MODELOS SEMI-PARAMÉTRICOS UTILIZANDO SPLINES

Joyce da Silva Lopes
Orientador: Prof. Clécio da Silva Ferreira

Departamento de Estatística – UFJF

Universidade Federal de Juiz de Fora

INTRODUÇÃO

Modelo semi-paramétrico

Um modelo semi-paramétrico é uma abordagem estatística que combina elementos de modelos paramétricos e não paramétricos para capturar tanto relações lineares quanto padrões mais complexos e flexíveis entre variáveis. Nesse tipo de modelo, parte das variáveis explicativas é tratada de forma paramétrica, seguindo uma forma funcional conhecida, enquanto outra parte é tratada de maneira não paramétrica, permitindo que a relação seja modelada de forma mais flexível.

INTRODUÇÃO

A estrutura de um modelo semi-paramétrico pode ser expressa da seguinte maneira:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{g(t)} + \boldsymbol{\epsilon}$$

Onde:

- **Y** é o vetor de variáveis dependentes (ou variável resposta);
- **X** é a matriz de variáveis independentes (ou variáveis explicativas paramétricas);
- **β** é o vetor de parâmetros paramétricos a serem estimados;
- **g(t)** é a função não paramétrica que representa a relação flexível entre variáveis não paramétricas;
- **ϵ** é o vetor de erros aleatórios
- **ϵ** N $(0, \sigma^2 I)$, temos então: **Y** N $(X\beta + N\gamma, \sigma^2 I)$

Neste trabalho, vamos abordar o modelo semi-paramétrico com a variável resposta possuindo distribuição normal.

Universidade Federal de Juiz de Fora

INTRODUÇÃO

<u>O que são splines?</u>

Splines são funções matemáticas suaves e segmentadas que são usadas para modelar relações não lineares entre variáveis em análises estatísticas e interpolações. Se baseiam em polinômios cúbicos (de grau 3) que são unidos de forma contínua para formar uma curva suave.

São particularmente úteis quando se deseja ajustar uma curva a um conjunto de dados, mas essa curva não pode ser facilmente representada por uma única função polinomial simples.

INTRODUÇÃO

Rugosidade da curva

A rugosidade da curva deve ser considerada, além do bom ajuste aos dados;

Uma maneira de medir a rugosidade: $\int_b^a (g''(t))^2 \, dt$;

Dada qualquer função duas vezes diferenciável g(t) definida entre [a,b] e um parametro de suavização **α > 0**, a soma dos quadrados penalizada é:

$$S(g) = \Sigma \{Yi - g(ti)\}^2 + \alpha \int_b^a (g''(t))^2 \, dt$$

INTRODUÇÃO

Seleção de **α**

O parâmetro de suavização **α** controla o grau de penalização aplicado à função não paramétrica durante o processo de estimação.

- Valores mais altos de **α:** penalidade mais forte, curvas mais suaves;
- Valores mais baixos de **α:** reduzem a penalização, levam a um ajuste mais flexível, mas também mais instável e com maior risco de *overfitting*.

Necessidade de um método automático para selecionar o parâmetro de suavização: Validação cruzada

Universidade Federal de Juiz de Fora

DESENVOLVIMENTO

<u>Estimadores de Máxima Verossimilhança Penalizada</u>

Considerando o modelo semi-paramétrico, temos que:

$\theta = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2})$ são os parâmetros a serem estimados.

Supondo que $\epsilon \quad N(0, \sigma^2 I)$, temos então: $Y \quad N(X\beta + N\gamma, \sigma^2 I)$

O logaritmo da máxima verossimilhança penalizada é dado por:

$$\mathbf{lp}(\theta) = -n/2 \log(2\pi\sigma^2) - (2\sigma^2)^{-1} (Y - X\beta - N\gamma)^T (Y - X\beta - N\gamma) - \alpha/2\, \gamma^T K \gamma$$

$$\hat{\beta} = (X^T X)^{-1} X^T (Y - N\hat{\gamma})$$

$$\hat{\gamma} = (N^T N + \hat{\sigma}^2 \alpha K)^{-1} N^T (Y - X\hat{\beta})$$

$$\hat{\sigma}^2 = (Y - X\hat{\beta} - N\hat{\gamma})^T (Y - X\hat{\beta} - N\hat{\gamma})\, n^{-1}$$

Universidade Federal de Juiz de Fora

DESENVOLVIMENTO

Matriz de Informação de Fisher

A matriz de informação de Fisher permite calcular o desvio padrão das estimativas de β, γ, σ^2 por meio da seguinte relação:

$$\mathbf{Ep(\hat{\theta}) = diag(MI^{-1})^{1/2}}$$

$$MI = \begin{bmatrix} -\frac{\partial^2 l_p(\theta)}{\partial\beta\partial\beta^T} & -\frac{\partial^2 l_p(\theta)}{\partial\beta\partial\gamma} & -\frac{\partial^2 l_p(\theta)}{\partial\beta\partial\sigma^2} \\ -\frac{\partial^2 l_p(\theta)}{\partial\gamma\partial\beta} & -\frac{\partial^2 l_p(\theta)}{\partial\gamma\partial\gamma^T} & -\frac{\partial^2 l_p(\theta)}{\partial\gamma\partial\sigma^2} \\ -\frac{\partial^2 l_p(\theta)}{\partial\sigma^2\partial\beta} & -\frac{\partial^2 l_p(\theta)}{\partial\sigma^2\partial\gamma} & -\frac{\partial^2 l_p(\theta)}{\partial\sigma^2\partial\sigma^2} \end{bmatrix}$$

Figura 1 - Matriz de Informação de Fisher

DESENVOLVIMENTO

<u>Simulação</u>

Temos:

- **Y** = **Xβ** + **g(t)** + **ϵ**
- Uma variável explicativa paramétrica **x**   U (0,1)
- **X** é uma matriz com 2 colunas: primeira coluna de 1's e a segunda formada pela variável **x**
- **β0** = 2; **β1** = 5, onde **β0** é o intercepto
- **g(t)** = $\cos(4\pi t) * \exp(-t^2/2)$
- **ϵ**   N (0, 0.25 I) (σ^2 = 0.25)

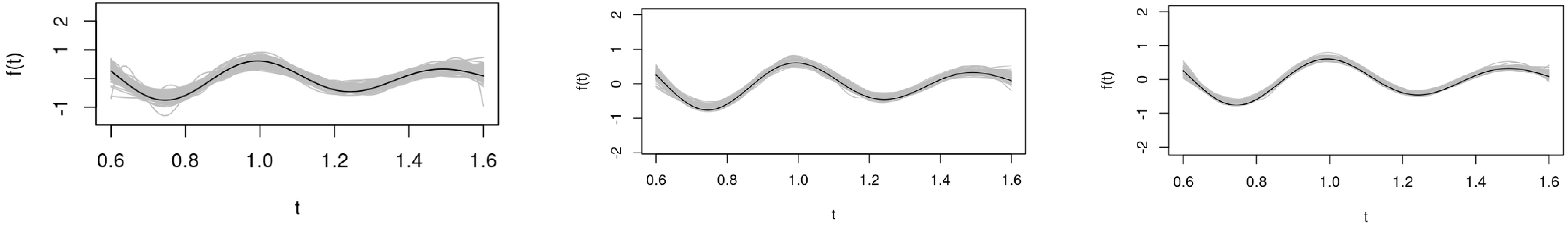
Universidade Federal de Juiz de Fora

DESENVOLVIMENTO

Resultados (Simulação)

		n=100			n=500			n=1000		
Parâmetro	Valor Real	Estimativa	SD	SDe	Estimativa	SD	SDe	Estimativa	SD	SDe
β_0	5.00	5.0279	0.2117	0.1791	5.0067	0.0691	0.0755	5.0087	0.0564	0.0549
β_1	2.00	2.0046	0.1758	0.1847	1.9903	0.0856	0.0776	2.0013	0.0576	0.0546
σ^2	0.25	0.2264	0.0336	0.0327	0.2410	0.0160	0.0153	0.2465	0.0117	0.0110

Tabela 1 - Estimativas e desvios padrão dos parâmetros



a b c

Gráficos 1, 2 e 3 - Modelo normal semi-paramétrico, utilizando função com 100 replicações. Curvas estimadas (linhas cinza), curva verdadeira (linha preta): a) n=100 b) n=500 e c) n=1000

UNIVERSIDADE FEDERAL DE JUIZ DE FORA

DESENVOLVIMENTO

<u>Aplicação com dados reais</u>

Conjunto de dados: **Onions{SemiPar}**

O conjunto de dados contém 84 observações de uma experiência envolvendo a produção de cebolas brancas espanholas em dois locais do sul da Austrália.

As variáveis do conjunto são:

Density: densidade areal de plantas (plantas por metro quadrado)

Yield: rendimento de cebola (gramas por planta)

Location: indicador de localização: 0=Purnong Landing, 1=Virgínia

$$\log(\text{Yield}) = \beta(\text{Location}) + g(\text{Density})$$

Universidade Federal de Juiz de Fora

DESENVOLVIMENTO

Resultados (Aplicação com dados reais)

	Estimativa	Sd
β	-0.3344	0.0229
σ^2	0.0105	0.0016

Tabela 2 - Estimativas e desvios padrão dos parâmetros

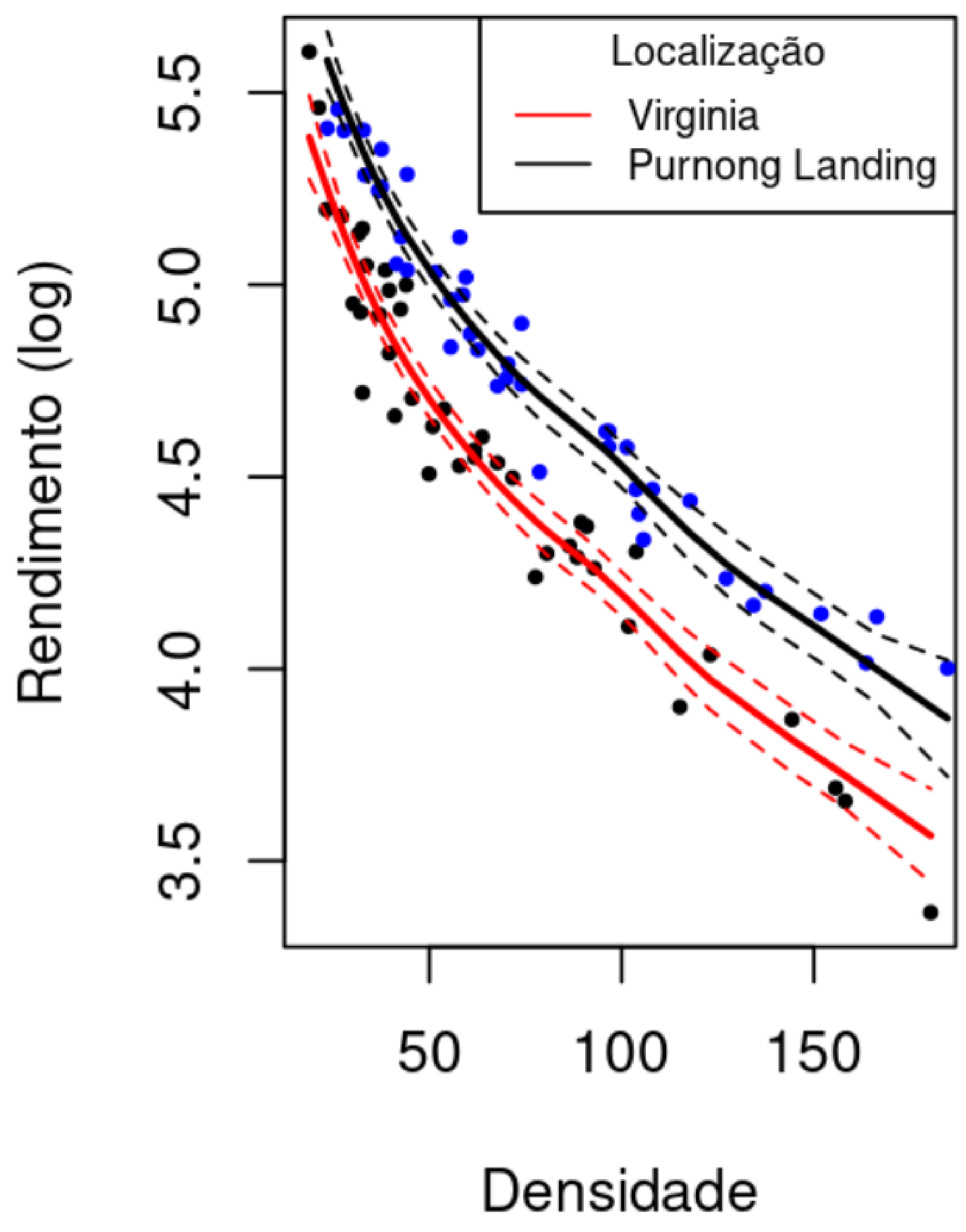


Gráfico 4 - Rendimentos(log) x Densidade

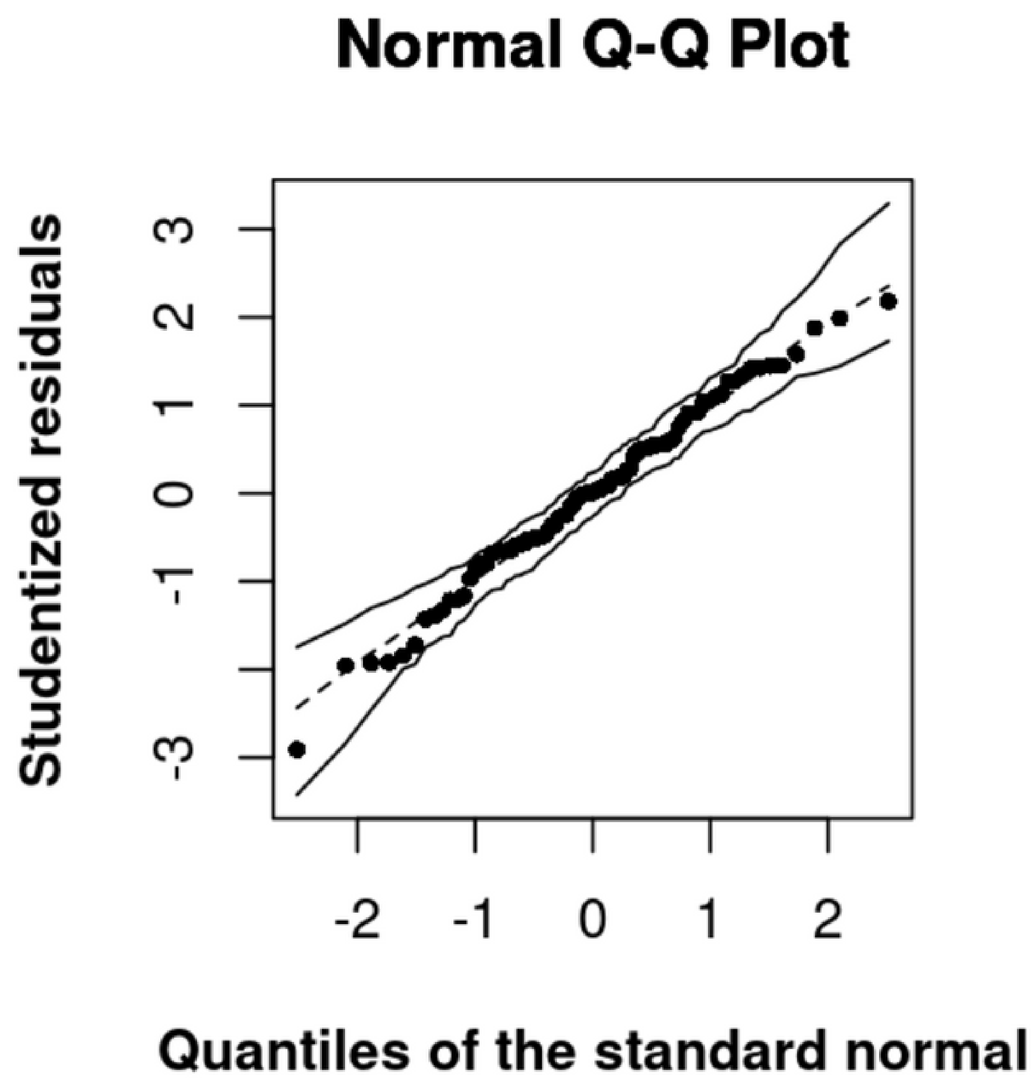


Gráfico 5 - Envelope de resíduos: Modelo semi-paramétrico (Onions)

CONCLUSÕES

- De maneira geral, o objetivo do trabalho foi alcançado, sendo possível utilizar splines para modelar curvas complexas;
- Desenvolvemos os cálculos dos estimadores de máxima verossimilhança penalizada e da matriz de informação de Fisher para o modelo semi-paramétrico;
- Problemas com os cálculos dos erros padrões das estimativas por meio da matriz de informação de Fisher, devido a não inversão da matriz em alguns casos, foram resolvidos adotando outra técnica, o B*ootstrap*;

UNIVERSIDADE FEDERAL DE JUIZ DE FORA

CONCLUSÕES

- Foram feitas simulações para comprovar as propriedades assintóticas dos estimadores e uma aplicação que nos permitiu verificar a eficiência do modelo ajustado;
- O trabalho ainda segue em desenvolvimento, avançando para estimação de curvas de modelos parcialmente lineares, que apresentam mais de uma função não paramétrica, sendo mais complexos que os modelos semi-paramétricos e outras extensões

Universidade
Federal de Juiz de Fora

OBRIGADA!