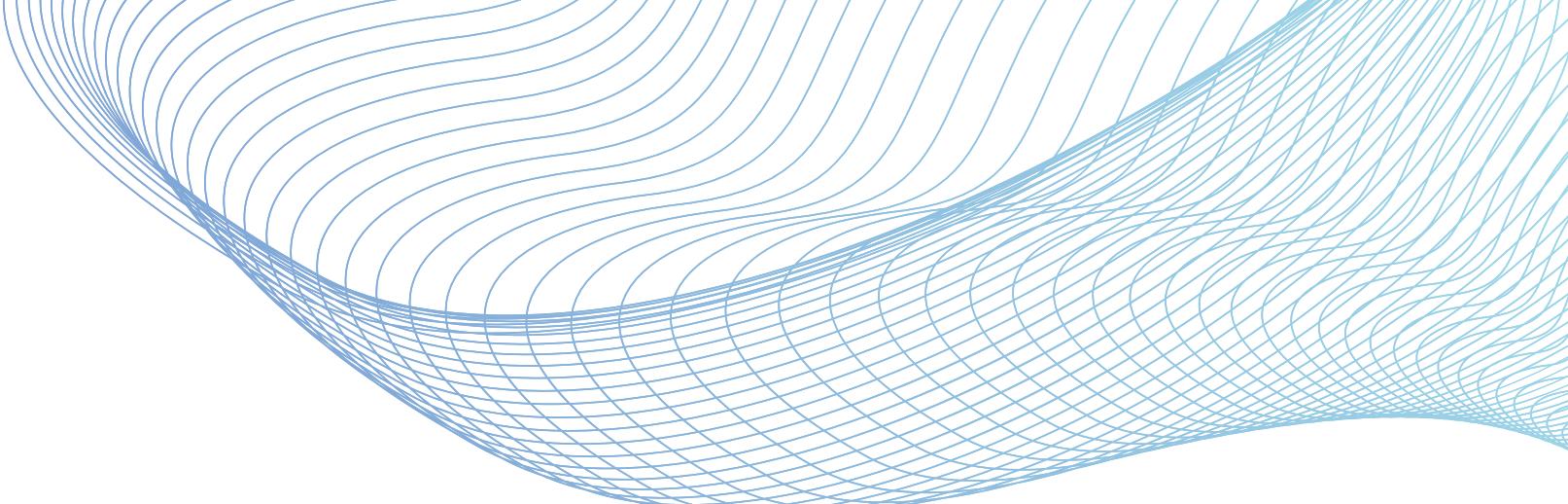


MODELOS ADITIVOS PARCIALMENTE LINEARES

Aluna: Camila Galhardo Toledo
Orientador: Clécio da Silva Ferreira



O QUE SÃO?



Modelos Aditivos Parcialmente Lineares são uma abordagem estatística que combina características de modelos lineares e modelos não lineares para análise de dados. Essa abordagem é frequentemente usada quando se suspeita que a relação entre as variáveis independentes e dependentes não seja completamente linear, mas também não seja completamente não linear.

Modelo de Regressão Linear

- + de 1 variável explicativa linear

Modelo Semi-paramétrico

- + de 1 variável explicativa linear
- 1 variável explicativa não linear

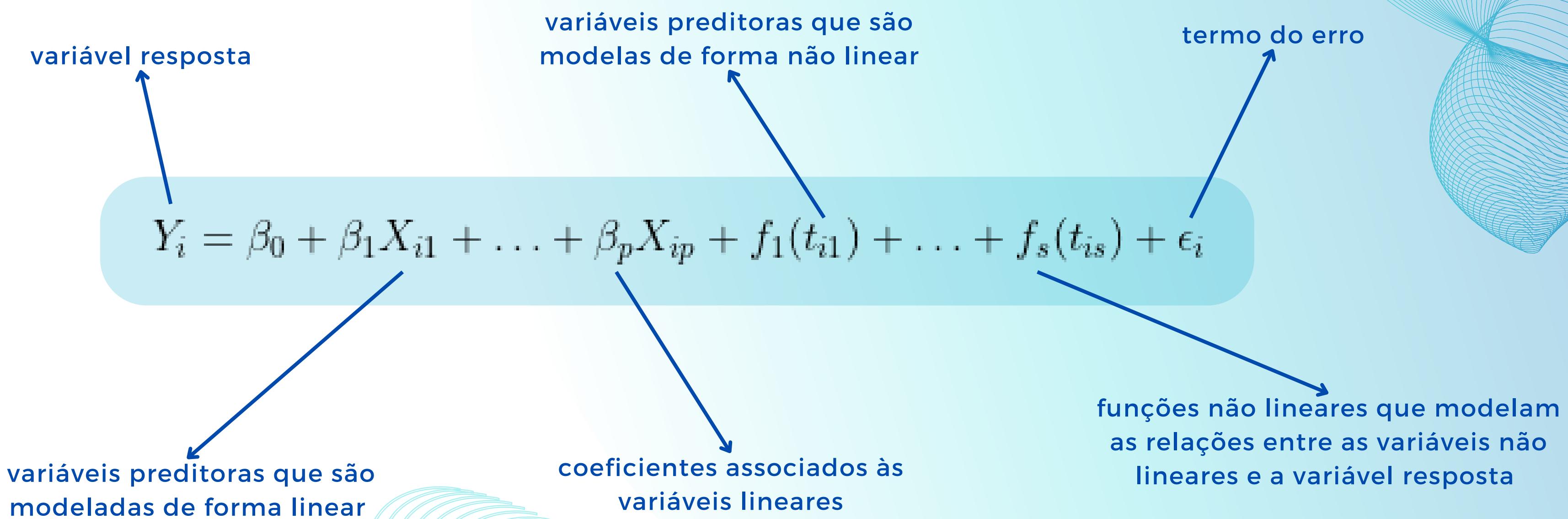
Modelo Aditivo

- + de 1 variável explicativa não linear

Modelo Aditivo Parcialmente Linear

- + de 1 variável explicativa linear
- + de 1 uma variável explicativa não linear

REPRESENTAÇÃO DO MODELO



SPLINES

Splines são uma técnica matemática utilizada para criar curvas suaves e contínuas a partir de pontos de controle. Um B-spline de grau q consiste de pedaços polinomiais, conectados de uma forma especial pelos chamados nós.

Neste trabalho utilizamos B-splines para descrever as componentes não paramétricas.

$$f_j(t_{ji}) = \sum_{l=1}^{q_j} \theta_{lj} B_l(t_{ji}), \quad j = 1, \dots, s$$

Reescrevendo o modelo:

$$Y = X\beta + \sum_{j=1}^s N_j \gamma_j + \epsilon$$

$$\gamma_j = (f_j(t_{1j}), \dots, f_j(t_{nj}))$$

matriz de incidência

ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA PENALIZADA

Sob suposição de normalidade dos erros aleatórios, a função de log-verossimilhança associada ao modelo é dada por

$$\ell(\beta, \sigma^2, \gamma_1, \dots, \gamma_s) \approx -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y} - \mathbf{X}\beta - \sum_{j=1}^s \mathbf{N}_j \gamma_j)^T (\mathbf{Y} - \mathbf{X}\beta - \sum_{j=1}^s \mathbf{N}_j \gamma_j)$$

Maximizar a equação acima sem impor restrições sobre as funções não lineares pode causar sobre-ajuste ou não-identificabilidade no parâmetro β . Um procedimento bem conhecido é o de estimação de máxima verossimilhança penalizada, que consiste em incorporar na função de log-verossimilhança uma função penalty.

$$\ell(\beta, \sigma^2, \gamma_1, \dots, \gamma_s, \alpha) \approx \ell(\beta, \sigma^2, \gamma_1, \dots, \gamma_s) - 1/2 \sum_{j=1}^s \alpha_j \gamma_j^T K_j \gamma_j$$

parâmetro de suavização

matriz penalty obtida como função apenas dos nós

ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA PENALIZADA

Através de métodos iterativos chega-se aos EMVP de $\theta = (\beta, \sigma^2, \gamma_1, \dots, \gamma_s)$

$$\begin{aligned}\hat{\beta}^{(k)} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \sum_{j=1}^s \mathbf{N}_j \hat{\gamma}_j^{(k)}) \\ \hat{\gamma}_i^{(k)} &= (\mathbf{N}_i^T \mathbf{N}_i + \hat{\sigma}^{2(k)} \alpha_i \mathbf{K}_i)^{-1} \mathbf{N}_i^T (\mathbf{Y} - \mathbf{X} \hat{\beta}^{(k)} - \sum_{j=1, j \neq i}^s \mathbf{N}_j \hat{\gamma}_j^{(k)}), \quad i = 1, \dots, s \\ \hat{\sigma}^{2(k)} &= \frac{1}{n} (\mathbf{Y} - \mathbf{X} \hat{\beta}^{(k)} - \sum_{j=1}^s \mathbf{N}_j \hat{\gamma}_j^{(k)})^T (\mathbf{Y} - \mathbf{X} \hat{\beta}^{(k)} - \sum_{j=1}^s \mathbf{N}_j \hat{\gamma}_j^{(k)})\end{aligned}$$

Além disso, o parâmetro de suavização é selecionado através do critério BIC.

MATRIZ DE INFORMAÇÃO DE FISHER

Os erros padrão da estimativa de $\theta = (\beta, \sigma^2, \gamma_1, \dots, \gamma_s)$ podem ser obtidos através da matriz de informação de Fisher, através da relação:

$$Ep(\theta) = \sqrt{diag(MI^{-1})}$$

Em casos onde a matriz de informação de Fisher não é inversível, uma outra abordagem é utilizada para o cálculo dos erros padrão das estimativas dos parâmetros, o bootstrap paramétrico. Utilizando esta técnica, são feitas reamostragens dos erros aleatórios sob os parâmetros estimados e a partir das estimativas obtidas nestas reamostragens, são calculados os erros padrão de cada parâmetro.



SIMULAÇÃO

Um estudo de simulação foi realizado através do método de Monte Carlo com 1000 replicações para os tamanhos amostrais iguais a 200, 500 e 1000. O modelo simulado é composto de duas variáveis explicativas não lineares e uma variável explicativa linear.

$$Y_i = \beta X_i + f_1(t_{1i}) + f_2(t_{2i}) + \epsilon_i, \quad i = 1, \dots, n,$$

onde

$$X \sim U(0, 1)$$

$$\beta = 5$$

$$f_1(t_1) = \cos(t_1)$$

$$f_2(t_2) = \cos(4\pi t_2) e^{-t_2^2/2}$$

$$\epsilon_i \sim N(0, 0.05)$$

$$t_1 \sim U(2\pi, 6\pi)$$

$$t_2 \sim U(0.6, 1.6)$$

SIMULAÇÃO

A Figura 1 apresenta as 1000 curvas estimadas da curva 1, enquanto a Figura 2 apresenta as 1000 curvas estimadas da curva 2.

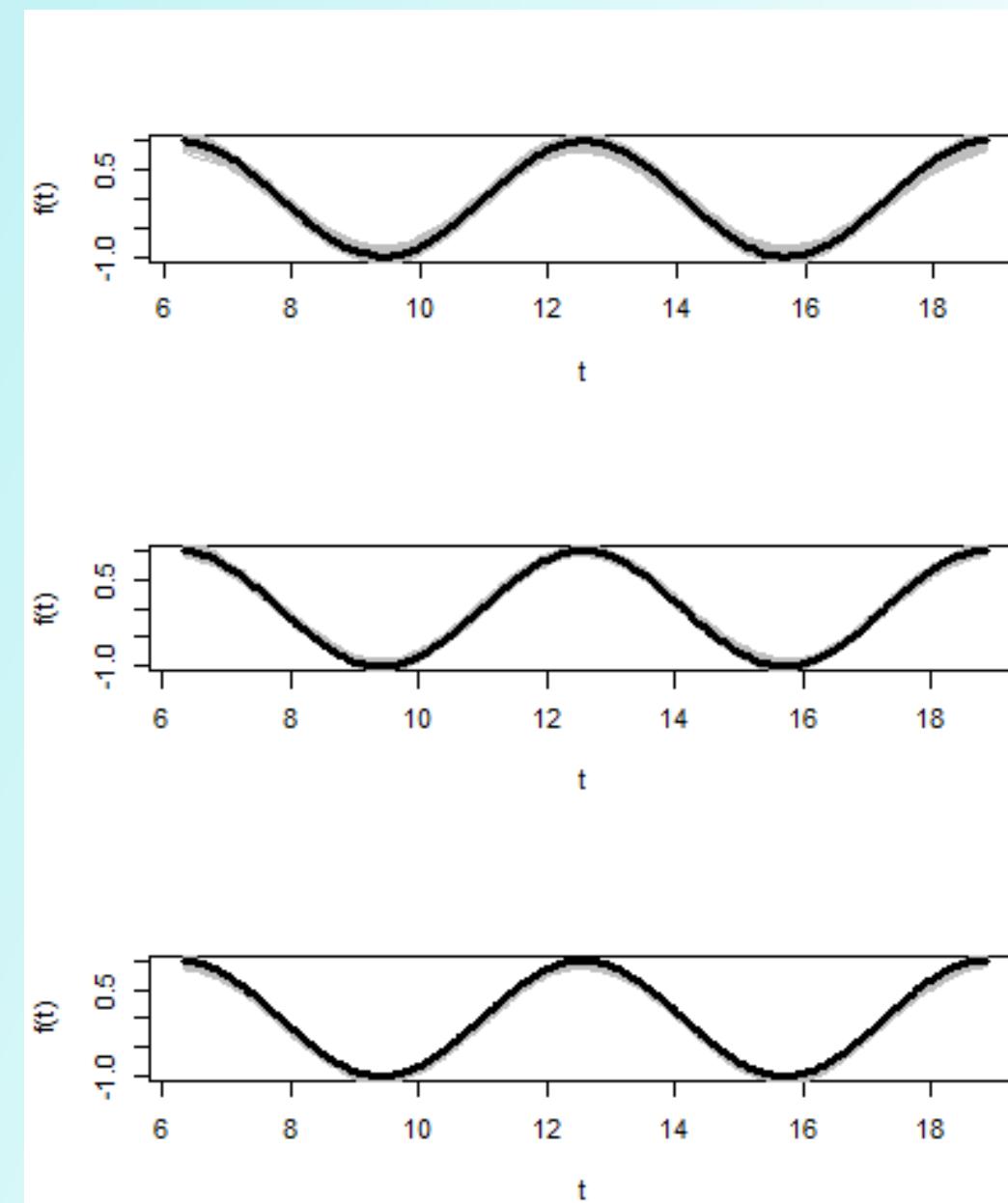


Figura 1: $f(t) = \cos(t)$

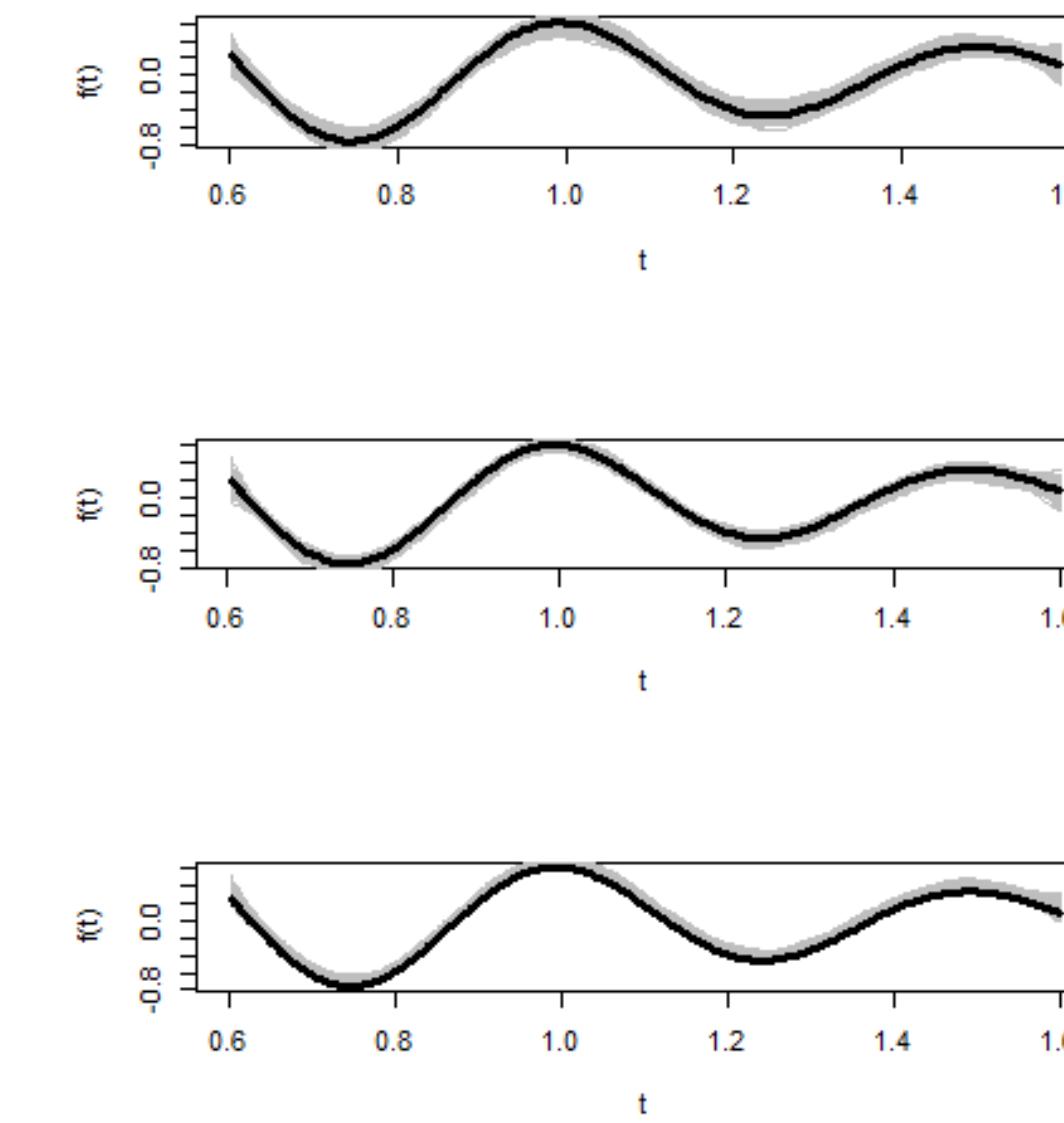


Figura 2: $f(t) = \cos(4\pi t)e^{-t^2/2}$

SIMULAÇÃO

Pela Tabela 1, nota-se um comportamento satisfatório do modelo, uma vez que as curvas estimadas estão muito próximas da curva real. Além disso, observa-se que ao aumentar o tamanho amostral, os desvios padrão e os desvios padrão empíricos dos estimadores de β e σ diminuem, e a média de suas estimativas se aproximam de seus valores reais, alcançando as propriedades assintóticas do Estimador de Máxima Verossimilhança (EMV).

	n = 200				n = 500				n = 1000				
	Valor Real	Média	DP	DP Emp	Média	DP	DP Emp	Média	DP	DP Emp	Média	DP	DP Emp
σ	0.05	0.046	0.005	0.005	0.048	0.003	0.003	0.049	0.002	0.002	0.049	0.002	0.002
β	5.00	4.995	0.056	0.055	5.001	0.036	0.034	5.000	0.024	0.025	5.000	0.024	0.025

Tabela 1: Médias e desvios padrão das estimativas e média dos desvios padrão obtidos pela matriz de informação dos parâmetros.

APLICAÇÃO

Para a aplicação foi utilizado o conjunto de dados "milan.mort" do pacote **SemiPar** disponível no software R, o qual contém dados de 3652 dias consecutivos (10 anos consecutivos: de 1º de janeiro de 1980 a 30 de dezembro de 1989) para a cidade de Milão, Itália. O objetivo é modelar a variabilidade do total de mortes utilizando como variáveis explicativas não lineares a média de temperatura e a humidade relativa e como variáveis lineares partículas totais em suspensão no ar ambiente (TSP) e número de dias (t) desde 31 de dezembro de 1979.

$$\sqrt{mortality} = \beta_0 + \beta_1 TSP + \beta_2 t + f_1(temperature_t) + f(humidity_t) + \epsilon$$

APLICAÇÃO

Parâmetro	Estimativa	Erro Padrão
σ^2	0.3256739	0.0265091
β_0	5.803937	0.0001389
β_1	0.000837	0.0000093
β_2	-0.000173	0.0490978
α_1	100	-
α_2	2.758055	-
$\ell(\theta)$	-3482.079	-
SIC	7070.462	-

Tabela 2: Mortalidade: Estimativas de máxima verossimilhança penalizada dos parâmetros e seus erros-padrão.

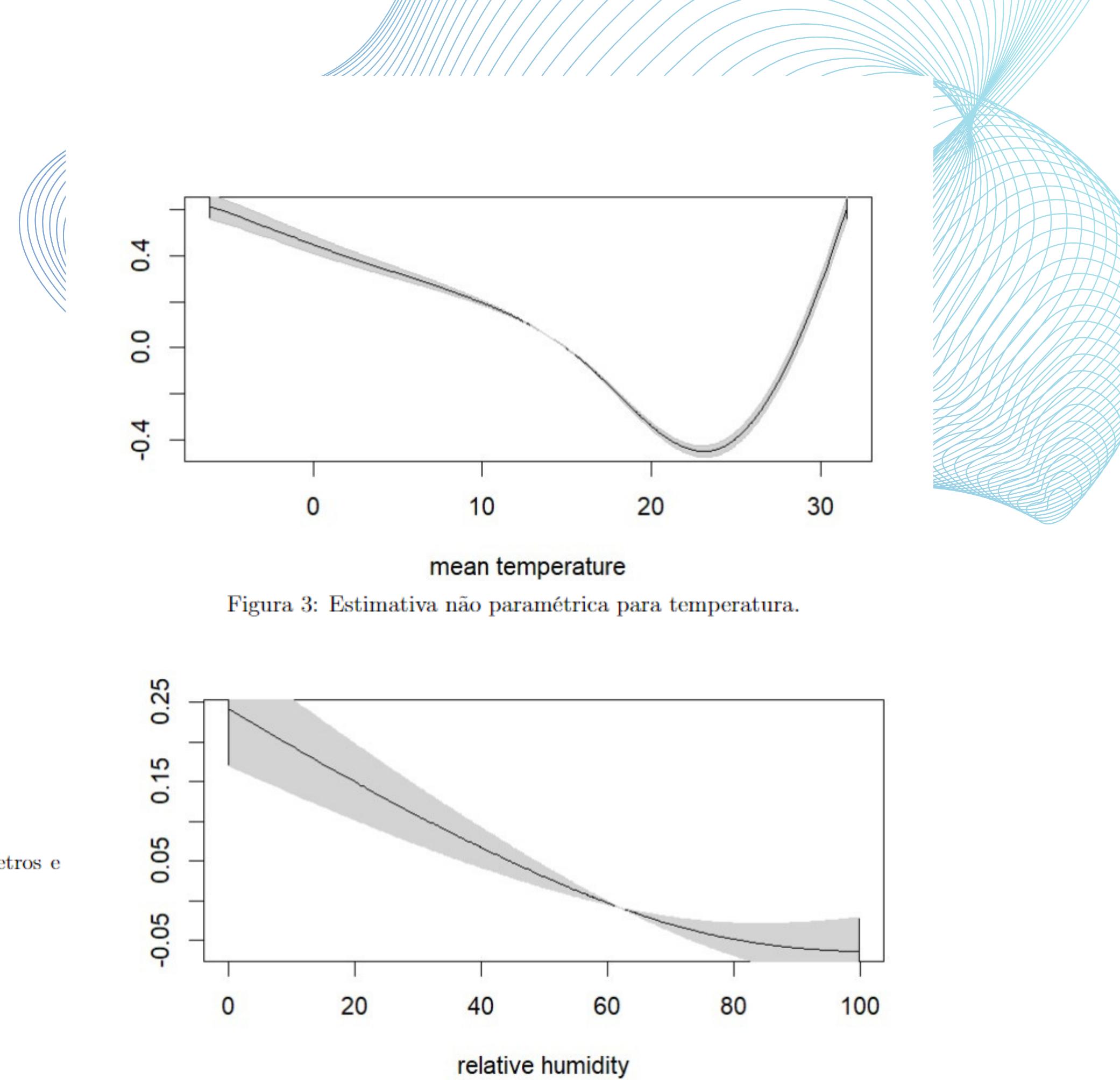


Figura 3: Estimativa não paramétrica para temperatura.

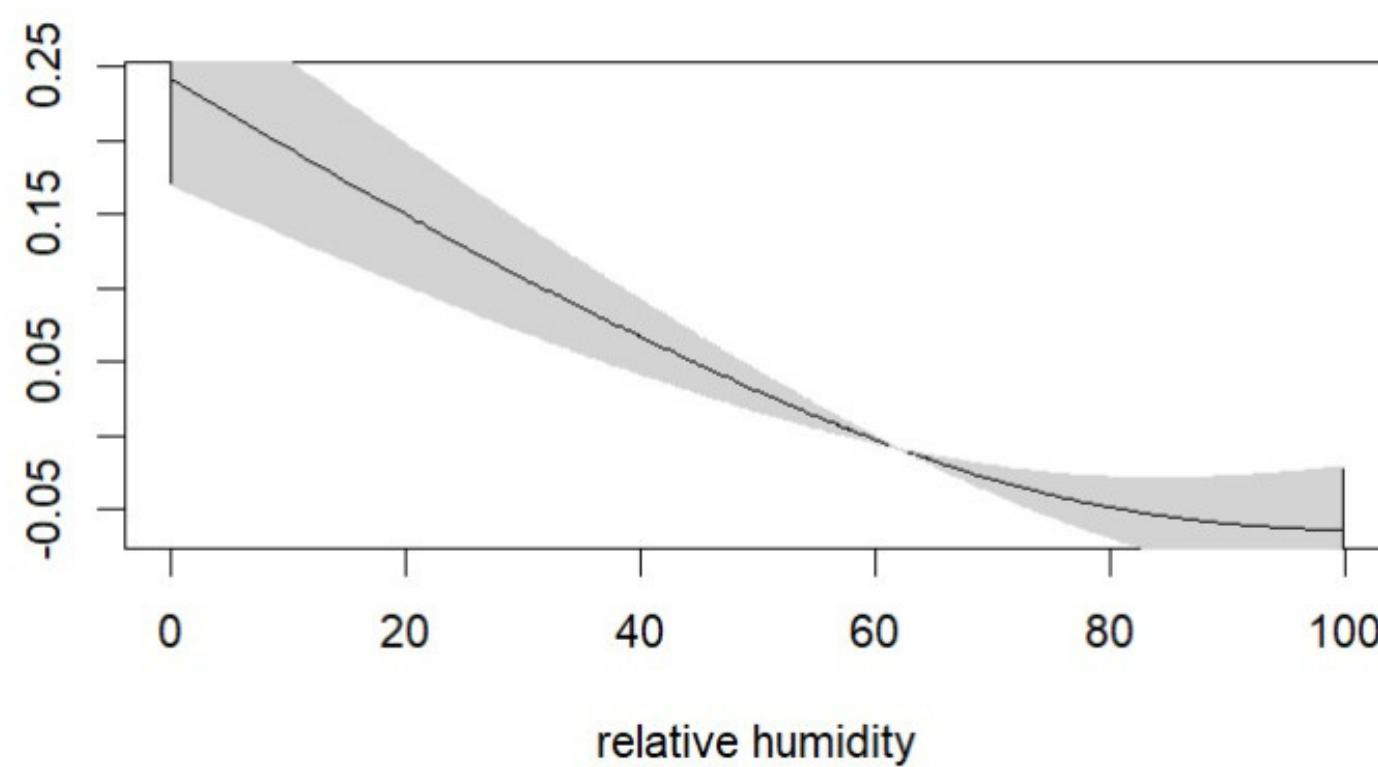


Figura 4: Estimativa não paramétrica para humidade.

APLICAÇÃO

Uma das maneiras de avaliarmos o ajuste do modelo é analisando o gráfico envelope, apresentado na Figura 5. Analisando o gráfico, temos evidências de que o modelo foi bem ajustado, pois os resíduos encontram-se entre as faixas do gráfico envelope.

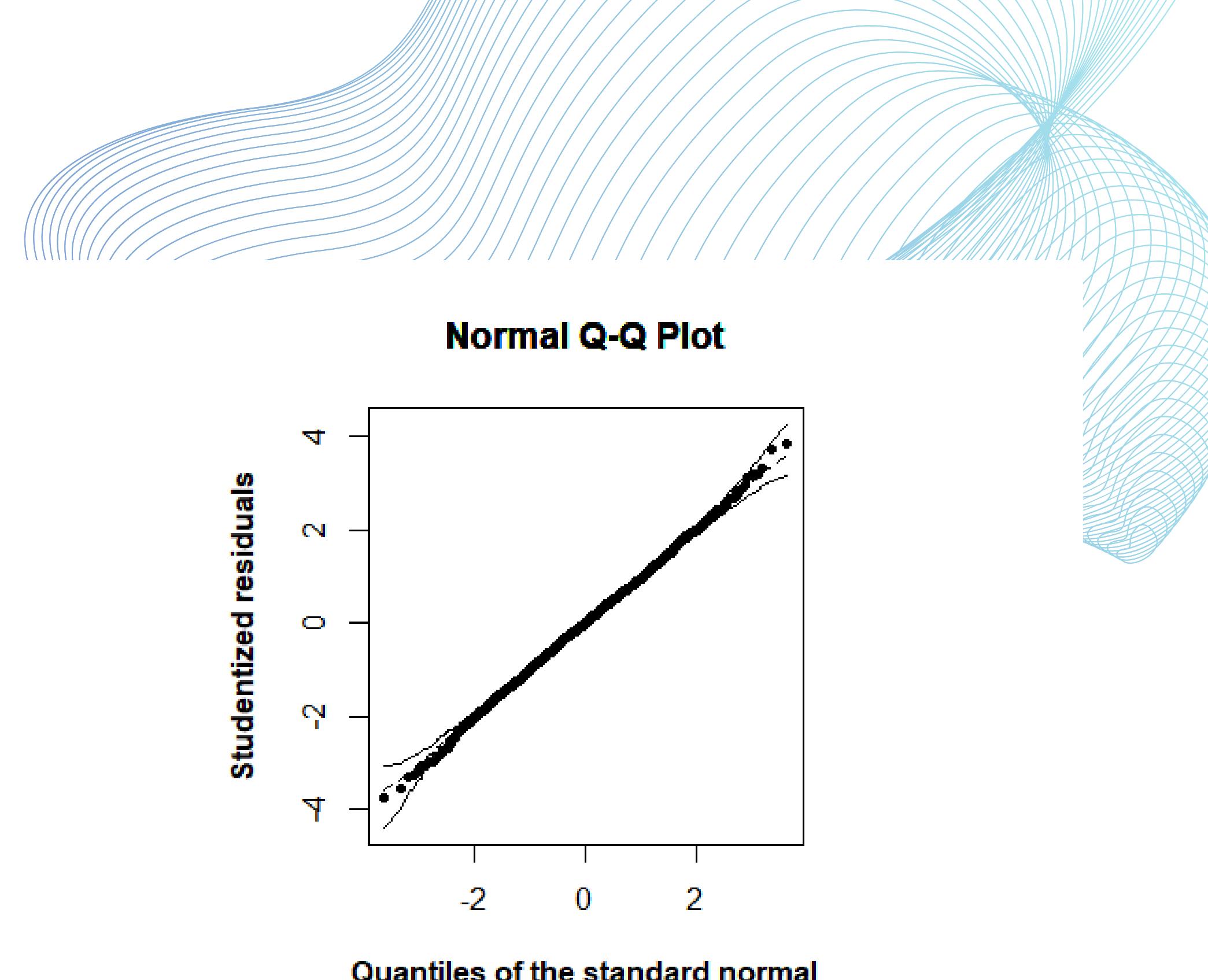
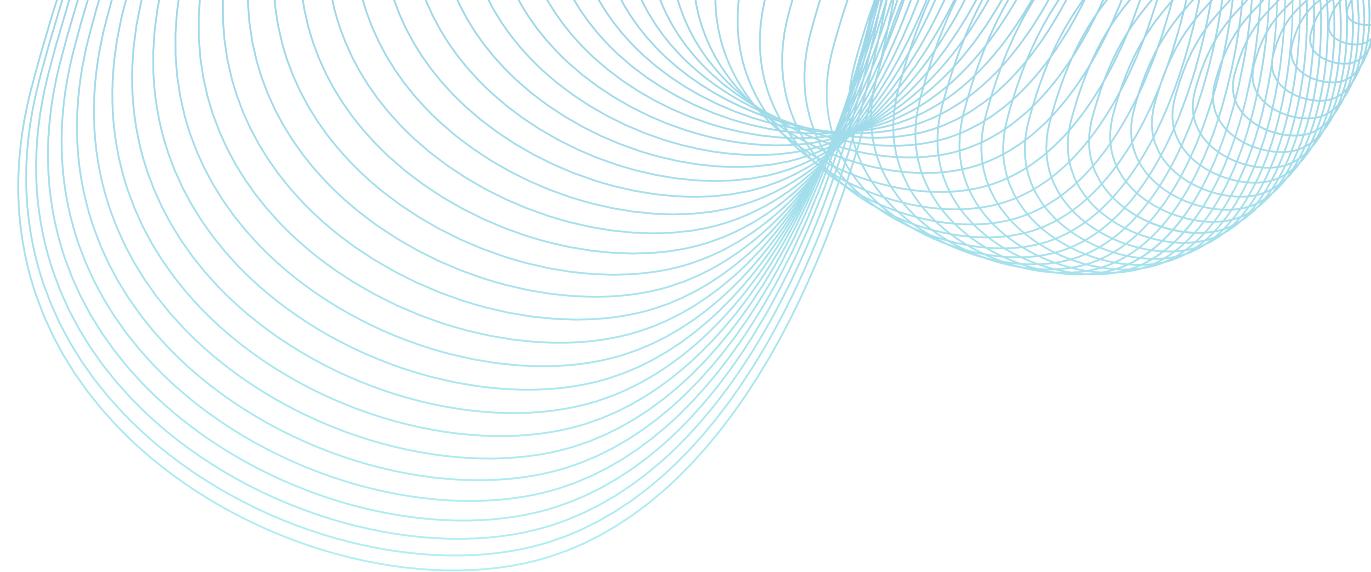


Figura 5: Mortalidade: Gráfico de envelope simulado

CONCLUSÃO



Neste trabalho desenvolvemos os estimadores de máxima verossimilhança e a matriz de informação de Fisher para os modelos aditivos parcialmente lineares. Um estudo de simulação foi realizado para verificar as propriedades assintóticas do EMVP. O modelo foi aplicado a um conjunto de dados reais sobre mortalidade na cidade de Milão (Itália), mostrando a eficiência do modelo.

OBRIGADA :)