

com

$$\text{resto}(t) = \frac{1}{(m-1)!} \int_a^b f^{(m)}(x)(t-x)_+^{m-1} dx.$$

Assim, o modelo $y_j = f(t_j) + \varepsilon_j$ é equivalente ao modelo

$$y_j = \sum_{i=0}^{m-1} \theta_i t^i + \text{resto}(t) + \varepsilon_j \ , j = 1, \ldots, n.$$

Se o resto(t_j) é proximo de zero, então parece razoável estimar f usando um estimador de regressão polinomial de ordem m. Caso contrário, há dúvidas sobre a aplicabilidade deste modelo polinomial. Veremos agora como o critério acima está relacionado com o modelo de regressão polinomial. Seja o funcional $J_m(f) = \int_a^b (f^{(m)}(t))^2 dt$, então aplicando a desigualdade de Cauchy-Schwartz no resto(t) temos:

$$\text{resto}(t)^2 \leq \frac{(b-a)^{2m-1}}{[(2m-1)(m-1)!]^2} J_m(f).$$

Note que existe uma constante que depende de m mas não de f de tal modo que

$\sup_{t\in[a,b]} |\text{resto}(t)| \leq CJ_m(f)^{1/2}$. Em particular, $|\text{resto}(t_j)| \leq CJ_m(f)^{1/2}$, $j = 1,\ldots,n$. Desta forma, o tamanho de $J_m(f)$ está ligado aos termos $\text{resto}(t_1),\ldots,\text{resto}(t_n)$, pois, se $J_m(f)$ é pequeno, isto é, $J_m(f) \leq \rho$, tal que $\rho > 0$, teremos grande credibilidade no modelo polinomial. Por exemplo, seja $m = 2$, daí $J_2(f) \equiv 0$ se e somente se f é linear. A condição $J_2(f) \leq \rho$ governa o quão longe f pode se afastar do modelo linear. Assim se a priori sabe-se que ρ é pequeno, então isto indicará que f é aproximadamente linear. Observe que $\dot{f}$ representa a taxa de variação da inclinação da curva de regressão.

Podemos incluir a condição acima para determinar um procedimento de estimação. Por exemplo, encontre $f \in W_2^m[a,b]$ que minimize $EQM(f) = \frac{1}{n}\sum_{j=1}^n (y_j - f(t_j))^2$ com a condição que $J_m(f) \leq \rho$. Usando multiplicadores de Lagrange, precisamos encontrar $f_\lambda \in W_2^m[a,b]$ tal que $EQM(f) + \lambda(J_m(f) - \rho)$ seja mínimo. Note que este critério é em essência o mesmo que foi definido anteriormente.

Pode-se mostrar que se $\phi_1,\ldots,\phi_n$ formam uma base para o espaço $NS^{2m}(t_1,\ldots,t_n)$, com $n \leq m$, e $m \in I\!N$, então exite para $0 < \lambda < \infty$ fixo, um único estimador f_λ sob o critério de mínimos quadrados penalizados quando $f \in W_2^m[a,b]$.

Mais ainda, $f_\lambda \in NS^{2m}(t_1,\ldots,t_n)$ e consequentemente este estimador pode ser escrito com uma combinação linear de $\phi_1,\ldots,\phi_n$. Isto é, $f_\lambda = \sum_{j=1}^n \theta_{\lambda j}\phi_j(t)$ tal que o vetor $\theta_{\lambda j} = (\theta_{\lambda 1},\ldots,\theta_{\lambda n})^t$ satisfaz a equação

$$(X^tX + n\lambda\Omega)\theta_\lambda = X^ty,$$

onde a matriz Ω possui entradas $\Omega_{i,j} = \int_a^b \ddot{\phi}_i(t)\ddot{\phi}_j(t)dt$, com $i,j = 1,\ldots,n$.

A escolha comum para as bases $\phi_1,\ldots,\phi_n$ é a sequência formada pelos bem conhecidos B-splines (De Boor (1978)).

4 A escolha de um base

Vimos que f_λ pode ser escrita como uma combinação linear de funções que formam um base. Obviamente existem várias maneiras de construir uma base apropriada para obtenção de de f_λ. Entretanto, vamos dar maior ênfase naquela é mais fácil de ser computada e consequentemente a mais utilizada nos programas de computadores.

A escolha comum para uma base $\phi_1, \ldots, \phi_n$ é a sequência formada pelos B-splines (de Boor, 1978).

$$N_{k,m}(x) = \sum_{j=0}^{m} \alpha_{j,m}^{[k]} (t_{k+m} - x)_+^{m-1},$$

onde $(t-x)_+^{m-1} = \max(0, (t-x)^{m-1})$ e os α's são os coeficientes das diferenças divididas avaliadas nos pontos t_i. Os $N_{k,m}$ são os bem conhecidos B-splines. Figura 4.1 exibe um exemplo da forma dessas bases.

Seja, $Y = (Y_1, \ldots, Y_n)'$, $\theta = (\theta_1, \ldots, \theta_n)'$, $X_{i,j} = N_{j,2}(t_i)$ e $\Omega_{i,j} = \int \ddot{N}_{i,2} \ddot{N}_{j,2}$ para $i, j = 1, \ldots, n$. Então o problema de minimização anterior passa a ser de dimensão finita e é dado por:

$$||Y - X\theta||^2 + \frac{\lambda}{2} \theta' \Omega \theta,$$

com solução única
$\theta_\lambda = (X'X + \lambda\Omega)^{-1} X'Y$.

5 Parâmetro de Suavização

Note que a solução apresentada na seção anterior θ_λ depende do parâmetro de suavização λ. Podemos observar o efeito deste parâmetro na figura 5.1 e da necessidade de obtermos bons métodos para estimá-lo.

A motivação básica para validação cruzada vem em termos de predição. Assumindo que os erros têm média zero. A verdadeira curva de regressão tem

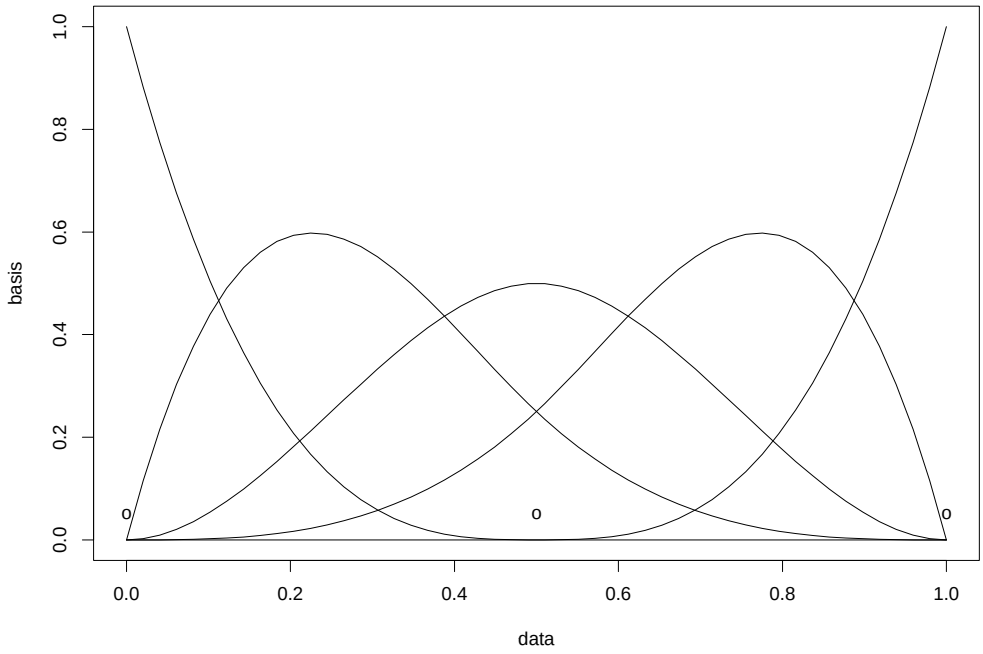

Figure 4.1: B-splines com 3 knots

a propriedade de que, se uma observação Y é feita num ponto t, o valor de $f(t)$ é o melhor preditor de Y em termos do erro médio quadrático. Assim, uma boa escolha para o estimador $\hat{f}$ seria aquele que tivesse o menor valor de $(Y - \hat{f}(t))^2$ para uma nova observação Y no ponto t.

Na prática, quando aplicamos métodos de suavização não existe nova observação disponível. A técnica de validação cruzada produz a situação de uma nova observação. Para λ fixo, considere a observação Y_i no ponto t_i como sendo uma nova observação. por omiti-la do conjunto de dados que vai ser utilizado na estimação de f. Denote $f_\lambda^{-i}(t)$ a solução por

$$\sum_{j \neq i} (Y_i - f(t_j))^2 + \lambda \int (\ddot{f})^2.$$

A qualidade de $f_\lambda^{(-i)}$ como preditor da nova observação pode ser julgada por como o valor de $f_\lambda^{(-i)}(t_i)$ prediz Y_i. Desde que a escolha de qual observação

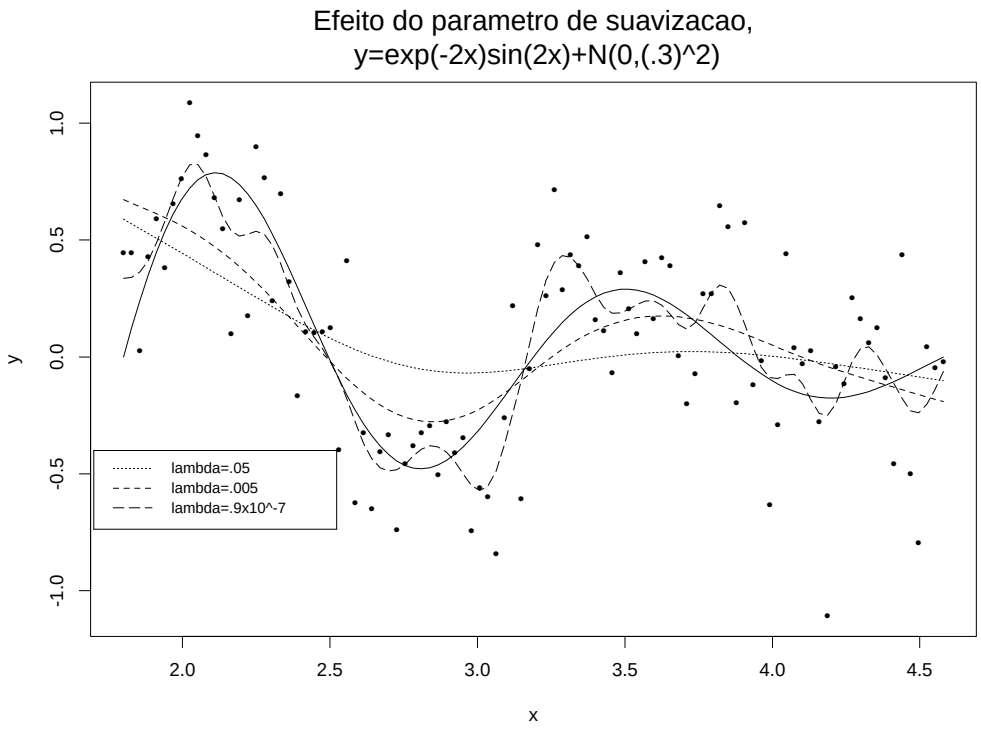

Figure 5.1: relação peso/altura com idade em meses

vai ser retirada é arbitrária, a eficiência do procedimento com λ pode ser quantificada pela função de validação cruzada

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\lambda^{(-i)}(t_i))^2$$

e então escolhe-se λ que minimiza $CV(\lambda)$.

Não se pode garantir que a função $CV(\lambda)$ tenha um único mínimo. Então deve-se tomar cuidado com sua minimização. Qualquer que seja o método a ser usado, envolverá o cálculo de $CV(\lambda)$ para vários valores de λ e daí um procedimento para computar CV é muito importante.

À primeira vista parece que temos que resolver n problemas diferentes de suavização para obter n curvas f^{-i}. Mas isto não é o caso. Note que,

$$f_\lambda = H(\lambda)Y,$$

onde $H(\lambda) = X(X'X + \lambda\Omega)^{-1}X'Y$. Pode-se mostrar que

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - f_\lambda(t_i)}{1 - h_{ii}(\lambda)})^2.$$

Validação Cruzada Generalizada é uma modificação do príncipio de validação cruzada (Veja Craven and Wahba (1979)). A idéia básica do GCV é substituir os fatores $1 - h_{ii}(\lambda)$ por sua médias $1 - \frac{1}{n}\sum_{i=1}^{n} h_{ii}$. Assim,

$$GCV(\lambda) = \frac{1}{n}\,\frac{\sum_{i=1}^{n}(Y_i - f_\lambda(t_i))^2}{(1 - \frac{1}{n}Tr(H))^2}.$$

Se os h_{ii} forem iguais e os $t_i's$ forem igualmente espaçados num intervalo sujeito a condições de fronteiras periódicas então, GCV=CV.

6 O ponto de vista Bayesiano para suavização

Suponha que as observações $Y_i \sim N(f(t_i, \sigma^2)$ sejam independentes $i = 1, \ldots, n$. Então o logaritmo da função de verossimilhança é dado por

$$l(f) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - f(t_i))^2.$$

A maximização sem restrições de $l(f)$ mostra que a solução é a interpolação. Intuitivamente, a justificativa Bayesiana para penalizar a verossimilhança é colocar uma densidade a priori proporcional a $\exp(-(\alpha/2)\int(\ddot{f})^2)$ sobre o espaço de todas as funções suaves. Com esta a priori, a log posteriori é,

$$l_p(f) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - f(t_i))^2 - \frac{\alpha}{2}\int(\ddot{f})^2,$$

onde $\alpha = (\lambda/\sigma^2)$ e λ é o parametro de suavização.
Maximizar $l_p(f)$ sé equivalente a minimizar $A_\lambda(f)$

6.1 O paradoxo de Wahba

O espaço de curvas f que desejamos escolher como a curva de regressão que melhor se ajusta aos dados é obviamente de dimensão infinita. Mesmo se

escolhermos curvas que são limitadas num intervalo $[a, b]$ e que tenham $\int_a^b (\ddot{f})^2 < \infty$. Porém, este tipo de formulação leva ao paradoxo de Wahba (Wahba, 1983).

Assuma que $[a, b] = [0, 1]$ e seja $\mathcal{S}$ o espaço de funções em $[0, 1]$ tais que $\forall f \in \mathcal{S}$, $\int (\ddot{f})^2 < \infty$. Então existe uma sequência de funções ortonormais $\{\phi_j\}$ e uma sequência crescente de autovalores ν_j com $0 = \nu_1 = \nu_2 < \nu_j$ para $j \geq 3$ tal que $\forall f \in \mathcal{S}$, $f = \sum a_j \phi_j$, onde os a_j são os coeficientes da expansão. E que, $\forall f \in \mathcal{S}$, $\int_0^1 (\ddot{f})^2 = \sum a_j^2 \nu_j$. Assim a densidade a priori é proporcional à $\exp((-\alpha/2) \int (\ddot{f})^2)$ pode ser escrita como

$$\exp(-\frac{\alpha}{2} \sum_{j \geq 3} a_j^2 \nu_j) = \prod_{j \geq 3} \exp(-\frac{\alpha}{2} a_j^2 \nu_j).$$

Assumindo, que ν_1 e ν_2 tem aprioris impróprias uniformes $(-\infty, \infty)$ e que ν_j para $j \geq 3$ tenham distribuições normais independentes com média zero e variância $(\alpha \nu_j)^{-1}$, temos

$$\begin{aligned} \int (\ddot{f})^2 = \sum_{j=3}^{\infty} a_j^2 \nu_j &= \sum_{j=3}^{\infty} \nu_j (N(0, (\alpha \nu_j) - 1))^2 \\ &= \alpha^{-1} \sum_{j=3}^{\infty} (N(0, 1))^2 = +\infty \end{aligned}$$

com probabilidade 1.

6.2 Dimensão Finita

Um modelo simples de dimensão finita desenvolvido por Silverman(1985) pode ser descrito como se segue. Denote por simplicidade os B-splines por $N_j(t)$ como uma base para os splines naturais. Defina, matrizes $n \times n$, X com $X_{i,j} = N_j(t_i)$ e $\Omega_{i,j} = \int \ddot{N}_i(t) \ddot{N}_j(t) dt$. Então se assumirmos um modelo Bayesiano no qual os dados tem distribuição Gaussiana com média $X\theta$ e varância $\sigma^2 I$ e o vetor de coeficientes θ com distribuição multivariada Gaussiana a priori com média zero e varância $\tau \Omega^-$, onde $\tau = (\sigma^2/\lambda)$. Segue que a média da distribuição a

posteriori de θ é $E(\theta|Y) = \hat{\theta} = (X'X + \lambda\Omega)^{-1}X'Y$ e a matriz de covariância é $cov(\theta|Y) = (\Omega/\tau + \sigma^{-2}X'X)^{-1}$. Dada a distribuição a posteriori de θ, podemos fazer inferências sobre todos os aspectos do spline cúbico $\sum_{j=1}^{n} N_j(t)\theta_j$. Em particular, podemos computar a média e variância (a posteriori) de $f_\lambda = X\theta$. A média a posteriori é $\hat{f}_\lambda = H(\lambda)Y$ e a covariância posteriori é dada por $H(\lambda)\sigma^2$.

7 Exemplos

Abaixo mostramos alguns exemplos de modelos obtidos através da metodologia exposta acima.

Figura 7.1 e 7.2 mostram o ajuste obtido por suavização por splines e um intervalo de confiança 95% para o modelo simulado $y = (\sin(2\pi x^3))^3 3 + N(0, .25)$. Já as figuras 7.3 e 7.4 exibem ajustes feitos com dados reais.

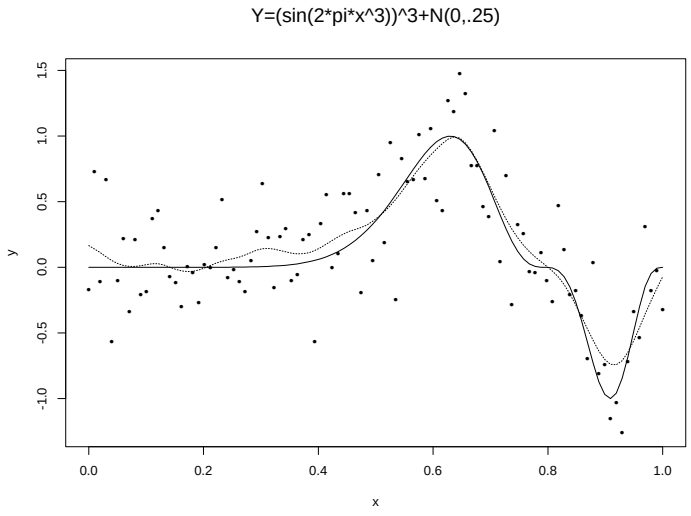

Figure 7.1: linha cheia é o modelo simulado e a linha pontilhada é o modelo ajustado

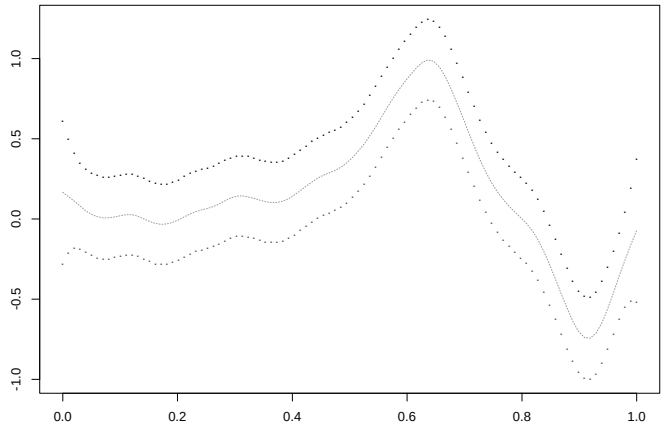

Figure 7.2: Intervalo de Confiança 95% para $y = (\sin(2\pi x^3))^3 3 + N(0, .25)$

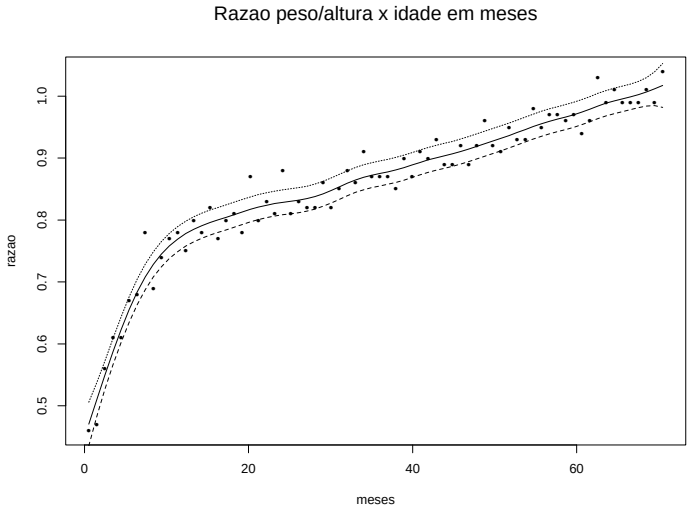

Figure 7.3: razão peso/altura com idade em meses

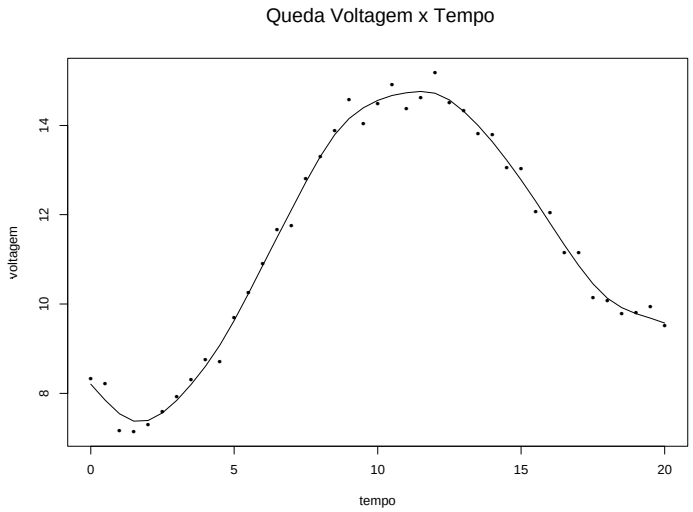

Figure 7.4: modelo ajustado para a relação queda de voltagem vs. tempo

References

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**: 377–403.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, New York.

Reinch, C. (1967). Smoothing by spline functions, *Numerische Mathematik* **10**: 177–183.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline, *JRSS-B, Methodological* **45**: 133–150.

Whittaker, E. T. (1923). On new method of graduation, *Proc. Edinburgh Math. Soc.* **41**: 63–75.