UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Joysce da Silva Lopes

Seleção de Modelos Aditivos Parcialmente Lineares

Juiz de Fora
2025

Joysce da Silva Lopes

Seleção de Modelos Aditivos Parcialmente Lineares

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador: Prof. Dr. Clécio da Silva Ferreira

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Lopes, Joysce da Silva.
Seleção de Modelos Aditivos Parcialmente Lineares / Joysce da Silva Lopes. – 2025.
78 f. : il.

Orientador: Clécio da Silva Ferreira
Trabalho de Conclusão de Curso (graduação) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas . Departamento de Estatística, 2025.

1. modelos aditivos parcialmente. 2. p-splines. 3. segunda derivada. I. Ferreira, Clécio da Silva, orient. II. Título.

Joysce da Silva Lopes

Seleção de Modelos Aditivos Parcialmente Lineares

Trabalho de Conclusão de Curso apresentado ao Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em 24 de janeiro de 2025

BANCA EXAMINADORA

Prof. Dr. Clécio da Silva Ferreira - Orientador
Universidade Federal de Juiz de Fora

Prof[a]. Dr[a]. Camila Borelli Zeller
Universidade Federal de Juiz de Fora

Prof. Dr. Tiago Maia Magalhães
Universidade Federal de Juiz de Fora

Dedico este trabalho à minha família e aos meus amigos,
que trilharam este caminho comigo.

AGRADECIMENTOS

Agradeço a Deus por me guiar e permitir a conclusão deste processo, pois nada disso seria possível sem Ele. Aos meus pais, Adriano e Josiane, por sonharem comigo e me apoiarem incondicionalmente. À minha avó, Inês, por todo o amor e suporte. Ao meu irmão, Adryan, por transformar momentos difíceis em risadas inesquecíveis. Ao meu orientador, Clécio, pela orientação excepcional e paciência. Aos professores do departamento, pelo ensino e partilha de conhecimentos. E aos colegas de turma, pela amizade e aprendizado mútuo ao longo da jornada.

“Grandes coisas fez o SENHOR por nós, e, por isso, estamos alegres.” (Salmos 126:3 - Almeida Revista e Corrigida).

RESUMO

As covariáveis em um modelo podem apresentar uma relação linear ou não linear com a variável resposta, resultando em diferentes configurações modelares. Por meio da comparação das diversas combinações possíveis à medida que as covariáveis são ajustadas nas partes paramétrica e não paramétrica do modelo, este trabalho tem como principal objetivo selecionar o modelo mais adequado, ou seja, aquele que oferece o melhor ajuste aos dados analisados. O processo de comparação consiste, primeiramente, em analisar as curvas da derivada de segunda ordem que cada covariável apresenta, identificando aquelas que possivelmente compõem cada parte do modelo. Em seguida, comparam-se os modelos resultantes das diferentes combinações, utilizando técnicas como o teste de razão de verossimilhança e o critério de informação bayesiano (BIC). Essa abordagem metodológica ainda não foi explorada por nenhum estudo até o momento, evidenciando sua contribuição e avanço significativos para o tema. Neste trabalho, são abordadas diferentes categorias de modelos, que vão desde os Modelos de Regressão Linear Múltipla até os Modelos Aditivos Parcialmente Lineares (APLMs, do inglês Additive Partial Linear Models), sendo este último o principal foco do estudo. Além da seleção dos modelos, foi possível atingir outros objetivos importantes para a construção dos modelos, tais como: obter os estimadores de máxima verossimilhança penalizada para os Modelos Aditivos Parcialmente Lineares, construir as curvas referentes à parte não paramétrica por meio de P-splines e determinar o parâmetro de suavização por meio de validação cruzada. Os estudos de simulação, bem como as aplicações em dados reais apresentadas neste trabalho, geraram resultados que demonstram o alcance dos objetivos, oferecendo contribuições valiosas para pesquisas futuras.

Palavras-chave: modelos aditivos parcialmente lineares; p-splines; segunda derivada.

ABSTRACT

The covariates in a model may exhibit either a linear or a nonlinear relationship with the response variable, leading to different model configurations. By comparing the various possible combinations as the covariates are assigned to the parametric and nonparametric parts of the model, this study aims to select the most appropriate model—i.e., the one that provides the best fit to the analyzed data. The comparison process begins with an analysis of the second-order derivative curves for each covariate, identifying those that are likely to compose each part of the model. Then, the resulting models from different combinations are compared using techniques such as the likelihood ratio test and the Bayesian Information Criterion (BIC). This methodological approach has not yet been explored in any study, highlighting its significant contribution and advancement in the field. In this study, different categories of models are examined, ranging from Multiple Linear Regression Models to Additive Partial Linear Models (APLMs), the latter being the primary focus of this research. In addition to model selection, other important objectives were achieved for model construction, such as obtaining the penalized maximum likelihood estimators for Additive Partial Linear Models, constructing the curves related to the nonparametric part using P-splines, and determining the smoothing parameter through cross-validation. The simulation studies, as well as the real-data applications presented in this work, produced results that demonstrate the achievement of the objectives, offering valuable contributions to future research.

Keywords: additive partially linear models; P-splines; second derivative.

LISTA DE FIGURAS

Figura 1 – Modelos em função do número de variáveis lineares e não lineares. . . 19
Figura 2 – Contorno suave obtido por meio de uma spline. 20
Figura 3 – Histogramas das 1000 estimativas dos parâmetros β_1,β_2 e σ^2, respectivamente, para amostra de tamanho igual a 500. 37
Figura 4 – Histogramas das 1000 estimativas dos parâmetros β_1,β_2 e σ^2, respectivamente, para amostra de tamanho igual a 1000. 37
Figura 5 – Histogramas das 1000 estimativas dos parâmetros β_1,β_2 e σ^2, respectivamente, para amostra de tamanho igual a 2000. 38
Figura 6 – Estimativas não paramétricas para as variáveis t_1 (à esquerda) e t_2 (à direita) para amostra de tamanho igual a 500. 39
Figura 7 – Estimativas não paramétricas para as variáveis t_1 (à esquerda) e t_2 (à direita) para amostra de tamanho igual a 1000. 40
Figura 8 – Estimativas não paramétricas para as variáveis t_1 (à esquerda) e t_2 (à direita) para amostra de tamanho igual a 2000. 40
Figura 9 – Estimativas da segunda derivada das 1000 curvas estimadas no modelo aditivo (em cinza) e curva real da segunda derivada (em preto). 42
Figura 10 – Estimativas da segunda derivada das 1000 curvas estimadas no modelo aditivo (em cinza) e curva real da segunda derivada (em preto). 42
Figura 11 – Estimativas da segunda derivada das 1000 curvas estimadas no modelo aditivo (em cinza) e curva real da segunda derivada (em preto). 43
Figura 12 – P-valores do teste de razão de verossimilhanças entre os modelos linear e APML. 45
Figura 13 – P-valores do teste de razão de verossimilhanças entre os modelos APML e aditivo. 46
Figura 14 – Qui-quadrado teórica vs. Estatísticas observadas (n=500). 47
Figura 15 – Qui-quadrado teórica vs. Estatísticas observadas (n=1000). 47
Figura 16 – Qui-quadrado teórica vs. Estatísticas observadas (n=2000). 48
Figura 17 – BICs dos modelos 1, 2 e 3 para n=500. 49
Figura 18 – BICs dos modelos 1, 2 e 3 para n=1000. 50
Figura 19 – BICs dos modelos 1, 2 e 3 para n=2000. 50
Figura 20 – Estimativas da curva para TSP 53
Figura 21 – Estimativa da 2ª derivada para TSP 54
Figura 22 – Estimativa da curva para day.num 54
Figura 23 – Estimativa da 2ª derivada para day.num 55
Figura 24 – Estimativa da curva para mean.temp 55
Figura 25 – Estimativa da 2ª derivada para mean.temp 56
Figura 26 – Estimativa da curva para rel.humid 56

Figura 27 – Estimativa da 2ª derivada para rel.humid 57
Figura 28 – Milan.Mort: Envelope dos resíduos 58
Figura 29 – Estimativa da curva para temperature 60
Figura 30 – Estimativa da 2ª derivada para temperature 60
Figura 31 – Estimativa da curva para wind 61
Figura 32 – Estimativa da 2ª derivada para wind 61
Figura 33 – Estimativa da curva para day.in.seas 62
Figura 34 – Estimativa da 2ª derivada para day.in.seas 62
Figura 35 – Estimativa não paramétrica para day.in.seas 65
Figura 36 – Ragweed: Envelope dos resíduos 66
Figura 37 – Estimativa da curva para tax 68
Figura 38 – Estimativa da 2ª derivada para tax 68
Figura 39 – Estimativa da curva para dis 69
Figura 40 – Estimativa da 2ª derivada para dis 69
Figura 41 – Estimativa da curva para lstat 70
Figura 42 – Estimativa da 2ª derivada para lstat 70
Figura 43 – Estimativa não paramétrica para a variável dis 72
Figura 44 – Estimativa não paramétrica para a variável lstat 73
Figura 45 – Boston: Envelope dos resíduos 74

LISTA DE TABELAS

Tabela 1 – Médias e erros padrão das estimativas dos parâmetros do modelo. . . . 36
Tabela 2 – P-valores dos testes de normalidade dos EMVPs dos parâmetros para cada tamanho de amostra. 38
Tabela 3 – Proporção média de pontos das bandas de confiança (95%) contendo o valor zero para cada curva 44
Tabela 4 – Milan.mort: Estimativas de máxima verossimilhança penalizada dos parâmetros e seus erros-padrão 57
Tabela 5 – Ragweed: p-valores dos testes de razão de verossimilhanças 63
Tabela 6 – Ragweed: BIC dos modelos ajustados 64
Tabela 7 – Ragweed: Estimativas de máxima verossimilhança penalizada dos parâmetros, seus erros-padrão e p-valores 64
Tabela 8 – Boston: p-valores dos testes de razão de verossimilhanças 71
Tabela 9 – Boston: BIC dos modelos ajustados 71
Tabela 10 – Boston: Estimativas de máxima verossimilhança penalizada dos parâmetros, seus erros-padrão e p-valores 71

LISTA DE ABREVIATURAS E SIGLAS

AIC	Akaike Information Criteria
BIC	Bayesian Information Criteria
EMVPs	Estimadores de Máxima Verossimilhança Penalizada
GAMs	Modelos Aditivos Generalizados
APLMs	Modelos Aditivos Parcialmente Lineares

SUMÁRIO

1 INTRODUÇÃO 13
2 MODELOS ADITIVOS PARCIALMENTE LINEARES 17
2.1 SPLINES CÚBICAS 20
2.2 PENALIZAÇÃO DA RUGOSIDADE DA CURVA 23
2.3 SELEÇÃO DO PARÂMETRO DE SUAVIZAÇÃO (λ) 25
2.3.1 *Validação cruzada* 25
2.4 ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA PENALIZADA 27
2.5 MATRIZ DE INFORMAÇÃO DE FISHER 28
2.6 MATRIZ DE INFORMAÇÃO OBSERVADA 29
2.7 MATRIZ DE INFORMAÇÃO ESPERADA 29
3 SELEÇÃO DE VARIÁVEIS 31
3.1 DERIVADA DE SEGUNDA ORDEM DAS CURVAS ESTIMADAS . . 32
3.2 TESTE DA RAZÃO DE VEROSSIMILHANÇAS 33
3.3 GRAUS DE LIBERDADE 34
4 SIMULAÇÃO 36
4.1 VERIFICAÇÃO DAS PROPRIEDADES ASSINTÓTICAS DOS EMVPs 36
4.2 SELEÇÃO DAS VARIÁVEIS COM RELAÇÕES LINEARES E NÃO LINEARES 40
5 APLICAÇÃO 52
5.1 MILAN.MORT 52
5.2 RAGWEED 58
5.3 BOSTON 66
6 CONCLUSÃO 75
REFERÊNCIAS 77

1 INTRODUÇÃO

Em diversas áreas do conhecimento, como economia, biologia e medicina, é comum tentar explicar uma variável dependente com base em variáveis independentes, pois essa abordagem permite entender como fatores externos ou internos influenciam determinado fenômeno. Por exemplo, na economia pode-se querer entender como o salário de um indivíduo é influenciado por sua idade e nível de escolaridade. Em seu livro, Wooldridge (2012) explora amplamente o uso de modelos de regressão na análise econômica. Ele apresenta aplicações práticas, como a investigação de fatores que influenciam salários e outros tópicos relacionados à economia aplicada. Assim como na economia, há inúmeros exemplos em outras áreas, como biologia, psicologia e marketing, onde existe o interesse em compreender as relações entre a variável de interesse e as variáveis explicativas.

Uma das relações mais comuns é aquela que se dá de maneira linear, ou seja, a mudança nas variáveis indendepentes está associada a uma mudança proporcional na variável resposta. Quando este tipo de relação ocorre, pode-se facilmente representá-la por meio de um Modelo de Regressão Linear Simples ou um Modelo de Regressão Linear Múltipla, de acordo com o número de variáveis estudadas. Entretanto, sabe-se que essas relações podem não ser exclusivamente lineares, tornando os Modelos de Regressão Linear menos eficazes para representar a relação entre as variáveis. Sendo assim, é necessário o uso de modelos estatísticos que permitam representar a relação existente de forma mais apropriada.

Os Modelos Aditivos Generalizados (GAMs, do inglês Generalized Additive Models) representam uma vasta categoria de modelos de regressão, onde o impacto de cada variável preditora na variável resposta é representado de maneira muito flexível através de uma função g não previamente definida. Algumas vantagens destes modelos apontadas por Hastie e Tibshirani (1990) são: ao invés de impor uma relação estritamente linear entre a variável preditora e a resposta, os GAMs permitem que cada preditor tenha uma relação não paramétrica individual com a resposta. Isso é particularmente útil em contextos onde a relação funcional entre as variáveis é desconhecida ou altamente complexa e também elimina a necessidade de testar manualmente diferentes transformações para cada variável individualmente. Apesar de sua flexibilidade, os GAMs preservam uma estrutura aditiva, o que facilita a interpretação. É possível examinar o efeito de uma variável enquanto as demais são mantidas constantes, tornando-os úteis para estudos de inferência. Estes modelos também apresentam desvantagens como: o modelo é restrito a ser aditivo. Com muitas variáveis, interações importantes podem não ser capturadas. Em situações onde as interações são importantes, é necessário incluir explicitamente termos de interação, como $X_j \times X_k$, ou suavizadores bidimensonais. Além disso, para conjuntos de dados muito grandes ou um número elevado de variáveis, o ajuste de GAMs pode ser computacionalmente caro, embora avanços em algoritmos de ajuste, como o método de

Wood (2017), tenham melhorado essa eficiência. Os GAMs são modelos importantes e de muita utilidade nos casos em que há relações não lineares entre as variáveis preditoras e a variável resposta. Podemos citar um estudo feito por Schwartz (1994) que aborda exemplos de aplicação dos GAMs nas relações entre a poluição por partículas no ar e a mortalidade diária em Birmingham, Alabama; entre poluentes (partículas, ozônio e dióxido de enxofre) e internações hospitalares por doenças respiratórias na Filadélfia e entre episódios de tosse em crianças e poluição (ozônio e partículas) em seis cidades do leste dos EUA. No contexto da epidemiologia, por exemplo, os GAMs são amplamente utilizados para modelar padrões não lineares, como os efeitos do clima em surtos de doenças transmissíveis (Hastie e Tibshirani (1990); Wood (2017)).

Enquanto os modelos lineares assumem que todas as variáveis têm efeitos estritamente lineares, e os GAMs permitem uma flexibilidade total ao modelar os efeitos de forma não linear, os Modelos Aditivos Parcialmente Lineares (APLMs) oferecem uma abordagem híbrida, combinando características de ambos. Os APLMs são particularmente úteis em cenários onde algumas covariáveis apresentam relações lineares com a variável resposta, enquanto outras têm relações mais complexas e não lineares. A vantagem em utilizá-los nestes casos, encontra-se no fato de que por meio dessa estrutura é possível obter maior flexibilidade do que nos modelos lineares tradicionais, capturando padrões complexos em algumas variáveis enquanto preserva a linearidade em outras. Essa característica híbrida é proveitosa em áreas como economia, biologia e ciências sociais, onde as relações entre variáveis frequentemente exibem esses padrões.

Uma das possíveis maneiras de representar as relações lineares e não lineares entre as variáveis explicativas e a variável resposta no APLM é dada por:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{s} g_j(\boldsymbol{t}_j) + \boldsymbol{\epsilon},$$

onde $\boldsymbol{Y} = (y_1, \ldots, y_n)$ denota a variável dependente ou resposta de um conjunto de dados com n observações; $\boldsymbol{X} = [X_1, ..., X_p]$ é a matriz, de dimensão $n \times p$, sendo as p colunas variáveis explicativas que se relacionam de maneira linear com a variável $\boldsymbol{Y}$; $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ é o vetor de coeficientes destas covariáveis; $g_1, ..., g_s$ denotam as relações não paramétricas entre as covariáveis $\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_s$ e a variável resposta do modelo; e $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ representa o vetor de erros aleatórios do modelo, que neste trabalho apresenta distribuição normal, com média igual a $\underset{\sim}{\mathbf{0}}$ e variância σ^2, ou seja, $\boldsymbol{\epsilon} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}}, \sigma^2\mathbf{I}_n)$, onde $\mathbf{I}_n$ é a matriz identidade de ordem n. Logo, consequentemente, a variável resposta também possui distribuição normal.

Na literatura, podemos encontrar alguns exemplos de aplicação utilizando os APLMs em diversas áreas de estudo, como: Liu et al. (2011) aplicaram um APLM para investigar a relação entre beta-caroteno e tipos de câncer, considerando a idade e índices de colesterol como componentes não lineares; Ibacache-Pulgar et al. (2013) modelaram

o preço de casas em Boston, utilizando a taxa de imposto como componente linear e as variáveis porcentagem de menor status da população e distância ponderada a cinco centros de emprego da cidade como componentes não lineares e Wang et al. (2019) estudaram o valor médio aplicado pelo plano de saúde Medicare nos 50 estados dos Estados Unidos, utilizando várias variáveis demográficas de forma linear e, como não lineares, o número de beneficiários distintos do plano, o número de serviços prestados e o número de beneficiários distintos do Medicare por dias de serviço.

Diante disso, este trabalho busca estudar os Modelos Aditivos Parciais Lineares (APLMs), dada sua significativa aplicabilidade em diversas áreas do conhecimento. A implementação de um algoritmo que permita a estimação dos parâmetros usando Estimadores de Máxima Verossimilhança Penalizada (EMVPs), o uso de splines cúbicas como técnica para construção das curvas estimadas, o cálculo da variância dos estimadores pela matriz de informação de Fisher e a obtenção do parâmetro de suavização via validação cruzada são alguns dos principais assuntos apresentados neste trabalho.

Além disso, este estudo propõe uma abordagem, ainda não explorada na literatura, para a identificação das covariáveis com relações lineares e não lineares nos APLMs. A proposta é identificar as covariáveis que compõem cada parte do modelo, analisando a derivada de segunda ordem da curva estimada, que representa a relação entre cada covariável e a variável resposta. E, em seguida, há a comparação entre os diferentes modelos ajustados, utilizando o teste de razão de verossimilhanças e BIC como critérios para encontrar o melhor modelo.

Feita a apresentação da metodologia, o trabalho aborda dois estudos de simulação no Capítulo 4. Um primeiro estudo para mostrar as propriedades assintóticas dos Estimadores de Máxima Verossimilhança Penalizada (EMVPs) obtidos para os APLMs, na Seção 4.1, e um segundo estudo para apresentar como pode-se identificar variáveis com relações lineares e não lineares analisando as segundas derivadas das curvas estimadas, na Seção 4.2.

Considerando a relevância das aplicações a dados reais para validar a eficácia dos métodos propostos e demonstrar sua aplicabilidade em contextos práticos, este trabalho apresenta três estudos de caso utilizando os conjuntos de dados “Milan.mort”, “Ragweed” e “Boston”, disponíveis nos pacotes SemiPar (2018) e MASS (2002) do software R Core Team (2024).

O pacote SemiPar (2018) foi desenvolvido para realizar análises de regressão semiparamétrica, complementando o conteúdo do livro de Ruppert et al. (2003). Neste pacote encontra-se funções que auxiliam na modelagem mais flexível de dados complexos além de conjuntos de dados como “Milan.mort” e “Ragweed”. O pacote MASS (2002), é mais popular quando comparado ao SemiPar (2018). Foi desenvolvido pelos estatísticos Venables e Ripley e fornece diversas funções para estatística aplicada, incluindo métodos para modelagem estatística, estimação de densidade, regressão robusta e análise de

variáveis categóricas. Entre os conjuntos de dados encontardos neste pacote, está o conjunto “Boston”, também utilizado na aplicação do presente trabalho.

A ideia na primeira aplicação é modelar as mortes ocorridas em Milão, na Itália, no período de 1º de janeiro de 1980 a 30 de dezembro de 1989. Na segunda aplicação, o objetivo é modelar os níveis diários de pólen em Kalamazoo, Michigan. Já na terceira e última aplicação, o intuito é modelar os preços das casas em Boston, EUA. Mais detalhes sobre os conjuntos de dados serão apresentados no capítulo 5.

Em todas as aplicações tem-se o uso da metodologia que o estudo apresenta para a identificação do tipo de relação que as variáveis preditoras apresentam com a variável resposta. Uma vez identificadas, há o ajuste e a análise de um modelo adequado que capture as relações lineares e não lineares entre as variáveis. Para introduzir a metodologia abordada neste trabalho, serão apresentados, no Capítulo 2, detalhes sobre os modelos aditivos, destacando as particularidades relacionadas à composição e à variação do número de covariáveis que integram as partes paramétrica e não paramétrica desses modelos. Como já mencionado, esse tipo de abordagem permite explorar como diferentes covariáveis contribuem para o ajuste do modelo, seja de forma linear ou não linear, garantindo maior flexibilidade na representação das relações entre as variáveis.

Ainda no Capítulo 2, serão descritas as técnicas utilizadas para a modelagem da parte não paramétrica dos modelos estudados. Essas técnicas são fundamentais para capturar padrões complexos nos dados sem a necessidade de impor uma estrutura funcional rígida. O uso de métodos baseados em splines, por exemplo, permite ajustar suavemente curvas às observações, promovendo um equilíbrio entre fidelidade aos dados e suavidade no ajuste.

2 MODELOS ADITIVOS PARCIALMENTE LINEARES

Em diversas áreas do conhecimento, é comum o uso de modelos estatísticos para entender, descrever ou prever o comportamento de uma ou mais variáveis em relação a outras. Na economia, por exemplo, esses modelos são utilizados para prever a inflação e analisar o impacto de políticas fiscais. Na medicina, ajudam a avaliar a eficácia de tratamentos e prever a progressão de doenças. Já na engenharia, auxiliam no controle de qualidade e na otimização de processos produtivos.

Quando é assumido que essa relação segue uma forma funcional que é fixa e que pode ser facilmente representada por uma função matemática já conhecida, podemos modelar usando os chamados modelos paramétricos. Um simples exemplo de modelos paramétricos é o Modelo de Regressão Linear Múltipla, podendo ser representado da seguinte maneira:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

em que $\boldsymbol{Y} = (y_1, \ldots, y_n)$ é um vetor coluna de dimensões $n \times 1$ que contém os valores da variável resposta para cada uma das n observações; $\boldsymbol{X}$ é uma matriz de dimensões $n \times p$, onde cada linha corresponde a uma observação e cada coluna corresponde a uma variável preditora; $\boldsymbol{\beta} = (\beta_1, ...\beta_p)$ é um vetor coluna de dimensões $p \times 1$ que contém os coeficientes a serem estimados para cada variável preditora. E por fim, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ é o erro aleatório do modelo. A simplicidade dos modelos de regressão paramétricos os tornam amplamente utilizados em várias áreas. Dentre os inúmeros exemplos, pode-se citar o artigo de Sassi et al. (2012), que apresenta ajustes de modelos de regressão linear múltipla aplicados a dados de conservação de frutas, utilizando os softwares R e STATISTICA.

Entretanto, uma vez que assumem uma forma funcional específica, estes modelos podem apresentar limitações quando são usados para modelar relações que sejam mais flexíveis. Diante disso, surgem os modelos não paramétricos. Os modelos não paramétricos permitem que a relação entre as variáveis que estamos estudando seja determinada diretamente a partir dos dados analisados, podendo assumir formas complexas. Isto faz com que estes modelos sejam especialmente úteis em situações em que a relação entre as variáveis é previamente desconhecida. Os modelos não paramétricos buscam estimar uma curva suave que apresenta o melhor ajuste aos dados ao invés de assumir uma forma matemática rígida e específica. De maneira generalizada, podemos representar a estrutura de um modelo não paramétrico como:

$$\boldsymbol{Y} = g(\boldsymbol{t}) + \boldsymbol{\epsilon} \tag{2.2}$$

em que $\boldsymbol{Y}$ e $\boldsymbol{\epsilon}$ possuem as mesmas definições mencionadas no modelo definido em (2.1) e $g(\boldsymbol{t})$ é o vetor resultante da aplicação da função g aos pontos da variável explicativa $\boldsymbol{t}$. Sendo assim, g é uma função não paramétrica desconhecida que representa a relação entre a variável resposta $\boldsymbol{Y}$ e a variável preditora $\boldsymbol{t}$.

Em situações com mais de uma variável explicativa, existem casos específicos em que uma única variável explicativa relaciona-se de forma desconhecida com a variável resposta, enquanto as demais relacionam-se de forma linear. Logo, temos a combinação de elementos dos modelos paramétricos e não paramétricos em um só modelo, chamado modelo semi-paramétrico. Neste modelo, o objetivo é representar as relações lineares e não lineares entre as variáveis, o que o torna uma abordagem interessante quando queremos analisar as relações complexas e lineares simultaneamente. De forma generalizada podemos representar a estrutura de um modelo semi-paramétrico como:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + g(\boldsymbol{t}) + \boldsymbol{\epsilon} \tag{2.3}$$

onde os termos seguem a mesma definição já vista nos outros modelos apresentados em (2.1) e (2.2). Este modelo é abordado no livro de Ruppert et al. (2003) que explica as técnicas e benefícios da regressão semiparamétrica de forma concisa e modular.

Outro caso específico ocorre quando todas as variáveis preditoras relacionam-se com a variável resposta de forma flexível e desconhecida a priori. Nesta situação, temos o modelo aditivo que se difere do modelo não paramétrico por apresentar $s(s \geq 2)$ covariáveis, enquanto no modelo não paramétrico há apenas uma variável preditora. A estrutura do modelo aditivo pode ser representada de forma bem semelhante ao modelo não paramétrico e é dada por:

$$\boldsymbol{Y} = \sum_{j=1}^{s} g_j(\boldsymbol{t}_j) + \boldsymbol{\epsilon} \tag{2.4}$$

onde g_j é a curva a ser estimada que representa a relação da variável preditora t_j com a variável Y.

Hastie e Tibshirani (1990) e Wood (2017) trazem, além de explicações mais profundas sobre este modelo, algumas aplicações como o uso deste modelo para modelar os efeitos do clima em doenças transmissíveis.

Por fim, temos os modelos aditivos parcialmente lineares (APLMs), em que há p variáveis preditoras que relacionam-se com a variável resposta de forma linear e $s(s \geq 2)$ variáveis explicativas que se relacionam de forma não linear com a variável dependente. A estrutura do modelo aditivo parcialmente linear pode ser representada da seguinte maneira:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \sum_{j=1}^{s} g_j(\boldsymbol{t}_j) + \boldsymbol{\epsilon} \tag{2.5}$$

onde os termos seguem a mesma definição já descrita nos outros modelos (2.1),(2.2),(2.3) e (2.4).

A Figura 1 traz uma representação resumida dos modelos apresentados.

Figura 1 – Modelos em função do número de variáveis lineares e não lineares.

Número de variáveis lineares	Número de variáveis não lineares	Modelo	Representação
nenhuma	uma	não paramétrico	$\boldsymbol{Y} = g(\boldsymbol{t}) + \boldsymbol{\epsilon}$
	duas ou mais	aditivo	$\boldsymbol{Y} = \sum_{j=1}^{s} g_j(\boldsymbol{t}_j) + \boldsymbol{\epsilon}$
uma ou mais	nenhuma	linear	$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$
	uma	semi paramétrico	$\boldsymbol{Y} = \boldsymbol{X\beta} + g(\boldsymbol{t}) + \boldsymbol{\epsilon}$
	duas ou mais	parcialmente linear	$\boldsymbol{Y} = \boldsymbol{X\beta} + \sum_{i=1}^{s} g_j(\boldsymbol{t}_j) + \boldsymbol{\epsilon}$

Fonte: Elaboração do autor.

Assume-se que os erros aleatórios dos modelos estudados possuem distribuição normal com média igual a zero e variância igual a σ^2. Sendo assim, consequentemente, nos APLMs temos que:

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta} + \textstyle\sum_{j=1}^{s} \mathbf{N}_j\boldsymbol{\gamma}_j, \sigma^2\boldsymbol{I}_n)$$

em que $\boldsymbol{N}_j$ é uma matriz de incidência que contém as funções base de B-spline (definidas na Seção 2.1) avaliadas nos pontos de dados t_i com dimensão $n \times q_j$ (onde q_j é o número de funções base dos B-splines); $\boldsymbol{\gamma}_j = (\gamma_1, \ldots, \gamma_{q_j})$ é um vetor coluna com q_j parâmetros a serem estimados para cada variável $\boldsymbol{t}_j$, e $\boldsymbol{I}_n$ é a matriz identidade de dimensão n $\times$ n.

Neste trabalho, o principal interesse é estudar o último modelo apresentado, ou seja, os Modelos Aditivos Parcialmente Lineares (APLMs). A ideia proposta pelo estudo é que ao analisar um conjunto de dados, seja possível identificar quais variáveis preditoras se relacionam de forma linear ou não linear com a variável resposta. Vale ressaltar que este processo de identificação é aplicado somente em variáveis numéricas. Portanto, variáveis categóricas não participam desta etapa de identificação, sendo incluídas diretamente na parte paramétrica dos modelos ajustados.

Uma vez que as variáveis forem identificadas, a etapa seguinte é construir um APLM que tenha a parte paramétrica formada pelas covariáveis com relações lineares e a parte não paramétrica formada pelas variáveis cuja relação se dá de forma flexível e desconhecida a priori. Entretanto, como já apresentado, há casos particulares dos APLMs que se diferenciam pelo número de covariáveis com relações lineares e não lineares. Sendo assim, sabe-se que é possível em um exemplo de aplicação uma única covariável ser identificada com uma relação não linear e as demais como lineares, tornando ideal a construção de um modelo semi-paramétrico para representar os dados, por exemplo.

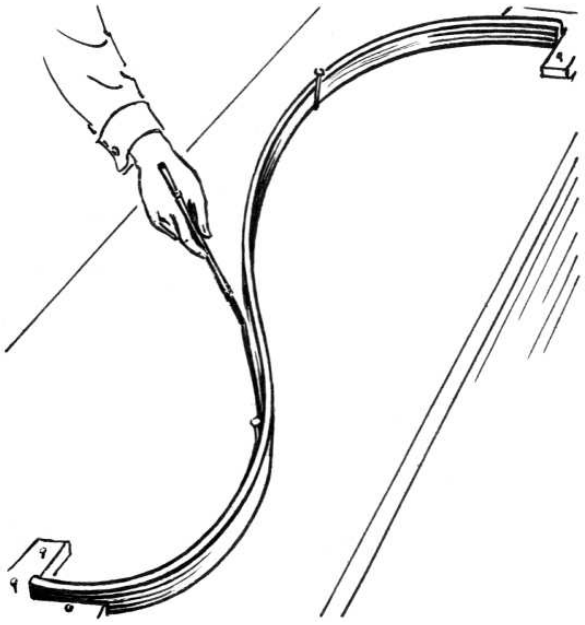
Nos modelos apresentados anteriormente, é comum a aplicação da regressão polinomial, que consiste no uso de um polinômio de baixo grau para modelar a parte não

paramétrica. Essa técnica, apesar de popular, apresenta algumas desvantagens. Uma delas é que observações individuais podem exercer influência, de maneira inesperada, em partes remotas da curva. Outro ponto que pode ser desvantajoso em alguns casos é que os polinômios de um determinado grau formam um conjunto de modelos finitos e de dimensão rígida, podendo restringir a forma como a relação realmente ocorre. Ao permitir que os dados determinem o modelo ajustado, espera-se que haja uma maior flexibilidade na curva a ser construída, logo a regressão polinomial, devido a suas desvantagens, passa a ser uma técnica menos interessante, sendo necessário o uso de uma técnica que apresente maior flexibilidade e estabilidade. Diante disso, no presente trabalho, optou-se em usar splines cúbicas penalizadas (ou P-splines), que são uma técnica amplamente aplicada para substituir a interpolação polinomial.

2.1 SPLINES CÚBICAS

O termo "spline" surgiu na engenharia mecânica e era o nome dado à uma régua flexível usada para desenhar curvas suaves e contínuas. A Figura 2 nos mostra o uso de uma spline para obter uma curva suavizada.

Figura 2 – Contorno suave obtido por meio de uma spline.



Fonte: https://$fr.m.wikipedia.org/wiki/Fichier$: $Splines moothing.png$

No século XX, as splines passaram a ser também um conceito matemático, que seria o de uma função construída por polinômios suavemente ligados em pontos denominados nós. A ideia central era construir uma curva que passasse por determinados pontos da

amostra, conhecidos como nós, buscando aproximá-la da forma que melhor representa o comportamento dos dados. Em vez de utilizar um único polinômio para modelar todas as observações, selecionam-se pontos específicos ao longo do intervalo dos dados (os nós) e define-se um polinômio para cada segmento, permitindo assim a representação de padrões mais complexos por meio de polinômios mais simples.

A partir disso, as splines se tornaram uma técnica frequentemente utilizada para substituir a interpolação polinomial, quando esta resultava em soluções com muita oscilação. A ideia era que a partir das splines fosse possível modelar curvas e superfícies em gráficos em que deseja-se obter curvas suavizadas. O uso de splines apresenta diversas vantagens, entre as quais se destacam sua maior flexibilidade no ajuste de modelos em comparação com a regressão linear ou polinomial, a capacidade de capturar padrões atípicos nos dados e a facilidade de ajuste quando a quantidade e a posição dos nós são previamente definidas.

Os B-splines, também conhecidos como splines básicos, foram introduzidos por De Boor (1978). Segundo a definição, esses splines são formados a partir de segmentos de polinômios que se conectam de maneira suave em pontos específicos chamados nós, evitando mudanças inesperadas na função entre esses pontos. O intervalo de dados $[a, b]$ é dividido em segmentos menores, delimitados pelos nós $r_1 < r_2 < \ldots < r_k, k < n$. Em vez de utilizar polinômios de alto grau como nos modelos polinomiais tradicionais, os B-splines ajustam polinômios de baixo grau em cada intervalo definido. Em casos onde esses polinômios possuem grau 3, como no presente trabalho, os splines formados são denominados splines cúbicos. Sendo assim, define-se um spline cúbico como sendo uma curva construída a partir de polinômios cúbicos, os quais são unidos de forma contínua nos nós. Ou seja, nos nós a primeira e a segunda derivadas da função serão contínuas.

Os splines cúbicos são amplamente utilizados nessa abordagem devido às suas ótimas propriedades de aproximação, possibilitando a obtenção de curvas suaves a partir de valores tabelados de uma função. Eles representam a menor ordem para a qual as descontinuidades nos nós são menos perceptíveis. Por esse motivo, há poucas justificativas para o uso de splines de grau superior.

Uma forma de especificar um spline cúbico é dar os 4 coeficientes polinomiais de cada peça cúbica. Assim, um spline de grau m para aproximar uma função g é dado por

$$g(t) = \beta_0 + \beta_1 t + \ldots + \beta_m t^m + \sum_{i=1}^{k} \beta_{m+i}(t - r_i)_+^m,$$

em que k é o número de nós, $r_1, \ldots, r_k$ são os nós fixos e conhecidos, m é o grau do polinômio e $(a)_+^m = a^m$, se $a \geq 0$ e $(a)_+^m = 0$, caso contrário. Em outras palavras, cada função base é 0 para valores $t < r_i$ e aumenta com um polinômio de grau m para valores de $t > r_i$. Os primeiros $m + 1$ termos da expressão acima representam um polinômio de grau m. Esse polinômio fornece a forma global da spline antes de levar em consideração os nós r_i. Os coeficientes β_{m+i} determinam a contribuição de cada $(t - r_i)_+^m$ para a spline.

Uma outra maneira de representar uma spline cúbica, utilizando B-splines, é proposta por De Boor (1978):

$$g(t_i) = \sum_{j=1}^{q} B_j^{(m)}(t_i)\gamma_j, \tag{2.6}$$

sendo $B_j^{(m)}(t_i)$ os B-splines definidos como

$$B_j^{(0)}(t_i) = \begin{cases} 1, & \text{se } r_j \leq t_i \leq r_{j+1} \\ 0, & \text{caso contrário} \end{cases}$$

$$B_j^{(m)}(t_i) = \frac{t_i - r_j}{r_{j+m} - r_j} B_j^{(m-1)}(t_i) + \frac{r_{j+m+1} - t_i}{r_{j+m+1} - r_{j+1}} B_{j+1}^{(m-1)}(t_i)$$

onde t_i é uma variável que contribui de forma não linear no modelo, para $i = 1, 2, ..., n$; $r_1, \ldots, r_k$, são k nós equidistantes; m é o grau do polinômio e $k = q + m + 1$ é o número de nós.

O número de nós a serem usados influencia na qualidade de ajuste do modelo. Ademais, a posição que estes assumem é um assunto amplo e discutido por vários autores, devido às diversas maneiras como isto pode ser feito. Um exemplo disso é que os nós podem ser equidistantes entre si ou não e a posição assumida pelo primeiro nó pode ser o primeiro valor da variável que será ajustada ou onde ocorre o primeiro ponto de mudança. Estas e outras abordagens podem ser encontradas em Yu e Ruppert (2002) e Vanegas e Paula (2016) com mais detalhes. Ainda em relação a escolha dos nós, Green e Silverman (1994) também apresentam um método para o uso de splines cúbicas no processo de ajuste de curvas suaves. A principal diferença entre os métodos de Green e Silverman (1994) e Eilers e Marx (1996) é como os nós são escolhidos e o tipo de penalização aplicada sobre as splines. Eilers e Marx (1996) sugerem que os k nós sejam equidistantes, enquanto Green e Silverman (1994) sugerem que os nós sejam os diferentes valores dos dados t_i ordenados. O método utilizado no presente trabalho é aquele sugerido por Eilers e Marx (1996).

Entretanto, uma desvantagem em utilizar os B-splines com nós equidistantes é que a suavização da curva é influenciada exclusivamente pela quantidade de nós usados no ajuste, que usualmente são valores pequenos, pois ao aumentar o número de nós ocorre o indesejado efeito de sobreajuste aos dados. Diante disso, Eilers e Marx (1996) propõem um método para solucionar esta questão: o uso de splines penalizadas ou P-splines.

Os P-splines são uma técnica estatística que combina B-splines com uma penalização para controlar a suavidade da curva ajustada a um conjunto de dados. A ideia central dos P-splines é fornecer um método flexível para ajustar curvas suavizadas, evitando o problema de sobreajuste, através da introdução de uma função de penalização de diferenças de ordem d, nos coeficientes adjacentes estimados (γ_j) dos B-splines. Utilizando o método de suavização P-splines, é possível atribuir um maior controle sobre a suavidade presente

na função não paramétrica. Mais detalhes sobre a penalização dos splines serão abordados na Seção 2.2.

2.2 PENALIZAÇÃO DA RUGOSIDADE DA CURVA

A penalização de rugosidade da curva é uma técnica utilizada na modelagem para obter curvas suaves. Ao ajustar modelos não paramétricos, há a necessidade de equilibrar a complexidade do modelo e a rugosidade da curva estimada, ou seja, deve-se considerar além do bom ajuste aos dados, a oscilação apresentada pela curva obtida. Este equilíbrio é necessário porque ao apresentar muita oscilação, a curva apresentará uma alta variabilidade e em casos extremos, pode até mesmo apresentar um sobreajuste. Por outro lado, se a curva é extremamente suave, há a perda de informação e, portanto, a curva pode não representar os dados de forma adequada. A penalização de rugosidade da curva ajuda a atingir esse equilíbrio, pois ao aplicá-la, evita-se que ocorra um sobreajuste do modelo aos dados (overfitting), garantindo que a curva modelada seja suficientemente suave e capture apenas as tendências relevantes dos dados, sem se ajustar aos ruídos.

Antes de abordar sobre a penalização, é necessário uma compreensão, ainda que breve, sobre a rugosidade de uma curva. A rugosidade de uma curva refere-se ao grau de ondulação que esta apresenta. De modo que, curvas mais suaves têm menos ondulações, enquanto curvas mais rugosas possuem mais ondulações. Uma das possíveis maneiras de quantificar ou medir a rugosidade de uma curva g, definida entre $[a, b]$ e duas vezes diferenciável, é obtendo sua integral do quadrado da segunda derivada $\int_a^b [g''(t)]^2 dt$. Vale ressaltar que outras medidas poderiam ser usadas para quantificar a rugosidade da curva, como o máximo absoluto de $|g''|$ ou o número de pontos de inflexão, porém a integral do quadrado da segunda derivada é preferida por ser uma medida global e por suas vantagens computacionais.

A penalização abordada por Green e Silverman (1994) consiste em utilizar o método de estimação por mínimos quadrados penalizado pela integral do quadrado da segunda derivada de g. Sendo assim, o interesse é minimizar a função

$$\sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \lambda \int g''(t) dt, \tag{2.7}$$

em que λ é um parâmetro de suavização usado para controlar a penalidade relativa a ser aplicada para produzir uma função suave.

Green e Silverman (1994) apontam uma motivação interessante para a escolha dessa medida, que surge da formalização de um dispositivo mecânico que era frequentemente usado para desenhar curvas suaves, antes dos gráficos computacionais. Eles abordam a seguinte ideia: se um pedaço fino de madeira flexível, chamado de spline, é dobrado na forma do gráfico de $g(t)$, então o termo principal na energia de deformação é proporcional a

$\int_a^b (g''(t))^2 dt$, fornecendo uma base física intuitiva para usar a expressão como uma medida de rugosidade.

Os mesmos autores demonstram que a medida de rugosidade da curva pode ser reescrita da seguinte maneira:

$$\int_a^b (g''(t))^2 dt = \boldsymbol{\gamma}^T \boldsymbol{K} \boldsymbol{\gamma}, \tag{2.8}$$

onde $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ são parâmetros a serem estimados e a matriz $\boldsymbol{K}$ depende exclusivamente dos nós.

O método proposto por Eilers e Marx (1996) sugere usar como penalização diferenças simples aplicadas diretamente aos parâmetros adjacentes dos B-splines. Os autores explicam que a principal vantagem em utilizar o método sugerido por eles é a possibilidade de redução da dimensionalidade do problema, pois dentre outras vantagens, são computacionalmente mais baratos.

Contudo, matematicamente, a matriz de penalização da segunda derivada contínua e a matriz de penalização baseada em diferenças finitas possuem formas semelhantes e levam a ajustes de suavidade equivalentes quando bem parametrizadas. Ou seja, apesar de possuírem formulações diferentes, ambas as abordagens controlam a variação da segunda derivada da spline. Sendo assim, os dois métodos têm como interesse minimizar a função

$$\sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{K} \boldsymbol{\gamma}. \tag{2.9}$$

Porém, a diferença na formulação de cada método resulta em uma construção distinta da matriz $\boldsymbol{K}$. Seguindo a proposta de Eilers e Marx (1996), escolhida para ser aplicada no presente trabalho, a matriz $\boldsymbol{K}$ é definida como uma matriz diagonal de tamanho igual a $m + k$ com as primeiras m diagonais nulas, onde m é o grau do polinômio e k é o número de nós. A matriz $\boldsymbol{K}$ é dada por $\boldsymbol{K} = \boldsymbol{D}_k^T \boldsymbol{D}_k = (\Delta^d)^T \Delta^d$, onde Δ^d é o operador de diferenças de ordem d, para $d = 1, 2, \ldots$. Considerando $d = 2$, o operador Δ^2 assume a seguinte expressão:

$$\Delta^2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots \\ 0 & 1 & -2 & 1 & \cdots \\ 0 & 0 & 1 & -2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Pelo método dos mínimos quadrados penalizados, pode-se enunciar a penalização de rugosidade para a estimativa de curvas da seguinte forma: dada qualquer função g duas vezes diferenciável definida em $[a, b]$, e um parâmetro de suavização $\lambda > 0$, define-se a soma dos quadrados penalizada

$$S_p(g) = \sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{K} \boldsymbol{\gamma} = (\boldsymbol{Y} - \boldsymbol{N}\boldsymbol{\gamma})^T (\boldsymbol{Y} - \boldsymbol{N}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^T \boldsymbol{K} \boldsymbol{\gamma}, \tag{2.10}$$

em que $\boldsymbol{N}$ é uma matriz de dimensão $n \times q$ formada pelos componentes $B_j^{(m)}(t_i)$, ou seja, as funções base do B-spline avaliadas nos pontos t_i. Essas funções são formadas a partir dos nós que são uma combinação de nós internos (gerados entre o mínimo e o máximo dos pontos t_i) e nós adicionais nas extremidades. Esses nós são adicionados para garantir que o spline tenha as propriedades desejadas nas extremidades.

A adição do termo de penalização é uma maneira de garantir que além da qualidade de ajuste da curva aos dados está sendo considerado também a sua rugosidade. Pode-se observar a partir da expressão acima que, se λ assume um alto valor, o principal termo de $S(g)$ é o de penalidade de rugosidade, e portanto, a curva apresentará pouca oscilação devido a alta penalidade aplicada sobre ela. Em contrapartida, se λ assume um baixo valor, o penalização aplicada sobre a curva é pequena, logo esta apresentará mais oscilações. A maneira como o valor do parâmetro de suavização λ pode ser selecionado é abordada na Seção 2.3.

2.3 SELEÇÃO DO PARÂMETRO DE SUAVIZAÇÃO (λ)

A livre escolha do parâmetro de suavização é algo a ser considerado vantajoso no processo de estimação da curva. Sendo assim, há a necessidade de selecioná-lo utilizando métodos automáticos onde os dados determinam qual o melhor valor a ser assumido pelo parâmetro.

O parâmetro de suavização λ controla o grau de penalização aplicado à função não paramétrica durante o processo de estimação. Como já abordado, valores mais altos de λ resultam em uma penalidade mais forte, o que favorece curvas mais suaves e evita overfitting, ou seja, a adaptação excessiva aos dados de treinamento. Sendo assim, no limite, quando $\lambda \to \infty$, a curva possui a tendência em não apresentar oscilações, se aproximando de uma reta. Por outro lado, valores menores de λ reduzem a penalização, permitindo que a função não paramétrica se ajuste de forma mais flexível aos dados. No limite, quando $\lambda \to 0$, a penalização aplicada tende a 0 e, portanto, a curva irá apresentar o maior nível de oscilações possível, causando um sobreajuste aos dados (overfitting).

Portanto, torna-se indispensável que o valor de λ seja aquele que resultará em uma penalização adequada sobre a curva. A escolha do valor de λ é geralmente obtida por meio de métodos automáticos, como a validação cruzada ou métodos de otimização que buscam encontrar o valor que melhor equilibra a suavidade da curva estimada com o ajuste aos dados disponíveis. No presente trabalho, optou-se em utilizar a técnica de validação cruzada do tipo k-folds para obter o valor do parâmetro de suavização.

2.3.1 *Validação cruzada*

A validação cruzada é uma técnica amplamente utilizada para avaliar a capacidade preditiva de um modelo. Este processo envolve a divisão do conjunto de dados em várias

partes (ou folds) e a execução de múltiplos treinamentos e testes do modelo em diferentes partes dos dados. Em cada iteração, um desses subconjuntos é usado para validação enquanto os restantes são utilizados para treinamento. O objetivo é estimar os parâmetros desejados usando os dados de treinamento e calcular a precisão (ou erro) do modelo com os dados de validação. Esse procedimento é repetido várias vezes, alternando ciclicamente o subconjunto usado para validação e os subconjuntos usados para treinamento. No final, seleciona-se o parâmetro que apresenta o menor erro. Para mais detalhes sobre a validação cruzada, ler Hastie et al. (2009), por exemplo.

Supondo que o erro aleatório tem média zero, a verdadeira curva g possui a característica de que se uma observação y for realizada em um ponto t, o valor de $g(t)$ é o melhor preditor de y, considerando o critério de menor erro quadrático médio. Logo, um bom estimador $\hat{g}(t)$ seria aquele que apresentasse um pequeno valor de $\{y - \hat{g}(t)\}^2$ para uma nova observação y no ponto t.

Considerando que o conjunto de dados é dividido em k subconjuntos, para cada fold S_i, as observações pertencentes a S_i, denotadas por $y_{[S_i]}$ e $t_{[S_i]}$, são reservadas como dados de validação, enquanto as demais observações são usadas para estimar a curva denotada por $\hat{g}_{[-S_i]}(t_{[-S_i]}; \lambda)$. Usa-se o valor de λ para o parâmetro de suavização de modo que $\hat{g}_{[-S_i]}(t_{[-S_i]}; \lambda)$ é o minimizador de $\sum_{j \notin S_i}\{y_j - g(t_j)\}^2 + \lambda \int g''^2$.

A partir disso, pode-se avaliar o quão bem o valor de $\hat{g}_{[-S_i]}(t_{[S_i]}; \lambda)$ prediz as observações $y_{[S_i]}$, obtendo a eficácia de λ que pode ser quantificada pela função da validação cruzada dada por

$$CV(\lambda) = n^{-1} \sum_{i=1}^{k} \sum_{j \in S_i} \{y_j - \hat{g}_{[-S_i]}(t_j; \lambda)\}^2. \tag{2.11}$$

A ideia é calcular $CV(\lambda)$ para os diferentes valores que λ assume dentro de um intervalo estabelecido pelo usuário. Como padrão o algoritmo usado neste trabalho assume que o intervalo para λ é composto por valores entre 0 e 1. Feito isso, o valor de λ selecionado é aquele que minimiza $CV(\lambda)$. Nos modelos em que existem mais de uma covariável que contribui de forma não linear, λ deixa de ser um escalar e passa a ser um vetor de valores, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)$, em que λ_s é o parâmetro de suavização da curva que representa a relação entre a variável Y e a covariável t_s. Neste contexto, o processo de seleção dos parâmetros para cada curva é semelhante aos casos em que há apenas uma covariável não linear. A diferença é que para a minimização de CV(λ), são testadas as combinações de s valores simultaneamente.

Até o momento foram apresentadas as técnicas e metodologias aplicadas para construir a parte não paramétrica dos APLMs. Dada a importância de também se compreender como a parte paramétrica destes modelos é construída, na Seção 2.4 serão apresentados os procedimentos aplicados para calcular os EMVPs para os APLMs.

2.4 ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA PENALIZADA

Uma das finalidades deste trabalho é a obtenção dos estimadores de máxima verossimilhança penalizada (EMVPs) para modelos aditivos parcialmente lineares. E uma vez obtidos, a verificação das propriedades destes estimadores por meio de um estudo de simulação, que será apresentado na Seção 4.1. Para isso, foi utilizado o método usual de obter os EMVPs dos parâmetros em que deriva-se o logaritmo da função de verossimilhança penalizada de $\boldsymbol{\theta}$, denotado por $\ell_p(\boldsymbol{\theta})$, e em seguida iguala-se a expressão obtida à zero.

Temos que $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$, em que $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s)$ são os parâmetros a serem estimados no modelo. Sob suposição de normalidade dos erros aleatórios, a função de log-verossimilhança associada ao modelo é dada por

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_s, \sigma^2) \approx -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{s} \mathbf{N}_j \boldsymbol{\gamma}_j)^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{s} \mathbf{N}_j \boldsymbol{\gamma}_j). \quad (2.12)$$

Contudo, maximizar esta expressão sem impor restrições sobre as funções $g_1, ..., g_s$ pode causar o efeito indesejado de overfitting. Um procedimento para evitar este problema é a estimação de máxima verossimilhança penalizada, que consiste em incorporar em $\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_s)$ uma função penalty, definida na Seção 2.2. Assim, a função de log-verossimilhança penalizada é representada da seguinte maneira:

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_s, \sigma^2, \boldsymbol{\lambda}) \approx \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_s, \sigma^2) - 1/2 \sum_{j=1}^{s} \lambda_j \boldsymbol{\gamma}_j^T \mathbf{K}_j \boldsymbol{\gamma}_j, \quad (2.13)$$

em que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)$ são os parâmetros de suavização para cada curva e $\mathbf{K}_j$ é a matriz penalty obtida como função apenas dos nós.

Uma vez obtidas as expressões analíticas para os EMVPs, implementa-se um algoritmo computacional iterativo por meio do qual chega-se às estimativas de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_s)$, os quais são dados pelas expressões:

$$\widehat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y} - \sum_{j=1}^{s} \mathbf{N}_j \widehat{\boldsymbol{\gamma}}_j^{(k)}), \quad (2.14)$$

$$\widehat{\boldsymbol{\gamma}_i}^{(k)} = (\mathbf{N}_i^T\mathbf{N}_i + \widehat{\sigma^2}^{(k)}\lambda_i\mathbf{K}_i)^{-1}\mathbf{N}_i^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} - \sum_{j=1, j\neq i}^{s} \mathbf{N}_j \widehat{\boldsymbol{\gamma}}_j^{(k)}), \quad i = 1, ..., s, \quad (2.15)$$

$$\widehat{\sigma^2}^{(k)} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} - \sum_{j=1}^{s} \mathbf{N}_j \widehat{\boldsymbol{\gamma}}_j^{(k)})^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)} - \sum_{j=1}^{s} \mathbf{N}_j \widehat{\boldsymbol{\gamma}}_j^{(k)}). \quad (2.16)$$

Como valores iniciais das estimativas utiliza-se:

$$\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \quad (2.17)$$

$$\widehat{\boldsymbol{\gamma}_j}^{(0)} = (\mathbf{N}_j^T\mathbf{N}_j + \widehat{\sigma^2}^{(0)}\lambda_j\mathbf{K}_j)^{-1}\mathbf{N}_j^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(0)}), \quad j = 1, ..., s, \quad (2.18)$$

$$\widehat{\sigma^2}^{(0)} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(0)})^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(0)}). \tag{2.19}$$

As iterações são repetidas até que uma regra de convergência apropriada seja atendida. A regra de convergência consiste em verificar se a norma dos parâmetros $\sqrt{\sum_i(\theta_i^{(k+1)} - \theta_i^{(k)})^2}$ é suficientemente menor que 10^{-5} ou se 5000 iterações já tenham sido realizadas. Quando uma dessas condições é satisfeita, o algoritmo iterativo é interrompido e assim, tem-se as estimativas dos parâmetros.

Além das expressões para os EMVPs, é importante obter também as expressões para calcular a variância destes estimadores. Para isso, utiliza-se a matriz de informação de Fisher, denotada por MI. Ela mede a quantidade de informação que uma amostra carrega sobre os parâmetros de um modelo estatístico.

2.5 MATRIZ DE INFORMAÇÃO DE FISHER

Uma das propriedades apresentadas pelos EMV é que eles são assintoticamente normais. Isso significa que, conforme o tamanho da amostra n aumenta, a distribuição do estimador de máxima verossimilhança se aproxima de uma distribuição normal. Especificamente, temos que

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, MI(\boldsymbol{\theta}_0)^{-1}),$$

em que $MI(\boldsymbol{\theta}_0)$ é a Matriz de Informação de Fisher avaliada no verdadeiro valor do parâmetro $\boldsymbol{\theta}_0$. Entretanto, o valor verdadeiro do parâmetro é desconhecido, logo, a matriz de informação de Fisher é, na prática, avaliada em $\hat{\boldsymbol{\theta}}$. Isto pode ser aplicado, pois assintoticamente $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta_0}, MI(\boldsymbol{\theta}_0)^{-1})$, ou seja, a medida que o tamanho amostral aumenta, $\hat{\boldsymbol{\theta}}$ tende ao valor verdadeiro do parâmetro.

Sendo assim, a matriz de informação de Fisher pode ser usada para calcular a variância dos estimadores de máxima verossimilhança e então, os erros padrão (ep) associados às estimativas de $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_j$ e σ^2 podem ser obtidos por meio da relação

$$ep(\boldsymbol{\theta}) = diag(MI(\hat{\boldsymbol{\theta}})^{-1})^{1/2}. \tag{2.20}$$

Calcular os erros padrão dos estimadores é fundamental na análise estatística porque eles fornecem uma medida da precisão dos estimadores e da incerteza associada às estimativas. No caso dos modelos que apresentam variáveis não lineares, como os APLMs, o cálculo dos erros padrão é importante não somente para avaliação dos parâmetros da parte paramétrica, bem como na construção das bandas de confiança das curvas estimadas. Como visto, a curva estimada $\hat{g}_j(\boldsymbol{t}_j)$ pode ser obtida da seguinte forma

$$\hat{g}_j(\boldsymbol{t}_j) = \boldsymbol{N}_j\hat{\boldsymbol{\gamma}}_j. \tag{2.21}$$

Como resultado, temos que:

$$cov(\hat{g}_j(\boldsymbol{t}_j)) = \boldsymbol{N}_j cov(\hat{\boldsymbol{\gamma}}_j)\boldsymbol{N}_j^T, \tag{2.22}$$

$$var(\hat{g}_j(\boldsymbol{t}_j)) = diag(\boldsymbol{N}_j cov(\hat{\boldsymbol{\gamma}}_j)\boldsymbol{N}_j^T), \tag{2.23}$$

$$ep(\hat{g}_j(\boldsymbol{t}_j)) = \sqrt{var(\hat{g}_j(\boldsymbol{t}_j))}. \tag{2.24}$$

A construção das bandas com $(1-\alpha)\%$ de confiança para a curva $\hat{g}_j(\boldsymbol{t}_j)$ é dada por

$$IC\Big(g_j(\boldsymbol{t}_j), 1-\alpha\Big) = \Big(\hat{g}_j(\boldsymbol{t}_j) \pm z_{\alpha/2} \times ep(\hat{g}_j(\boldsymbol{t}_j)\Big), \tag{2.25}$$

em que $\hat{g}_j(\boldsymbol{t}_j)$ é o valor previsto pela curva ajustada no ponto $\boldsymbol{t}_j$; $z_{\alpha/2}$ é o quantil da distribuição normal padrão, que depende do nível de confiança $1-\alpha$ desejado; e $ep(\hat{g}_j(\boldsymbol{t}_j))$ é o erro padrão da previsão no ponto $\boldsymbol{t}_j$. Vale destacar que, como nos modelos APLMs tem-se uma quantidade significativa de parâmetros a serem estimados, os dados usados em estudos de simulações e aplicações devem possuir um tamanho amostral sufucientemente grande ($n \geq 30$). Sendo assim, pode-se utilizar a distribuição normal para construir os intervalos de confiança, mesmo sem conhecer a distribuição exata do parâmetro, aplicando os resultados do teorema central do limite.

2.6 MATRIZ DE INFORMAÇÃO OBSERVADA

Temos que: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s, \sigma^2)$. Sendo assim, a matriz de informação de Fisher observada é construída de seguinte forma:

$$MI_{obs}(\boldsymbol{\theta}_0) = \begin{bmatrix} -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}_1^T} & \cdots & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}_s^T} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\sigma^2} \\ -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\beta}^T} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\gamma}_1^T} & \cdots & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\gamma}_s^T} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_1\partial\sigma^2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_s\partial\boldsymbol{\beta}^T} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_s\partial\boldsymbol{\gamma}_s^T} & \cdots & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_s\partial\boldsymbol{\gamma}_s^T} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_s\partial\sigma^2} \\ -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\sigma^2\partial\boldsymbol{\beta}} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\sigma^2\partial\boldsymbol{\gamma}_1} & \cdots & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\sigma^2\partial\boldsymbol{\gamma}_s} & -\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\sigma^2\partial\sigma^2} \end{bmatrix} \tag{2.26}$$

2.7 MATRIZ DE INFORMAÇÃO ESPERADA

No presente trabalho, foi implementada a matriz de informação esperada, com o objetivo de facilitar os cálculos e diminuir o custo computacional em obter os resultados. A matriz de informação esperada é construída calculando-se a esperança de cada um dos termos da matriz de informação observada, ou seja,

$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}) = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$,

$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}_j^T}) = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{N}_j$, $j = 1, \ldots, s$,

$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\sigma^2}) = \underset{\sim}{\boldsymbol{0}}$,

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\beta}^T}) = \frac{1}{\sigma^2}(\boldsymbol{X}^T \boldsymbol{N}_j)^T,\ j = 1, \ldots, s,$$

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\gamma}_j^T}) = \frac{1}{\sigma^2}\boldsymbol{N}_j^T \boldsymbol{N}_j + \lambda_j \boldsymbol{K}_j,\ j = 1, \ldots, s,$$

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\gamma}_i^T}) = \frac{1}{\sigma^2}\boldsymbol{N}_j^T \boldsymbol{N}_i, \text{ para } i \neq j \text{ e } i, j = 1, \ldots, s,$$

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_j \partial \sigma^2}) = \underset{\sim}{\mathbf{0}},\ j = 1, \ldots, s,$$

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\beta}}) = \underset{\sim}{\mathbf{0}},$$

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\gamma}_j}) = \underset{\sim}{\mathbf{0}},\ j = 1, \ldots, s,$$

$$E(-\frac{\partial^2 l_p(\boldsymbol{\theta})}{\partial \sigma^2 \partial \sigma^2}) = \frac{n}{2\sigma^4}\ .$$

Como pode-se observar o valor esperado de alguns termos é zero, o que torna o cálculo da matriz menos oneroso.

Além da implementação da matriz de informação de Fisher esperada, foi aplicado como método alternativo, a técnica de Bootstrap para a obtenção dos erros padrão das estimativas dos parâmetros em casos em que a matriz de informação não era inversível. Esta técnica consiste em fazer reamostragens dos erros aleatórios sob os parâmetros estimados, e a partir das novas estimativas obtidas nestas reamostragens, calcula-se as variâncias associadas a cada uma delas. Para maiores informações sobre a técnica Bootstrap, ver Davison e Hinkley (1997).

Vale destacar que as expressões encontradas para os EMVPs dos parâmetros dos APLMs, apresentados na Seção 2.4, bem como as matrizes de informação de Fisher observada e esperada foram calculadas a partir da função da log-verossimilhança, aplicando técnicas como derivadas de primeira e segunda ordem para encontrá-las. Isto foi necessário, pois na literatura explorada as expressões dos EMVPs para estes modelos não são encontradas. Com isso, o presente trabalho contribui de forma significativa no processo de construção dos APLMs.

Uma outra significativa contribuição que este estudo busca oferecer é um método que permita selecionar as covariáveis de um APLM, identificando quais variáveis formam a parte paramétrica e quais formam a parte não paramétrica do modelo. Sendo assim, é apresentado no próximo capítulo a metodologia desenvolvida para alcançar este objetivo proposto.

3 SELEÇÃO DE VARIÁVEIS

A principal proposta deste trabalho é apresentar uma metodologia que permita a seleção de APLMs, identificando quais covariáveis pertencem a parte paramétrica e quais pertencem a parte não paramétrica do modelo. Como já visto, pode-se construir diferentes modelos de acordo com o número de variáveis preditoras cujo as relações com a variável resposta são consideradas não lineares. Diante disso, é importante explorar o tipo de relação presente entre as covariáveis e a variável resposta do modelo. Uma das maneiras mais comuns para inicialmente explorar essas relações é analisando os gráficos de dispersão das variáveis preditoras com a variável resposta e também os coeficientes de correlação, onde uma correlação forte pode indicar uma relação linear, embora isso não seja garantido.

Sendo assim, uma outra alternativa é proposta no presente trabalho. Trata-se de identificar o tipo de relação, linear ou não linear, por meio da análise das derivadas de segunda ordem das curvas estimadas por um modelo aditivo. Esta proposta, até o presente momento, não foi explorada nos trabalhos literários que abordam modelos como o APLM. Sendo assim, trata-se de uma abordagem inovadora que contribui para a ampliação do conhecimento na área. Além disso, sua aplicação pode trazer avanços significativos na modelagem estatística, especialmente no contexto dos APLMs. Essa proposta não apenas preenche uma lacuna existente na literatura, como também pode oferecer uma alternativa mais eficiente para a análise de dados complexos.

A ideia consiste em primeiramente ajustar um modelo aditivo, ou seja, assume-se que todas as covariáveis contribuem de forma não linear para explicar a variável resposta. Em seguida, para cada curva estimada, obtem-se também a sua derivada de segunda ordem. Sabe-se que, nos casos em que a variável explicativa possui uma relação linear, a curva estimada no modelo aditivo é na realidade algo semelhante à uma reta. Ao obter a segunda derivada desta reta, espera-se que esta seja uma reta com valor zero em todos os pontos, pois como resultado temos que a derivada de segunda ordem de uma reta genérica é sempre zero. Nos casos em que a variável explicativa apresenta uma relação não linear, espera-se que a segunda derivada seja uma curva correspondente à segunda derivada da relação não linear existente entre a variável em questão e a variável resposta. A maneira como as curvas da segunda derivada são obtidas e o critério proposto para considerá-las como retas em zero ou curvas serão abordados com mais detalhes na Seção 3.1.

Após identificar quais covariáveis contribuem de forma linear, um segundo modelo é ajustado, agora considerando as variáveis identificadas na parte paramétrica do modelo e as demais na parte não paramétrica. Feito isso, o primeiro modelo (aditivo) e este segundo modelo são comparados a fim de verificar se houve alguma melhora no ajuste aos dados. Neste trabalho, optou-se em aplicar o teste de razão de verossimilhanças e o BIC para

comparar os modelos e encontrar aquele que ajusta-se melhor aos dados.

3.1 DERIVADA DE SEGUNDA ORDEM DAS CURVAS ESTIMADAS

A análise das curvas das derivadas de segunda ordem de cada covariável é uma etapa essencial no processo de seleção de variáveis. Esse procedimento visa classificar a relação entre cada covariável e a variável resposta, identificando se essa relação é linear ou não linear. Como já visto, as curvas do modelo são construídas a partir da relação $\hat{g}_j(\boldsymbol{t}_j) = \boldsymbol{N}_j\hat{\boldsymbol{\gamma}}_j$, em que $\boldsymbol{N}_j$ é a matriz de incidência e cada $\boldsymbol{\gamma}_j$ é um vetor coluna com k_j parâmetros a serem estimados para cada variável $\boldsymbol{t}_j$. A função utilizada para construir a matriz $\boldsymbol{N}$ (nomeada "splineDesign" do pacote {splines} do R Core Team (2024)) permite calcular tanto os valores das funções base (polinômios de B-spline) nos pontos t_i, quanto os valores de suas derivadas. Esses cálculos podem ser realizados especificando o argumento "derivs", que define a ordem da derivada a ser avaliada. A matriz obtida com derivs assumindo o valor zero, é aquela denotada por $\boldsymbol{N}$, usada para estimar as curvas do modelo. Caso o parâmetro da função assuma o valor dois, por exemplo, a matriz resultante, denotada por $\boldsymbol{N}''$, possui os valores da segunda derivada das funções base, permitindo a obtenção da segunda derivada das curvas estimadas por meio da expressão

$$\hat{g_j}''(\boldsymbol{t}_j) = \boldsymbol{N}''_j\hat{\boldsymbol{\gamma}}_j. \tag{3.1}$$

Este resultado não foi encontrado, até o presente momento, na literatura explorada para a base teórica do presente trabalho, sendo obtido por meio de testes empíricos feitos durante a pesquisa. Um estudo de simulação foi realizado e é apresentado na Seção 4.2, comprovando de maneira empírica esta relação para a obtenção da segunda derivada das curvas estimadas. Para cada curva $\hat{g_j}''(\boldsymbol{t}_j)$ é possível determinar as bandas de confiança de maneira semelhante a forma como é feito para as curvas $\hat{g}_j(\boldsymbol{t}_j)$ do modelo. O erro padrão da derivada de segunda ordem pode ser calculado como

$$cov(\hat{g_j}''(\boldsymbol{t}_j)) = \boldsymbol{N}''_j cov(\hat{\boldsymbol{\gamma}}_j)\boldsymbol{N}''^T_j, \tag{3.2}$$

$$var(\hat{g_j}''(\boldsymbol{t}_j)) = diag(\boldsymbol{N}''_j cov(\hat{\boldsymbol{\gamma}}_j)\boldsymbol{N}''^T_j), \tag{3.3}$$

$$ep(\hat{g_j}''(\boldsymbol{t}_j)) = \sqrt{var(\hat{g_j}''(\boldsymbol{t}_j))}. \tag{3.4}$$

Sendo assim, a construção das bandas com $(1-\alpha)\%$ de confiança para a curva $\hat{g}''_j(\boldsymbol{t}_j)$, é dada por:

$$IC\Big(g''_j(\boldsymbol{t}_j), 1-\alpha\Big) = \Big(\hat{g}''_j(\boldsymbol{t}_j) \pm z_{\alpha/2} \times ep(\hat{g}''_j(\boldsymbol{t}_j))\Big), \tag{3.5}$$

em que $\hat{g}_j''(\boldsymbol{t}_j)$ é o valor previsto pela curva da segunda derivada no ponto $\boldsymbol{t}_j$; $z_{\alpha/2}$ é o quantil da distribuição normal padrão que depende do nível de confiança do intervalo, e $ep(\hat{g}_j''(\boldsymbol{t}_j))$ é o erro padrão da previsão da derivada de segunda ordem no ponto $\boldsymbol{t}_j$.

Para cada $\hat{g}_j''(\boldsymbol{t}_j)$, é avaliado se as bandas de confiança da curva contêm o valor real zero em no mínimo $(1-\alpha)\%$ dos pontos. Caso isto ocorra, a segunda derivada $g_j''(\boldsymbol{t}_j)$ é considerada uma reta em zero e, portanto, a covariável correspondente é selecionada como tendo uma relação linear. Caso contrário, a segunda derivada $g_j''(\boldsymbol{t}_j)$ não é considerada uma reta em zero e a covariável correspondente é selecionada como tendo uma relação não linear com a variável resposta. Na prática, quando os percentuais de pontos das bandas que contêm o valor zero estão acima de 85%, é razoável optar em considerar a variável como sendo pertencente à parte paramétrica do modelo e comparar os modelos pelo teste de razão de verossimilhanças. A partir do resultado do teste, o usuário pode decidir o que deve ser considerado em relação à variável testada. Feito este processo para todas as variáveis preditoras numéricas, pode-se construir o modelo final de acordo com o número de variáveis classificadas como tendo relações não lineares.

3.2 TESTE DA RAZÃO DE VEROSSIMILHANÇAS

O teste de razão de verossimilhanças é uma técnica estatística utilizada para avaliar e comparar dois modelos estatísticos, normalmente um modelo mais básico (modelo reduzido) e um modelo mais complexo (modelo alternativo). A finalidade é verificar se o modelo mais complexo se ajusta significativamente melhor aos dados em comparação ao modelo mais simples. A partir disso, pode-se considerar o seguinte teste de hipóteses:

$$H_0 : \text{Os dados são melhor explicados pelo modelo reduzido.}$$
$$H_1 : \text{Os dados são melhor explicados pelo modelo alternativo.}$$

A partir disso, para cada um dos modelos, calcula-se a função de verossimilhança e então, determina-se a estatística de teste que é dada por:

$$\Lambda = -2ln\left(\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right),$$

em que $L(\hat{\boldsymbol{\theta}}_0)$ é a função de verossimilhança do modelo reduzido e $L(\hat{\boldsymbol{\theta}})$ é a função de verossimilhança do modelo alternativo (complexo). No caso em que compara-se modelos penalizados, como no presente trabalho, a estatística de teste é a mesma, porém as funções de verossimilhanças dos modelos comparados são penalizadas.

Em modelos clássicos, a estatística Λ segue, assintoticamente, uma distribuição qui-quadrado com graus de liberdade iguais à diferença (v) no número de parâmetros entre os modelos. Sendo assim, $\Lambda \sim \chi^2_v$ e, portanto, rejeita-se o teste em α quando a estatística de teste é maior que o quantil de ordem $1-\alpha$ da distribuição χ^2_v em v graus de liberdade.

No entanto, em modelos com penalização, o conceito de número de parâmetros deve ser substituído por graus de liberdade efetivos, que representam a complexidade real do modelo ajustado, podendo inclusive assumir valores fracionários. Tais graus de liberdade são obtidos a partir da estrutura do modelo penalizado e refletem o ajuste e a suavização induzida pela penalização. Os cálculos dos graus de liberdade para os modelos neste contexto serão abordados na Seção 3.3.

Além disso, quando se utilizam modelos penalizados ou mais complexos, a distribuição assintótica da estatística de teste pode ser afetada pela penalização, e a aproximação por uma distribuição qui-quadrado pode não ser adequada. Neste trabalho, inicialmente considerou-se essa aproximação, mas, diante dos resultados obtidos (apresentados na Seção 4.2), é proposta uma abordagem alternativa via simulação Bootstrap para construção da distribuição empírica da estatística sob a hipótese nula. Mais detalhes serão apresentados na seção dos estudos de simulação.

3.3 GRAUS DE LIBERDADE

Nos modelos paramétricos, os graus de liberdade do modelo podem ser obtidos a partir do traço da matriz chapéu, definida por $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. O traço dessa matriz, $tr(\boldsymbol{H})$, representa o número efetivo de parâmetros estimados pelo modelo. Na regressão linear, esse valor coincide com o número de parâmetros ajustados, isto é, p, ou seja, o comprimento do vetor $\boldsymbol{\beta}$. Em modelos onde há componentes lineares e não lineares, $p+s$ seriam os graus de liberdade totais, onde p é o número de parâmetros no componente linear (paramétrico) e s é o número de parâmetros no componente não linear (não paramétrico). No entanto, devido à penalização na função de verossimilhança, o "custo" real em termos de graus de liberdade efetivos pode ser menor que $p+s$.

Para determinar os graus de liberdade do modelo ($df(\lambda)$), Hastie e Tibshirani (1990) sugerem calcular o traço da matriz suavizadora $\boldsymbol{H}(\lambda)$. A matriz $\boldsymbol{H}(\lambda)$ pode ser decomposta em duas partes:

$$\boldsymbol{H}(\lambda) = \boldsymbol{H}_{\boldsymbol{\beta}}(\lambda) + \boldsymbol{H}_{\boldsymbol{\gamma}}(\lambda).$$

Logo,

$$df(\lambda) = tr\{\boldsymbol{H}_{\boldsymbol{\beta}}(\lambda)\} + tr\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\lambda)\},$$

em que $tr\{\boldsymbol{H}_{\boldsymbol{\beta}}(\lambda)\}$ corresponde aos graus de liberdade referentes à parte paramétrica do modelo e, segundo Eubank (1999), $tr\{\boldsymbol{H}_{\boldsymbol{\beta}}(\lambda)\} \approx p$. Já o termo $tr\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\lambda)\}$ corresponde aos graus de liberdade da parte não paramétrica do modelo. Diante destes resultados, pode-se definir $df(\lambda) = 1 + p + df_{\boldsymbol{\gamma}}(\lambda)$, em que 1 é para o parâmetro de dispersão σ^2, p para os parâmetros $\boldsymbol{\beta}$ e $df_{\boldsymbol{\gamma}}(\lambda)$ para os parâmetros $\boldsymbol{\gamma}$, da parte não paramétrica. Apesar de

nesta definição de $df(\lambda)$ ser considerado apenas uma variável com relação não linear, esta pode ser generalizada para os APLMs com s variáveis com relações não lineares, tal que

$$df(\lambda) = 1 + p + df_{\boldsymbol{\gamma}_1}(\lambda_1) + \ldots + df_{\boldsymbol{\gamma}_s}(\lambda_s).$$

Isso permite entender melhor como cada parte do modelo contribui para os graus de liberdade totais.

Eilers e Marx (1996) apresentam que

$$df_{\gamma}(\lambda_j) = tr\{\boldsymbol{N_j}(\boldsymbol{N_j}^T\boldsymbol{N_j} + \sigma^2\lambda_j\boldsymbol{K_j})^{-1}\boldsymbol{N_j}^T\}, j = 1, \ldots, s,$$

onde $\boldsymbol{N_j}, \boldsymbol{K_j}$ são matrizes já definidas anteriormente. Com isso, pode-se calcular os graus de liberdade dos APLMs e seus casos particulares.

4 SIMULAÇÃO

Neste capítulo é apresentado o estudo de simulação desenvolvido que foi dividido em duas partes: a primeira consiste em uma simulação feita com o objetivo de verificar as propriedades assintóticas dos EMVPs. Na segunda etapa, a ideia é apresentar como a segunda derivada das curvas estimadas são obtidas e uma vez representadas graficamente, como podemos avaliá-las de forma a selecionar quais variáveis compõem a parte linear e quais compõem a parte não linear do APLM.

4.1 VERIFICAÇÃO DAS PROPRIEDADES ASSINTÓTICAS DOS EMVPs

Esta etapa do estudo tem como objetivo verificar as propriedades assintóticas como a consistência e a normalidade dos EMVPs obtidos para um APLM. A simulação foi realizada usando o método de Monte Carlo com 1000 replicações para 3 diferentes tamanhos de amostra: 500, 1000 e 2000. O modelo simulado é composto de duas variáveis explicativas que apresentam relações lineares e duas variáveis explicativas com relações não lineares:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + g_1(t_{1_i}) + g_2(t_{2_i}) + \epsilon_i, \quad i = 1, ..., n,$$

em que $X \sim U(2\pi, 6\pi)$, ou seja, as variáveis explicativas da parte paramétrica do modelo, X_1 e X_2, possuem distribuição uniforme; os parâmetros $\boldsymbol{\beta} = (\beta_1, \beta_2)$ assumem os valores reais $\beta_1 = 3, \beta_2 = 4$; a parte não paramétrica é composta por duas curvas $g_1(t_1) = cos(t_1)$ (cosseno) e $g_2(t_2) = cos(4\pi t_2)e^{-t_2^2/2}$ (Doppler), sendo $t_1 \sim U(2\pi, 6\pi)$ e $t_2 \sim U(0.6, 1.6)$; e por fim, os erros aleatórios do modelo possuem distribuição normal, $\epsilon_i \sim N(0, 0.1)$, ou seja, $\sigma^2 = 0.1$.

Após a execução do algoritmo utilizando os dados da simulação, obteve-se os seguintes resultados.

Tabela 1 – Médias e erros padrão das estimativas dos parâmetros do modelo.

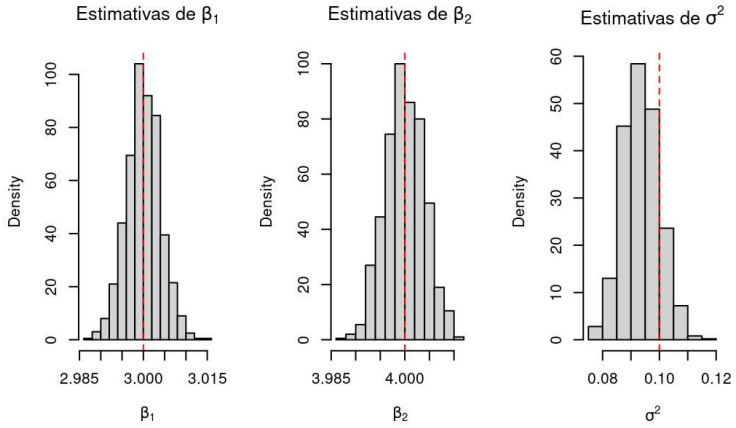
		n=500		n=1000		n=2000	
Parâm.	VR	Est.	DPe	Est.	DPe	Est.	DPe
σ^2	0.1	0.094	0.006	0.097	0.004	0.098	0.003
β_1	3.00	3.000	0.004	3.000	0.003	3.000	0.002
β_2	4.00	4.000	0.004	4.000	0.003	4.000	0.002

A Tabela 1 apresenta os valores de estimativas e erros-padrão associados aos parâmetros da parte linear do modelo. Para cada um dos três tamanhos de amostra, tem-se três colunas: média das estimativas do parâmetro (Est) e os desvios-padrão empíricos (DPe), obtidos por meio da técnica Bootstrap, para cada um dos parâmetros estimados no modelo. Analisando os valores apresentados na Tabela 1, percebe-se que as estimativas

obtidas nas simulações evidenciam a consistência assintótica dos EMVPs para o valor verdadeiro dos parâmetros. Além disso, nota-se que à medida que o tamanho amostral cresce, o desvio-padrão empírico das estimativas diminui, evidenciando a consistência dos EMVPs.

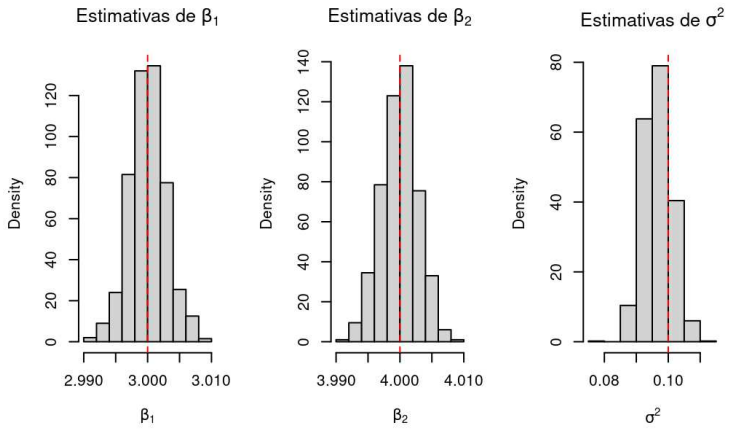
A normalidade assintótica também é uma propriedade dos EMVPs. Foram feitas algumas análises para verificação desta propriedade nos estimadores de cada parâmetro do modelo simulado.

Figura 3 – Histogramas das 1000 estimativas dos parâmetros β_1,β_2 e σ^2, respectivamente, para amostra de tamanho igual a 500.



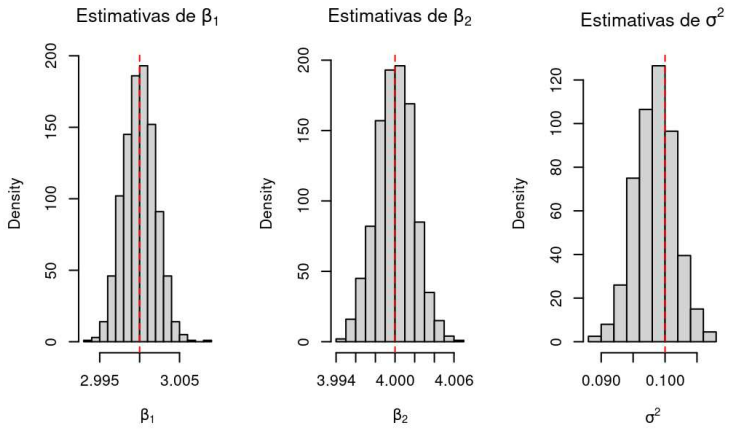
Fonte: Elaboração do autor.

Figura 4 – Histogramas das 1000 estimativas dos parâmetros β_1,β_2 e σ^2, respectivamente, para amostra de tamanho igual a 1000.



Fonte: Elaboração do autor.

Figura 5 – Histogramas das 1000 estimativas dos parâmetros β_1,β_2 e σ^2, respectivamente, para amostra de tamanho igual a 2000.



Fonte: Elaboração do autor.

Observando os histogramas das Figuras 3,4 e 5, percebe-se a propriedade da normalidade assintótica, que no caso dos estimadores de β_1 e β_2 é perceptível desde as amostras de tamanho 500 e para σ^2, a normalidade fica mais evidente à medida que o tamanho amostral aumenta.

Uma outra maneira utilizada para verificar a normalidade dos estimadores foi a aplicação do teste Shapiro-Wilk. Alguns trabalhos encontrados na literatura, como o de Torman et al. (2012), trazem comparações de desempenho deste teste com outros testes de normalidade. Como resultado, o teste de Shapiro-Wilk geralmente apresenta maior eficiência em amostras grandes, o que justifica a escolha da aplicação deste teste neste trabalho. Os p-valores do teste estão na Tabela 2.

H_0 : As estimativas apresentam distribuição normal.
H_1 : As estimativas não apresentam distribuição normal.

Tabela 2 – P-valores dos testes de normalidade dos EMVPs dos parâmetros para cada tamanho de amostra.

	n=500	n=1000	n=2000
Parâmetro	p-valor	p-valor	p-valor
σ^2	0.023	0.040	0.856
β_1	0.949	0.047	0.189
β_2	0.546	0.174	0.681

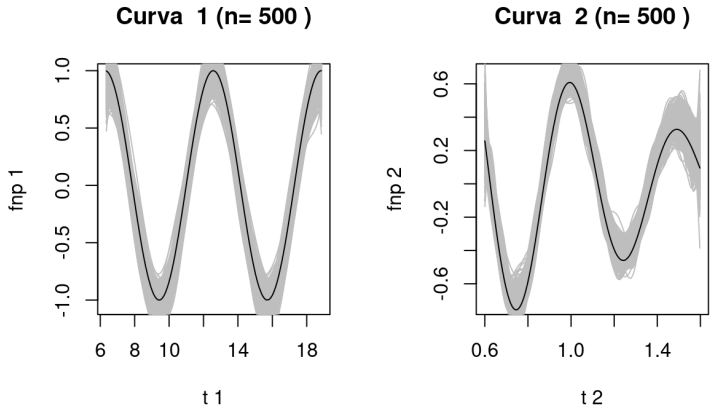
Dado que a hipótese nula do teste é que as estimativas assumem normalidade, a um nível de 5% de significância, a hipótese nula não é rejeitada para as estimativas

do estimador de β_2 nos três diferentes tamanhos de amostra. Já nas estimativas para o estimador de β_1, na amostra de tamanho igual a 1000, o p-valor obtido encontra-se um pouco abaixo de 0.05, o que acarreta na rejeição da hipótese de normalidade neste caso. Entretanto, para os demais tamanhos amostrais a hipótese nula não é rejeitada a um nível de 5% de significância. Portanto, conclui-se que os estimadores para β_1 e β_2 possuem a propriedade de normalidade assintótica.

Para o estimador de σ^2, tem-se que inicialmente a hipótese nula é rejeitada, a um nível de 5% de significância, para n=500 e n=1000. À medida que o tamanho da amostra aumenta (n=2000), a hipótese nula não é mais rejeitada a um nível de 5% de significância, evidenciando a propriedade de normalidade assintótica do estimador de σ^2.

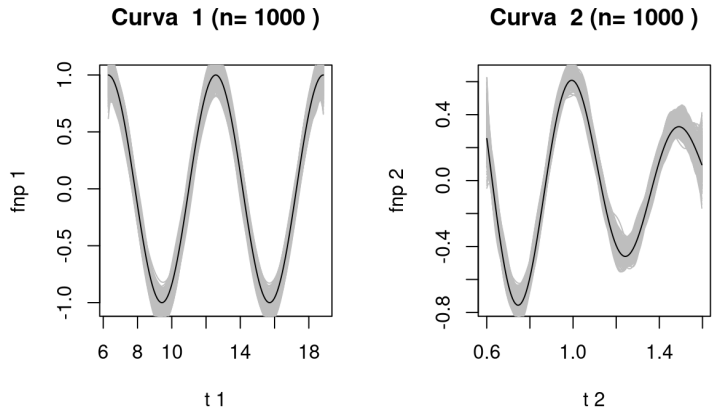
Na parte não paramétrica do modelo, é possível verificar visualmente o comportamento assintótico das curvas estimadas de se aproximarem da curva verdadeira à medida que o tamanho da amostra aumenta. A Figura 6 apresenta os gráficos com as 1000 curvas estimadas da curva $g_1(t_1)$ e $g_2(t_2)$, para o tamanho de amostra igual a 500. As Figuras 7 e 8 apresentam o mesmo, mas para os tamanhos de amostra 1000 e 2000, respectivamente.

Figura 6 – Estimativas não paramétricas para as variáveis t_1 (à esquerda) e t_2 (à direita) para amostra de tamanho igual a 500.



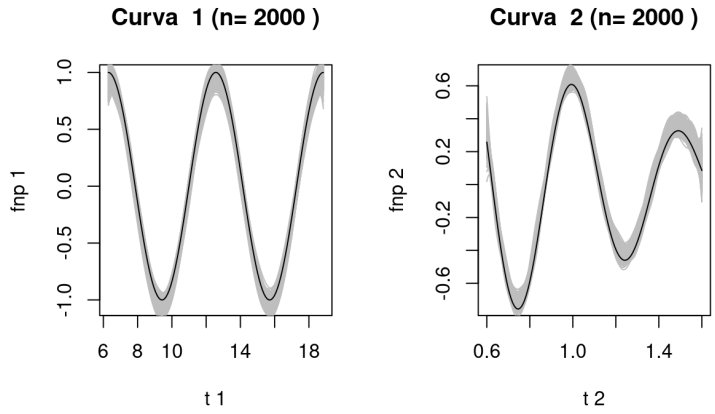
Fonte: Elaboração do autor.

Figura 7 – Estimativas não paramétricas para as variáveis t_1 (à esquerda) e t_2 (à direita) para amostra de tamanho igual a 1000.



Fonte: Elaboração do autor.

Figura 8 – Estimativas não paramétricas para as variáveis t_1 (à esquerda) e t_2 (à direita) para amostra de tamanho igual a 2000.



Fonte: Elaboração do autor.

Pelos gráficos acima, percebe-se que assintoticamente, as 1000 curvas estimadas nas simulações (em cinza) aproximam-se da curva verdadeira (em preto), evidenciando a redução da variabilidade à medida que aumenta-se o tamanho amostral.

4.2 SELEÇÃO DAS VARIÁVEIS COM RELAÇÕES LINEARES E NÃO LINEARES

Nesta segunda etapa da simulação, a ideia geral é conseguir identificar quais variáveis pertecem à parte linear e quais pertencem à parte não linear do modelo. Para

isso, aplicou-se a metodologia discutida anteriormente. Para a construção dos três modelos desta segunda etapa foram usados os mesmos dados da etapa anterior, apresentada na Seção 4.1.

No primeiro modelo, as variáveis denotadas como X_1 e X_2 são as covariáveis da parte paramétrica e as variáveis t_1 e t_2 foram inseridas na parte não paramétrica, o que resultaria em um modelo parcialmente linear

$$\text{modelo 1 (modelo verdadeiro): } Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + g_1(t_{1i}) + g_2(t_{2i}) + \epsilon_i.$$

Estas são as verdadeiras relações simuladas das covariáveis com a variável resposta Y, portanto, sabe-se que este seria o modelo ideal.

No segundo modelo, considerou-se que todas as variáveis relacionam-se de maneira linear com a variável resposta, ou seja, foi construído um modelo de regressão linear múltipla

$$\text{modelo 2: } Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3(t_{1i}) + \beta_4(t_{2i}) + \epsilon_i.$$

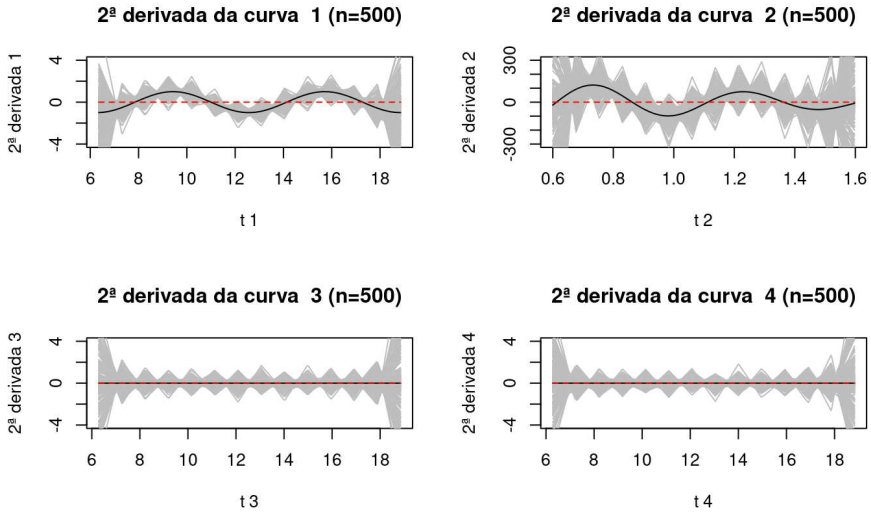
E por fim, um terceiro modelo aditivo, onde todas as variáveis relacionam-se de maneira não linear

$$\text{modelo 3: } Y_i = g_1(t_{1i}) + g_2(t_{2i}) + g_3(X_{1i}) + g_4(X_{2i}) + \epsilon_i.$$

Aplicando a metodologia apresentada, inicialmente avaliou-se as curvas estimadas para as derivadas de segunda ordem das curvas do modelo 3. Como as verdadeiras relações das covariáveis com a variável resposta são aquelas representadas no primeiro modelo, as curvas estimadas pelo modelo 3 são uma curva cosseno em $g_1(t_1)$, uma doppler em $g_2(t_2)$, uma reta em $g_3(X_1)$ e uma reta em $g_4(X_2)$.

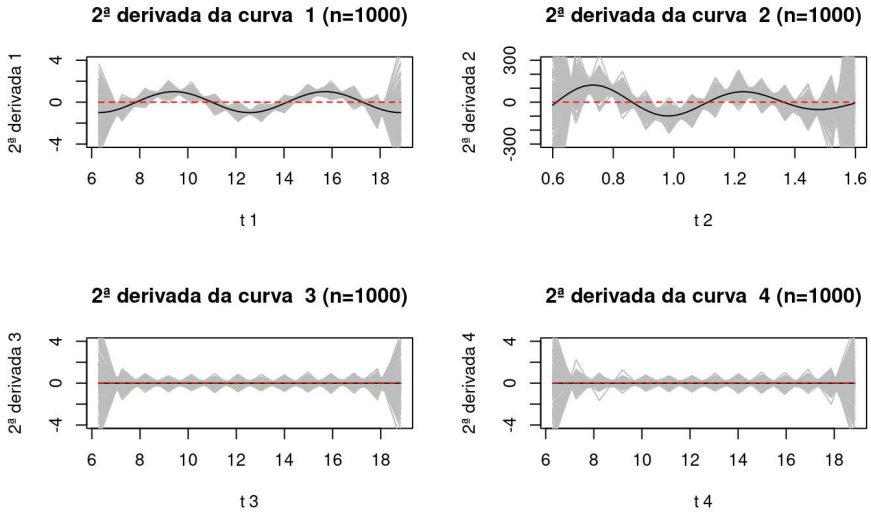
As duas últimas variáveis apresentam uma relação verdadeira com a variável resposta que é dada de forma linear, por isso, o modelo aditivo estima retas para representá-las. Portanto, para estas duas covariáveis, espera-se que as curvas da segunda derivada sejam na verdade algo que assemelha-se à uma reta em torno de zero, já que o valor real da segunda derivada de uma reta qualquer é zero em todos os pontos. A partir da relação $\hat{g_j}''(\boldsymbol{t}_j) = \boldsymbol{N}_j''\hat{\boldsymbol{\gamma}}_j$, foi possível construir as curvas da segunda derivada de cada curva estimada no modelo aditivo. As 1000 réplicas das derivadas de segunda ordem estão apresentadas nas Figuras 9, 10 e 11 para os três diferentes tamanhos de amostra considerados.

Figura 9 – Estimativas da segunda derivada das 1000 curvas estimadas no modelo aditivo (em cinza) e curva real da segunda derivada (em preto).



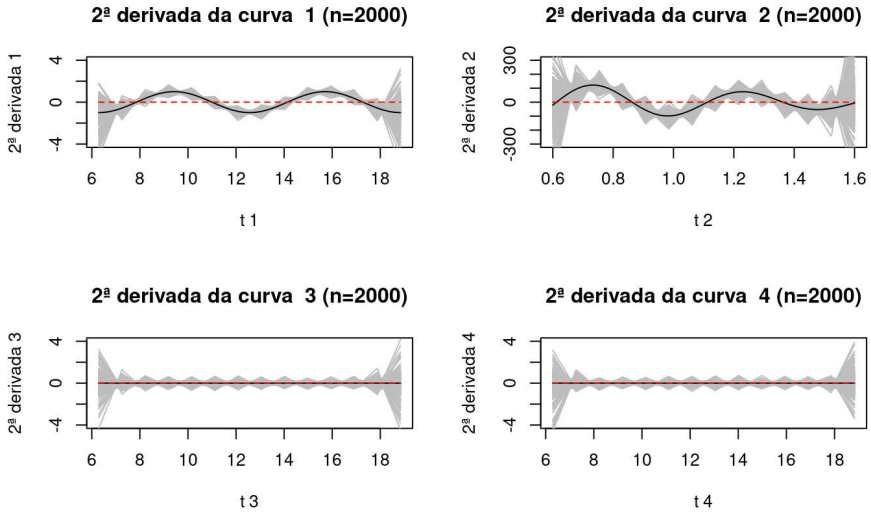
Fonte: Elaboração do autor.

Figura 10 – Estimativas da segunda derivada das 1000 curvas estimadas no modelo aditivo (em cinza) e curva real da segunda derivada (em preto).



Fonte: Elaboração do autor.

Figura 11 – Estimativas da segunda derivada das 1000 curvas estimadas no modelo aditivo (em cinza) e curva real da segunda derivada (em preto).



Fonte: Elaboração do autor.

Como é possível ver, analisando os gráficos apresentados nas figuras, para os três tamanhos de amostra considerados, temos que as curvas da segunda derivada em cinza, assemelham-se à curva real da segunda derivada, representada em preto. No caso da primeira curva, a relação original é $g_1(t_1) = cos(t_1)$, portanto, $g_1''(t_1) = -cos(t_1)$. Para a segunda curva, a segunda derivada real é dada pela expressão: $g_2''(t_2) = exp(-t_2^2/2) * t_2^2 * cos(4 * \pi * t_2) + 8 * \pi * exp(-t_2^2/2) * sin(4 * \pi * t_2) - 16 * \pi^2 * exp(-t_2^2/2) * cos(4 * \pi * t_2) - exp(-t_2^2/2) * cos(4 * \pi * t_2)$. Nas curvas 3 e 4, como a relação real é linear, as curvas estimadas da segunda derivada tendem à uma reta em zero, como esperado. É possível notar que as curvas em cinza apresentam oscilações em torno da curva real e, assintoticamente, essas oscilações são menores, ou seja, assim como nas curvas $g_1(t_1)$ e $g_2(t_2)$, as curvas estimadas da segunda derivada aproximam-se da curva real à medida que o tamanho da amostra aumenta. Ao comparar as curvas estimadas da segunda derivada com a reta em zero, representada pela linha vermelha no gráfico, pode-se identificar a diferença do comportamento das curvas para as variáveis não lineares e lineares. Apesar de não estarem nos gráficos, por uma questão de visualização, as bandas de 95% de confiança para cada estimativa de curva, foi construída. Logo, pode-se avaliar a proporção de pontos das bandas que contêm o valor real zero para cada curva estimada e obter a média dessas proporções para cada variável:

Tabela 3 – Proporção média de pontos das bandas de confiança (95%) contendo o valor zero para cada curva

	n=500	n=1000	n=2000
Curva	Prop. média	Prop. média	Prop. média
(cosseno)	0.273	0.196	0.168
(doppler)	0.304	0.287	0.270
(reta)	0.932	0.939	0.943
(reta)	0.928	0.933	0.934

Analisando os valores das proporções médias de pontos das bandas de confiança que contêm o valor zero, dispostos na Tabela 3, pode-se notar que as duas últimas curvas, cuja relação verdadeira é dada de forma linear, apresentam valores muito próximos de 95%, indicando que suas derivadas de segunda ordem podem ser iguais a zero, tornando as covariáveis correspondentes como pertencentes à parte paramétrica do modelo. Por outro lado, as duas primeiras apresentam valores baixos, indicando que suas derivadas de segunda ordem não são iguais a zero, tornando as covariáveis correspondentes como pertencentes à parte não paramétrica do modelo. Sendo assim, pelo método de análise da segunda derivada, o modelo escolhido é o modelo parcialmente linear (modelo verdadeiro simulado).

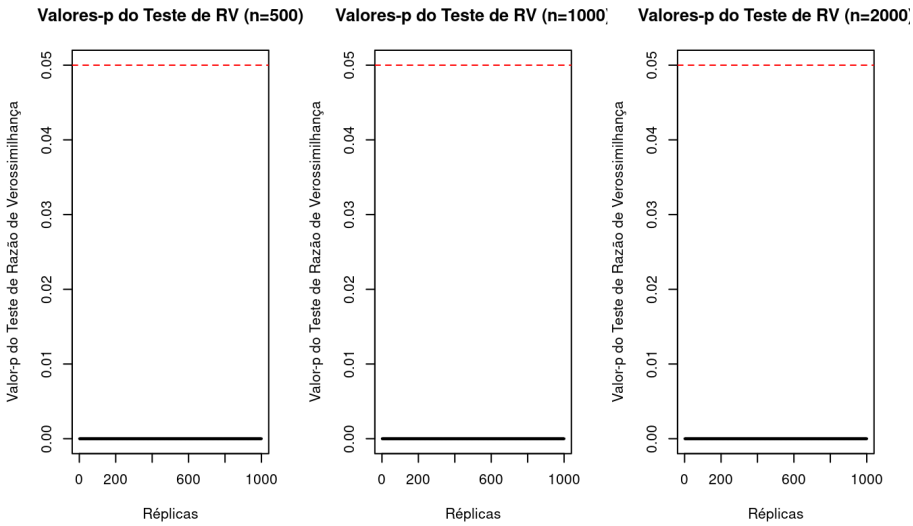
Para confirmar os resultados obtidos, foi aplicado o teste de razão de verossimilhanças entre os três modelos ajustados. O objetivo é comparar o primeiro modelo (APLM) com os demais, linear e aditivo, por meio do teste de razão de verossimilhanças. Na primeira aplicação do teste, os modelos APLM e de regressão linear múltipla foram comparados. Como o modelo linear é menos complexo que o APLM, pode-se formular o seguinte teste de hipóteses

H_0 : O modelo de regressão linear oferece um ajuste tão bom para os dados quanto o modelo parcialmente linear.

H_1 : O modelo de regressão linear não oferece um ajuste tão bom para os dados quanto o modelo parcialmente linear.

O teste foi repetido para cada uma das 1000 simulações considerando os três diferentes tamanhos de amostra. Os p-valores obtidos no teste foram plotados nos gráficos abaixo.

Figura 12 – P-valores do teste de razão de verossimilhanças entre os modelos linear e APML.



Fonte: Elaboração do autor.

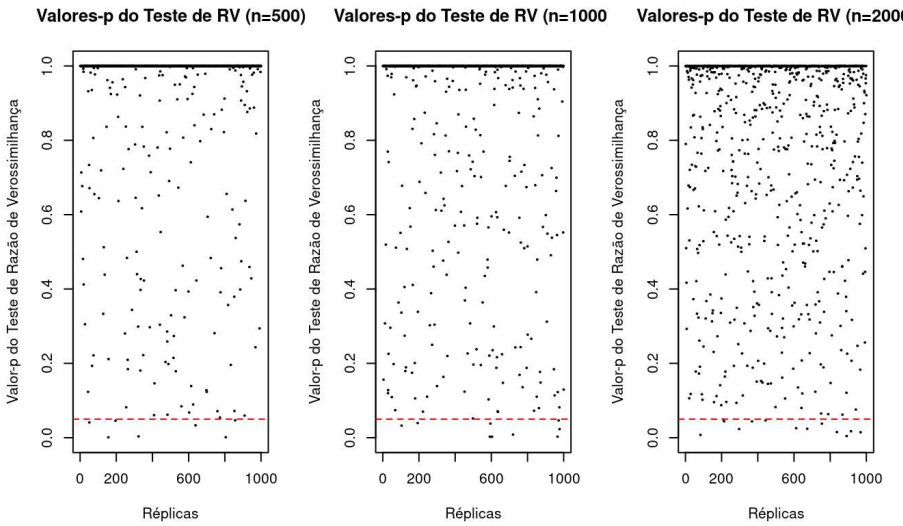
Analisando a Figura 12, percebe-se que a hipótese nula foi rejeitada à um nível de 5% de significância para os 1000 testes realizados para cada um dos três tamanhos de amostra. Logo, conclui-se que o modelo de regressão linear múltipla não oferece um bom ajuste aos dados, evidenciando que uma ou mais variáveis preditoras não possuem relação linear com a variável resposta.

Em seguida, os mesmos testes foram realizados novamente, mas agora comparando os modelos aditivo e parcialmente linear. Neste cenário, o modelo mais complexo é o modelo aditivo, onde todas as variáveis explicativas assumem uma relação não linear com a variável resposta. Logo, formula-se seguinte teste de hipóteses:

H_0 : O modelo parcialmente linear oferece um ajuste tão bom para os dados quanto o modelo aditivo.

H_1 : O modelo parcialmente linear não oferece um ajuste tão bom para os dados quanto o modelo aditivo.

Figura 13 – P-valores do teste de razão de verossimilhanças entre os modelos APML e aditivo.

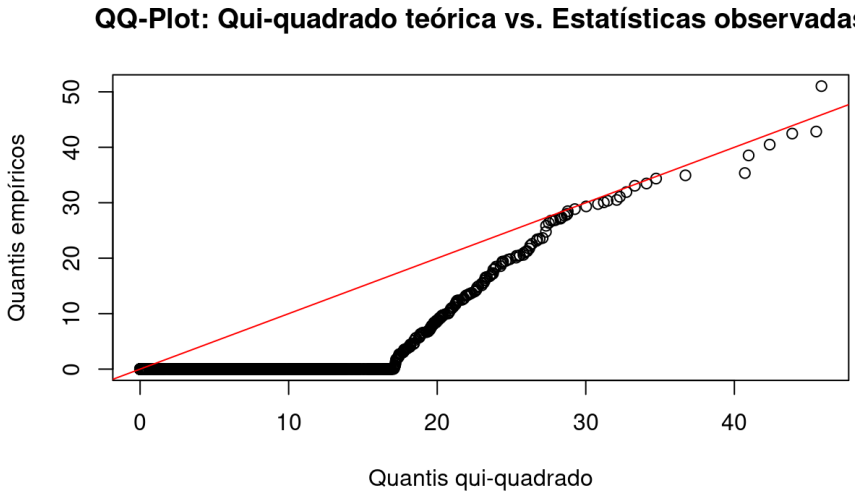


Fonte: Elaboração do autor.

Observando os resultados do teste, tem-se que para amostras de tamanho 500, a hipótese nula foi rejeitada em 0,7% dos 1000 testes realizados. Para os tamanhos de amostra 1000 e 2000, estes percentuais foram de 0,9% e 1,3%, respectivamente. Analisando os gráficos, de um modo geral, e considerando os resultados mencionados no parágrafo anterior, pode-se concluir que a hipótese nula não foi rejeitada à um nível de 5% de significância para os 1000 testes feitos em cada tamanho amostral. Logo, conclui-se que o modelo parcialmente linear oferece um bom ajuste aos dados, evidenciando que as variáveis preditoras X_1 e X_2 se relacionam de maneira linear e que as variáveis t_1 e t_2 se relacionam de forma não linear. Temos então que, pelo teste de razão de verossimilhanças, foi possível reforçar a conclusão já obtida pelas segundas derivadas: o APLM é o modelo ideal para modelar os dados simulados, como já esperado pois este é modelo verdadeiramente simulado.

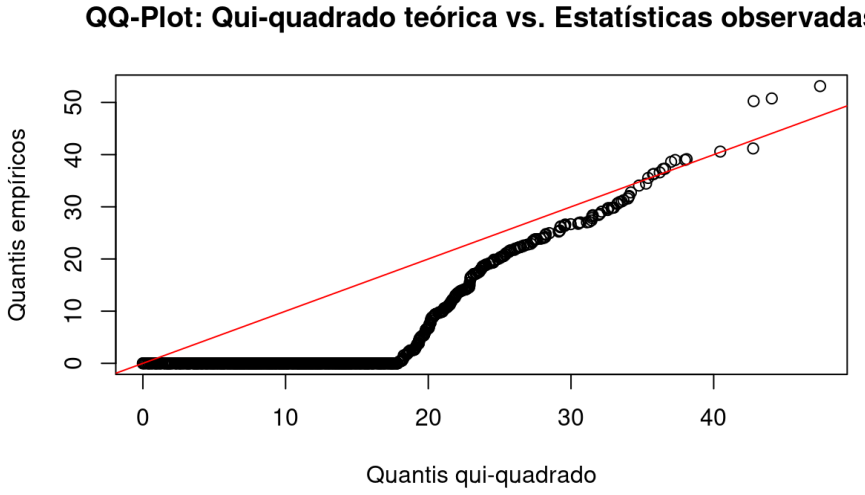
Após a aplicação destes testes, apesar dos resultados satisfatórios, houve a necessidade de analisar o comportamento das estatísticas dos testes de razão de verossimilhança. Como abordado na Seção 3.2, em casos em que os modelos são penalizados, as estatísticas calculadas para o teste de razão de verossimilhanças podem não convergir para uma distribuição qui-quadrado. Nas Figuras 14, 15 e 16, estão os gráficos que evidenciam este resultado de não convergência das estatísticas obtidas neste trabalho.

Figura 14 – Qui-quadrado teórica vs. Estatísticas observadas (n=500).



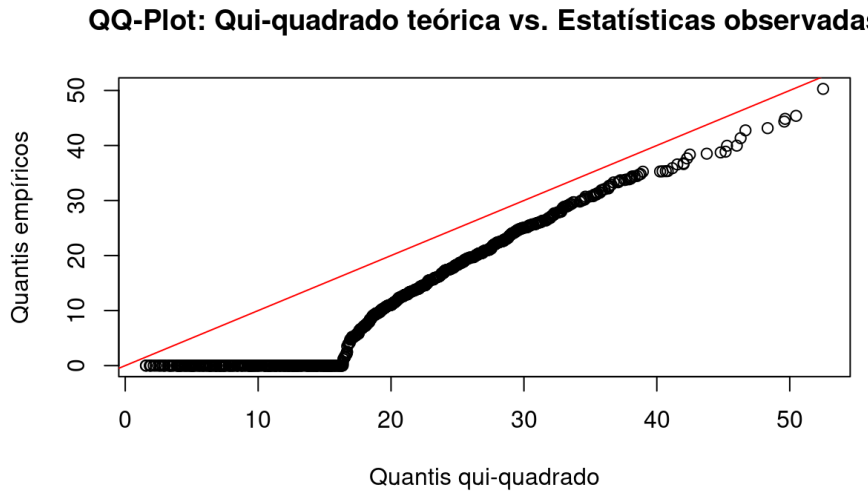
Fonte: Elaboração do autor.

Figura 15 – Qui-quadrado teórica vs. Estatísticas observadas (n=1000).



Fonte: Elaboração do autor.

Figura 16 – Qui-quadrado teórica vs. Estatísticas observadas (n=2000).



Fonte: Elaboração do autor.

Analisando os gráficos, percebe-se que os pontos não coincidem com a linha em vermelho, indicando uma não aderência da estatística de teste à distribuição qui-quadrado. Os quantis empíricos de valor zero resultam das observações que apresentaram valor negativo para a estatística de teste. Nestes casos, optou-se por substituir esses valores negativos por zero, com o objetivo de preservar a coerência teórica da estatística de teste e evitar distorções em sua distribuição, uma vez que a qui-quadrado assume valores não negativos.

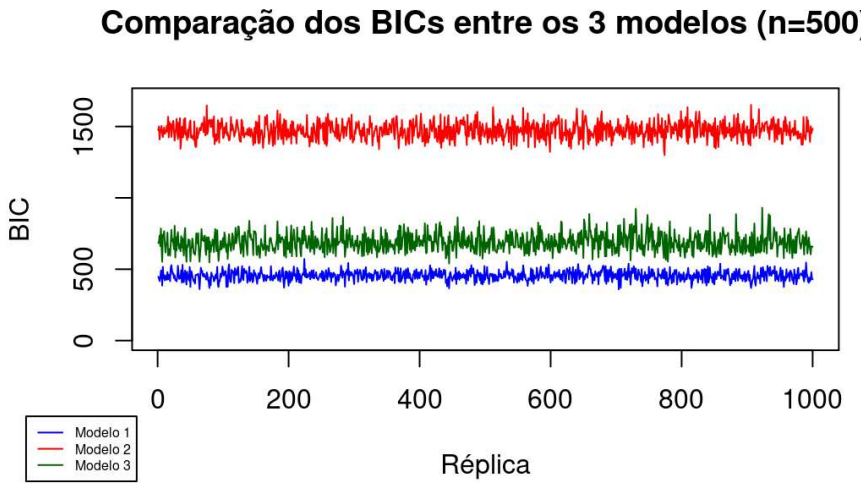
Diante destes resultados, a proposta alternativa sugerida é que os testes sejam feitos via Bootstrap. Esse procedimento permite obter uma distribuição empírica da estatística de teste sob a hipótese nula, sem depender da suposição de que a estatística segue exatamente uma distribuição qui-quadrado, o que pode não ser válido em contextos com penalizações. A ideia, de um modo geral, é utilizar a construção de uma distribuição empírica de Λ via Bootstrap sob a hipótese nula. Feito isso, calcula-se o quantil 1-α das observações desta distribuição empírica. Em seguida, compara-se as estatísticas observadas (aquelas usadas nos testes aplicados) com o quantil obtido, verificando a proporção de estatísticas calculadas que estão na área de rejeição da hipótese.

No entanto, para que o método Bootstrap seja aplicado corretamente, é necessário gerar diversas réplicas dos dados sob o modelo nulo e, para cada uma delas, reestimar os parâmetros de ambos os modelos (reduzido e alternativo), a fim de recalcular a estatística de teste. Embora essa abordagem seja teoricamente adequada e apresente bons resultados na literatura, sua aplicação prática pode demandar alto custo computacional, especialmente em modelos com alta complexidade ou grande volume de dados.

Por esse motivo, optou-se por não realizar a abordagem via Bootstrap neste trabalho, embora os resultados obtidos indiquem que seria um método potencialmente mais confiável em situações onde a distribuição assintótica da estatística de teste não apresenta boa convergência. A aplicação do Bootstrap permanece como uma alternativa viável para estudos futuros, caso se disponha de maior capacidade computacional ou de algoritmos mais eficientes para a reestimação dos parâmetros.

Diante dos desafios apresentados, optou-se em utilizar o BIC dos modelos como critério de comparação. Nas Figuras 17 a 19 são apresentados os gráficos com os BICs obtidos em cada réplica para cada um dos modelos ajustados (modelos 1, 2 e 3), considerando os diferentes tamanhos amostrais.

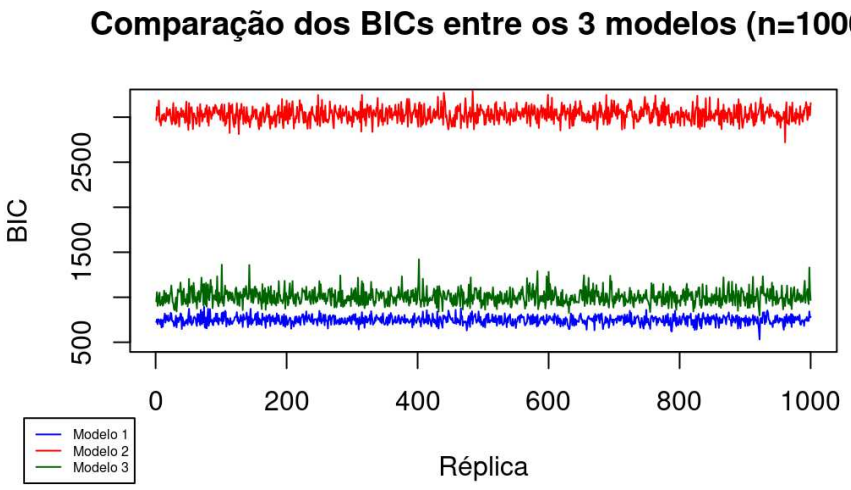
Figura 17 – BICs dos modelos 1, 2 e 3 para n=500.



.

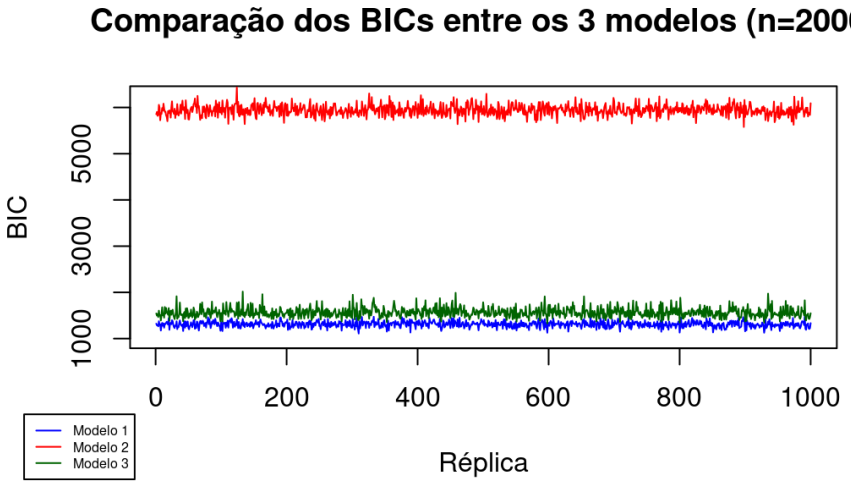
Fonte: Elaboração do autor.

Figura 18 – BICs dos modelos 1, 2 e 3 para n=1000.



Fonte: Elaboração do autor.

Figura 19 – BICs dos modelos 1, 2 e 3 para n=2000.



Fonte: Elaboração do autor.

Analisando as Figuras 17, 18 e 19, percebe-se que para os três diferentes tamanhos de amostra, o modelo 1 apresenta os menores valores para o BIC em todas as réplicas. Sendo assim, ao comparamos os modelos utilizando este critério, novamente o modelo 1 (modelo verdadeiro) é selecionado como aquele que oferece um melhor ajuste aos dados.

Com o objetivo de validar os resultados obtidos por meio das simulações, os métodos propostos também foram aplicados a conjuntos de dados reais. Dessa forma, o Capítulo 5

apresenta três aplicações distintas, realizadas em diferentes bases de dados, com o intuito de ilustrar o desempenho prático dos modelos estudados.

5 APLICAÇÃO

Uma maneira de validar a metodologia apresentada no presente trabalho é demonstrar a aplicação à dados reais e analisar os resultados obitidos. Sendo assim, neste capítulo são apresentados três aplicações utilizando variados conjuntos de dados reais.

5.1 MILAN.MORT

Nesta seção é apresentada a primeira aplicação utilizando os dados "Milan.mort" do pacote "{Semipar}" disponível no R Core Team (2024). A ideia é modelar a variabilidade das mortes em Milão, na Itália, registradas entre o período de 1° de janeiro de 1980 a 30 de dezembro de 1989. O conjunto de dados foi originalmente publicado por Vigotti et al. (1996) e é composto por 3652 observações e 9 variáveis, sendo:

- **day.num**: número de dias desde 31 de dezembro de 1979
- **day.of.week**: 1=Segunda-feira, 2=Terça-feira, 3=Quarta-feira, 4=Quinta-feira, 5=Sexta-feira, 6=Sábado, 7=Domingo.
- **holiday**: 1=feriado, 0=caso contrário.
- **mean.temp**: temperatura média diária em graus Celsius.
- **rel.humid**: umidade relativa.
- **tot.mort**: número total de mortes.
- **resp.mort**: número total de mortes respiratórias.
- **SO2**: medida do nível de dióxido de enxofre no ar ambiente.
- **TSP**: total de partículas suspensas no ar ambiente.

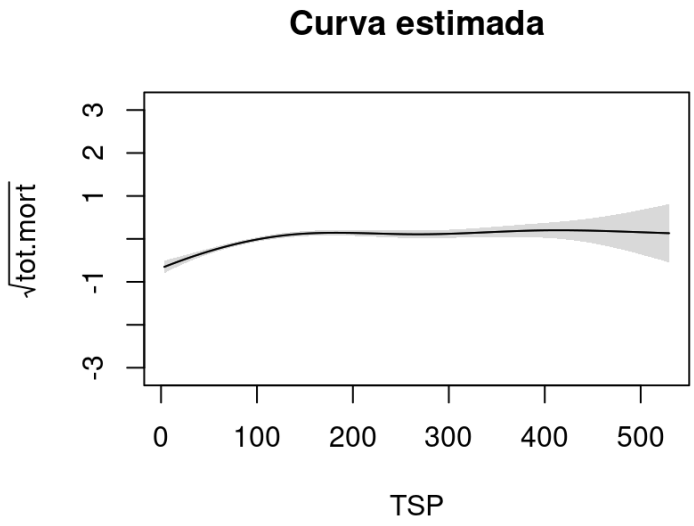
Ruppert et al. (2003) apresentam uma aplicação utilizando os dados de mortalidade em Milão, com o objetivo de demonstrar o comportamento das curvas em função do número de nós escolhidos. No modelo apresentado por eles, a raíz quadrada da mortalidade é modelada em função da temperatura média diária ($mean.temp$), o total de partículas suspensas no ar ambiente (TSP), (usada como a única variável linear do modelo), a humidade relativa ($rel.humid$) e o número de dias ($day.num$). Utilizando as mesmas variáveis consideradas pelos autores, foi construído um modelo aditivo aplicando a metodologia vista neste trabalho. Entretanto, o objetivo aqui é obter as estimativas dos parâmetros do modelo, bem como as curvas que representam a parte não linear, analisar a segunda derivada das curvas estimadas, com a finalidade de selecionar as variáveis lineares e não lineares do modelo, e por fim analisar a qualidade de ajuste do modelo final.

Inicialmente, seguindo a metologia apresentada no trabalho, o modelo proposto considera que todas as variáveis preditoras possuem relações não lineares com a variável resposta:

$$\sqrt{tot.mort_i} = g_1(TSP_i) + g_2(day.num_i) + g_3(mean.temp_i) + g_4(rel.humid_i) + \epsilon_i \quad (5.1)$$

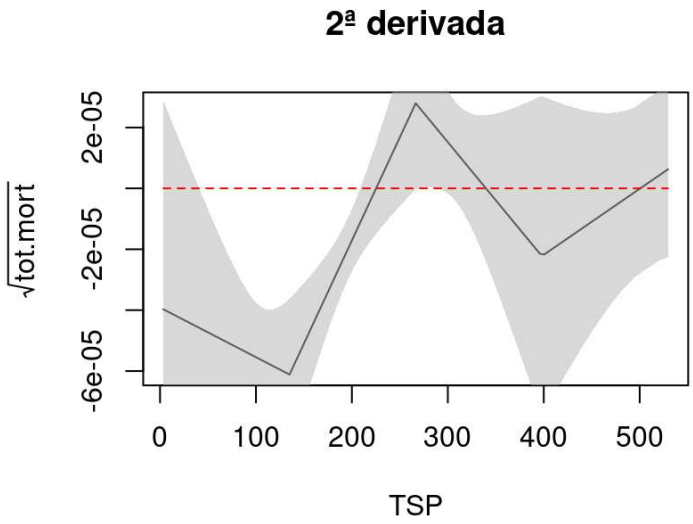
Como resultado da aplicação do algoritmo desenvolvido para a construção do modelo proposto, obtiveram-se as curvas estimadas e suas respectivas derivadas de segunda ordem para cada variável, apresentadas nas Figuras 14 a 21.

Figura 20 – Estimativas da curva para TSP



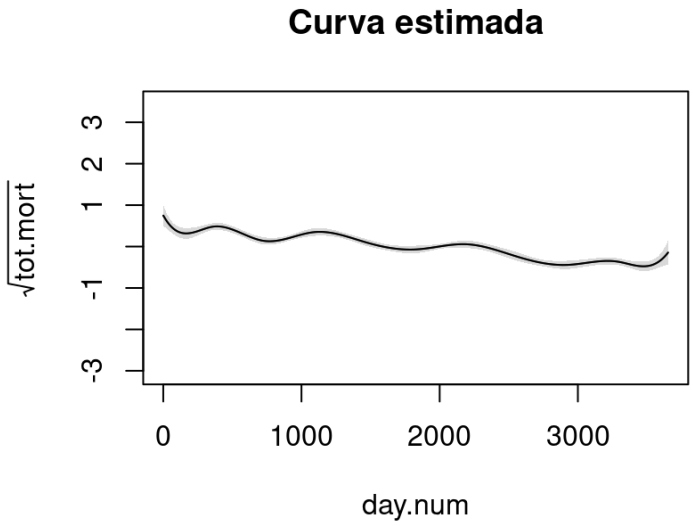
Fonte: Elaboração do autor.

Figura 21 – Estimativa da 2ª derivada para TSP



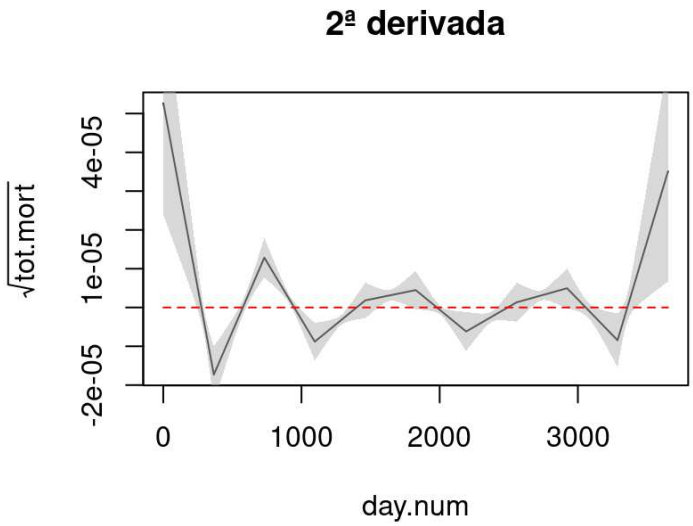
Fonte: Elaboração do autor.

Figura 22 – Estimativa da curva para day.num



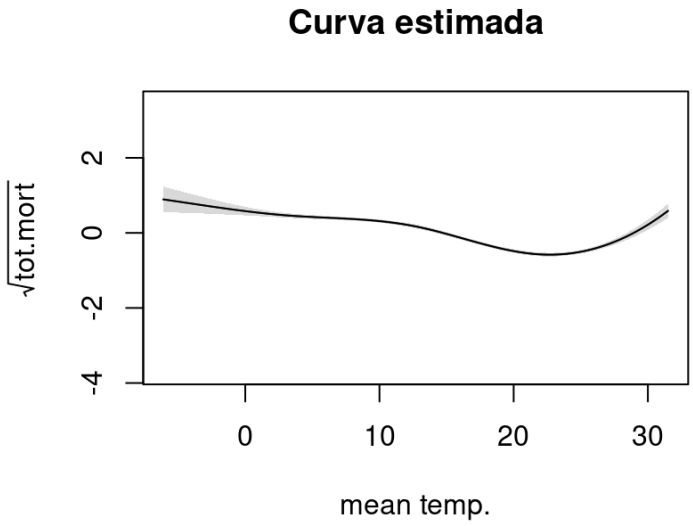
Fonte: Elaboração do autor.

Figura 23 – Estimativa da 2ª derivada para day.num



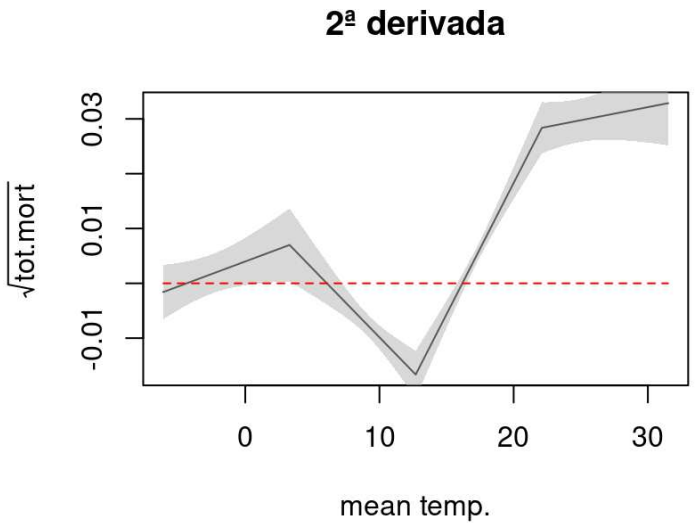
Fonte: Elaboração do autor.

Figura 24 – Estimativa da curva para mean.temp



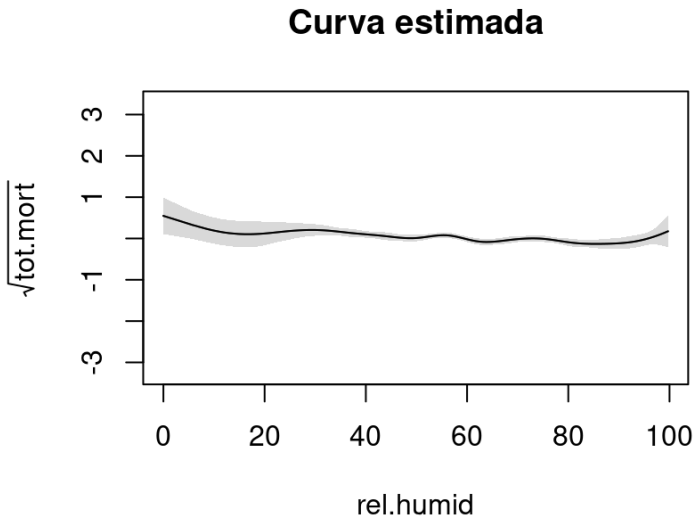
Fonte: Elaboração do autor.

Figura 25 – Estimativa da 2ª derivada para mean.temp



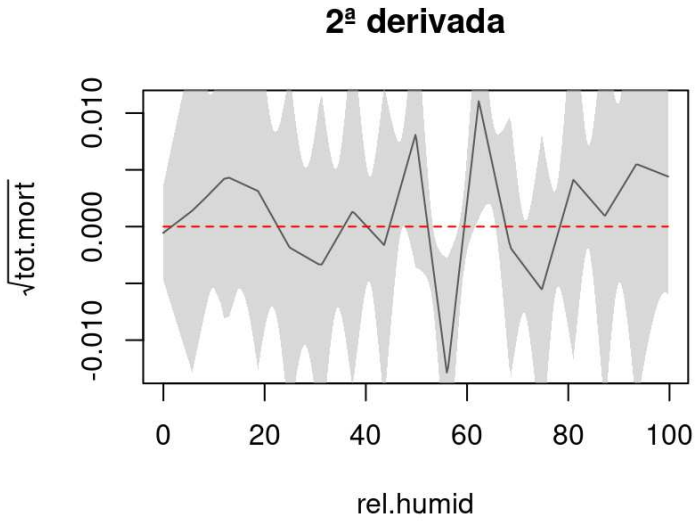
Fonte: Elaboração do autor.

Figura 26 – Estimativa da curva para rel.humid



Fonte: Elaboração do autor.

Figura 27 – Estimativa da 2ª derivada para rel.humid



Fonte: Elaboração do autor.

Avaliando os gráficos das derivadas de segunda ordem de cada variável preditora, constatou-se que os percentuais de pontos das bandas de confiança que contêm o valor zero são de: 16.6%, 31.8%, 20.6% e 79.5%, para as variáveis TSP, $day.num$, $mean.temp$ e $rel.humid$, respectivamente. Sendo assim, não há nenhuma curva da derivada de segunda ordem que possa ser considerada uma reta, segundo a metodologia aplicada, pois nenhum percentual está acima de 85%, evidenciando que as variáveis devem ser consideradas como tendo relações não lineares de acordo com a metodologia aplicada. Logo, o modelo considerado é o modelo aditivo ajustado inicialmente:

$$\sqrt{tot.mort_i} = g_1(TSP_i) + g_2(day.num_i) + g_3(mean.temp_i) + g_4(rel.humid_i) + \epsilon_i$$

Os valores dos parâmetros de suavização λ de cada curva e da variância dos erros σ^2 para o modelo (5.1) estão apresentados na Tabela 4:

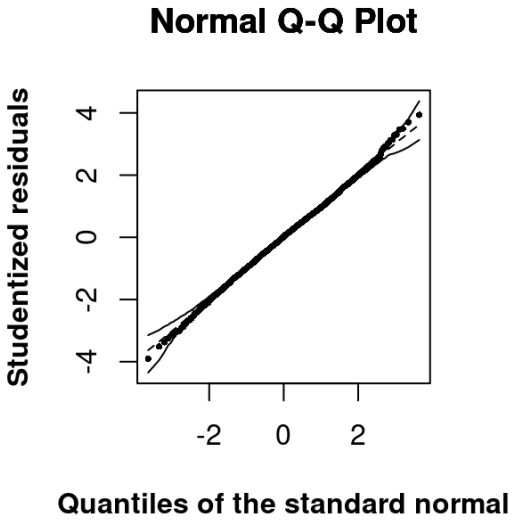
Tabela 4 – Milan.mort: Estimativas de máxima verossimilhança penalizada dos parâmetros e seus erros-padrão

Parâmetro	Estimativa	Erro-padrão
λ_1	0.5903	-
λ_2	0.0198	-
λ_3	1.0000	-
λ_4	1.0000	-
σ^2	0.6699	0.0002

Analisando os parâmetros de suavização, pode-se perceber que a curva estimada para variável day.num apresenta o menor valor para λ e portanto, seria a curva com mais oscilações. Isto pode ser constatado na Figura 22. As curvas estimadas para as variáveis mean.temp e rel.humid são semelhantes no grau de oscilação, pois os parâmetros λ_3 e λ_4 possuem o mesmo valor. Logo, estas são, entre as quatro, as curvas mais suaves. A estimativa associada ao parâmetro de variância dos resíduos do modelo também é apresentado na Tabela 4. O baixo valor do erro-padrão indica a confiança de que o valor estimado é preciso.

Para validar o modelo, foi feito o gráfico envelope dos resíduos:

Figura 28 – Milan.Mort: Envelope dos resíduos



Fonte: Elaboração do autor.

A Figura 28 apresenta um gráfico de envelope dos resíduos que evidencia um bom ajuste do modelo aos dados. O percentual de resíduos fora das bandas de confiança é de 0.027% do tamanho amostral. Logo, por ser um percentual abaixo de 5%, pode-se concluir que o modelo aditivo, oferece um bom ajuste aos dados Milan.mort.

5.2 RAGWEED

Com o objetivo de analisar a aplicação da metodologia em outros casos de modelos, foi utilizado o conjunto de dados *Ragweed* do pacote {*Semipar*} disponível no R Core Team (2024). Este conjunto contém dados sobre os níveis de ambrósia e variáveis meteorológicas por 335 dias em Kalamazoo, Michigan, EUA.

A variável de interesse é o nível de pólen de ambrósia (*ragweed*). Os dados foram registrados durante a temporada de ambrósia de 1993 em Kalamazoo, Michigan e foram

publicados por Stark et al. (1997). Como a prevenção desempenha um papel importante no tratamento de alergias relacionadas ao pólen, um dos principais objetivos da aerobiologia é o desenvolvimento de modelos de previsão precisos para os níveis diários de pólen. O conjunto de dados em questão consiste de:

- **ragweed**: nível de pólen de ambrósia para aquele dia (grãos/m^3);
- **temperature**: temperatura do dia seguinte (°F);
- **rain**: indicador de chuva significativa para o dia seguinte (1 = pelo menos 3 horas de chuva contínua ou chuva breve mas intensa, 0 = caso contrário);
- **wind**: previsão de velocidade do vento para o dia seguinte (nós);
- **day.in.seas**: número do dia na atual temporada de pólen de ambrósia.

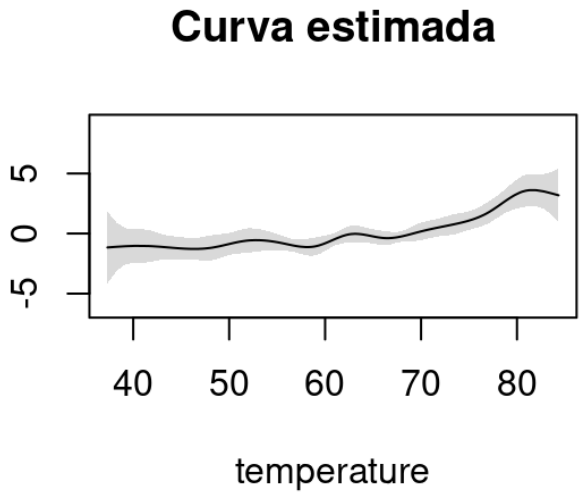
Ruppert et al. (2003) também abordam este conjunto de dados em seu livro, aplicando um modelo semi paramétrico para a modelagem. No modelo apresentado eles utilizam a raíz quadrada da variável *ragweed* como a variável resposta do modelo, as variáveis *temperature*, *rain* e *wind* pertencentes a parte paramétrica do modelo e a variável *day.in.seas* como não paramétrica.

Assim como na aplicação apresentada na Seção 5.1, o passo inicial é ajustar um modelo onde todas as covariáveis são consideradas como tendo relações não lineares, ou seja, um modelo aditivo. Como a variável preditora *rain* é uma variável *dummy*, esta não é avaliada na etapa de seleção de variáveis. Logo, neste caso, a variável será considerada de imediato como pertencente à parte paramétrica e o modelo ajustado passa a ser um APLM, dado da seguinte forma:

$$\sqrt{ragweed_i} = \beta_0 + \beta_1 rain_i + g_1(temperature_i) + g_2(wind_i) + g_3(day.in.seas_i) + \epsilon_i \quad (5.2)$$

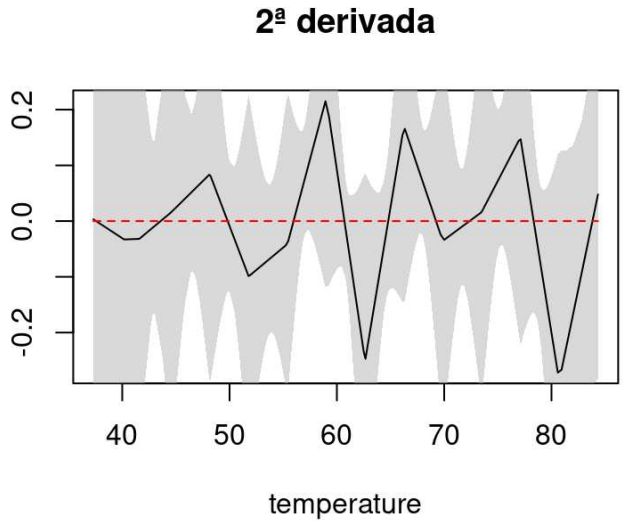
Ajustando o modelo acima, obteve-se as curvas estimadas e suas respectivas derivadas de segunda ordem para cada covariável. As curvas são apresentadas nas Figuras 29 a 34.

Figura 29 – Estimativa da curva para temperature



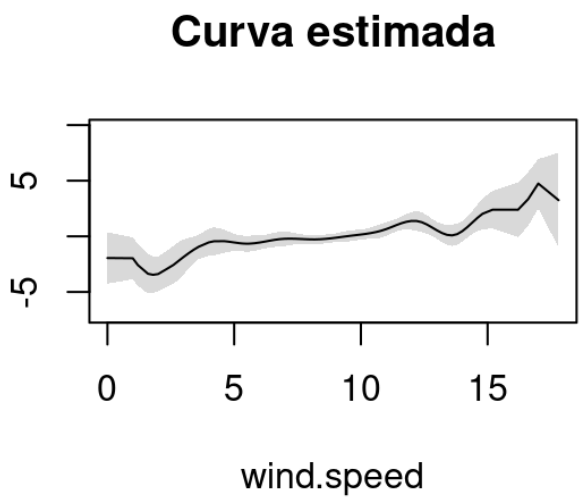
Fonte: Elaboração do autor.

Figura 30 – Estimativa da 2ª derivada para temperature



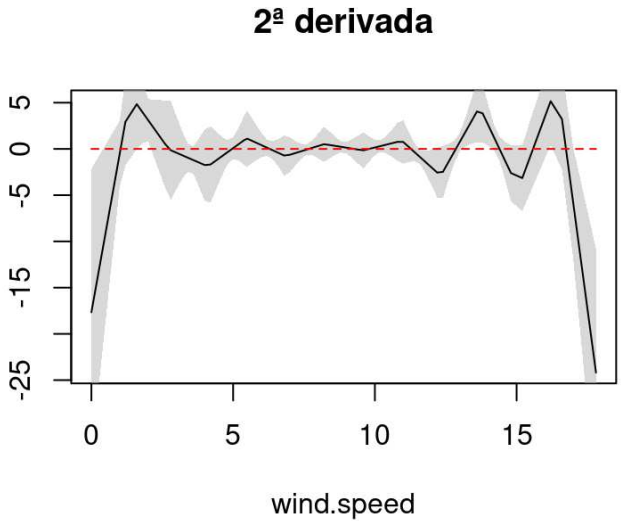
Fonte: Elaboração do autor.

Figura 31 – Estimativa da curva para wind



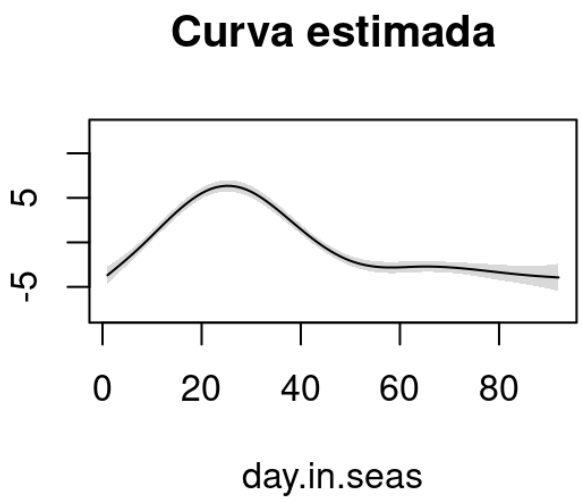
Fonte: Elaboração do autor.

Figura 32 – Estimativa da 2ª derivada para wind



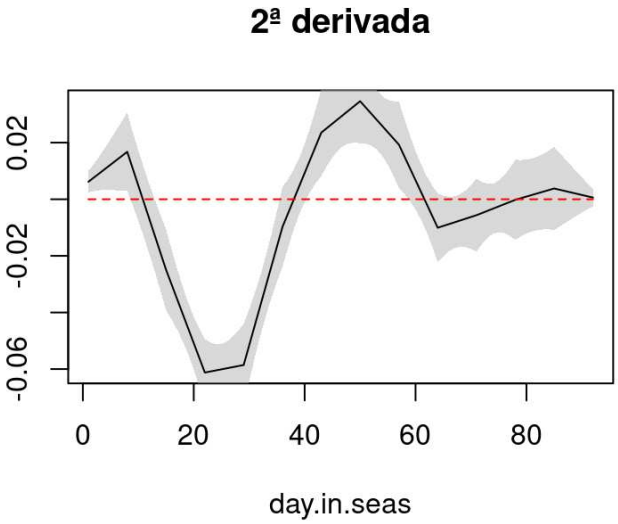
Fonte: Elaboração do autor.

Figura 33 – Estimativa da curva para day.in.seas



Fonte: Elaboração do autor.

Figura 34 – Estimativa da 2ª derivada para day.in.seas



Fonte: Elaboração do autor.

Verificando as bandas de 95% de confiança da segunda derivada, pode-se obter o percentual de pontos das bandas que contêm o valor zero. Os percentuais foram de 100%, 88.06% e 42.98%, para as variáveis *temperature*, *wind* e *day.in.seas*, respectivamente. Com estes resultados, é possível selecionar a variável *day.in.seas* como tendo uma relação não linear, por ter uma segunda derivada muito diferente de uma reta em zero. A variável

temperature é considerada como tendo uma relção linear com a variável resposta, pois apresentou uma segunda derivada muito semelhante à uma reta em zero. A variável *wind.speed* apresentou uma alta porcentagem de pontos na banda de confiança que englobam o valor zero. Apesar de não ter atingido os 95%, de acordo com a metodologia adotada, é interessante considerá-la como sendo pertencente à parte paramétrica, ajustar novos modelos e compará-los via teste de razão de verossimilhanças e BIC.

Sejam os seguintes modelos:

$$\sqrt{ragweed_i} = \beta_0 + \beta_1 rain_i + \beta_2 temp_i + g_1(wind_i) + g_2(day.in.seas_i) + \epsilon_i \tag{5.3}$$

$$\sqrt{ragweed_i} = \beta_0 + \beta_1 rain_i + \beta_2 wind_i + g_1(temp_i) + g_2(day.in.seas_i) + \epsilon_i \tag{5.4}$$

$$\sqrt{ragweed_i} = \beta_0 + \beta_1 rain_i + \beta_2 wind_i + \beta_3 temp_i + g_1(day.in.seas_i) + \epsilon_i \tag{5.5}$$

Após ajustar cada um dos modelos, o teste de razão de verosimilhanças foi aplicado a fim de comparar o modelo (5.5), com os demais modelos. Assim, temos a realização de três testes de hipóteses.

- Teste 1:

$$H_0 : \text{O modelo (5.5) oferece um ajuste tão bom quanto o modelo (5.2)}$$
$$H_1 : \text{O modelo (5.5) não oferece um ajuste tão bom quanto o modelo (5.2)}$$

- Teste 2:

$$H_0 : \text{O modelo (5.5) oferece um ajuste tão bom quanto o modelo (5.3)}$$
$$H_1 : \text{O modelo (5.5) não oferece um ajuste tão bom quanto o modelo (5.3)}$$

- Teste 3:

$$H_0 : \text{O modelo (5.5) oferece um ajuste tão bom quanto o modelo (5.4)}$$
$$H_1 : \text{O modelo (5.5) não oferece um ajuste tão bom quanto o modelo (5.4)}$$

Tabela 5 – Ragweed: p-valores dos testes de razão de verossimilhanças

Teste	Estatística de teste	p-valor
1	13.20635	0.6722348
2	10.46405	0.6012254
3	11.96491	0.2201461

Analisando os p-valores da Tabela 5, pode-se concluir que nos três testes o p-valor é maior que 0.05, portanto, a hipótese nula não é rejeitada a um nível de 5% de significância nos três testes. Conclui-se que, pelo teste de razão de verossimilhanças, o modelo (5.5)

oferece um bom ajuste aos dados mesmo sendo o modelo mais simples. Para reforçar esta conclusão, pode-se analisar o critério BIC, em que dá-se preferência ao modelo com menor BIC.

Tabela 6 – Ragweed: BIC dos modelos ajustados

Modelos	BIC
(5.2)	1742.873
(5.3)	1651.407
(5.4)	1642.283
(5.5)	1583.387

Os valores apesentados na Tabela 6 indicam um menor valor do BIC para o modelo (5.5). Logo, este critério também permite a conclusão de que o modelo ideal para os dados *ragweed* é aquele cujo a única variável com relação não linear é a *day.in.seas*, sendo portanto, um modelo semi-paramétrico:

$$\sqrt{ragweed_i} = \beta_0 + \beta_1 rain_i + \beta_2 wind_i + \beta_3 temp_i + g_1(day.in.seas_i) + \epsilon_i$$

Este mesmo modelo é proposto por Ruppert et al. (2003). Como resultado da aplicação do algoritmo desenvolvido para a construção do modelo proposto, tem-se:

Tabela 7 – Ragweed: Estimativas de máxima verossimilhança penalizada dos parâmetros, seus erros-padrão e p-valores

Parâmetro	Estimativa	Erro-padrão	p-valor
λ	0.381	-	-
σ^2	4.535	0.123	-
β_0	-5.079	0.827	4.183e-10
β_1	1.421	0.151	0
β_2	0.237	0.0014	0
β_3	0.101	0.0001	0

Os valores da Tabela 7 permitem a interpretação da parte paramétrica do modelo. Considerando todas as demais variáveis inalteradas, em dias onde há mais de 3 horas de chuva contínua, a variável rain assume valor 1, contribuindo para o aumento de aproximadamente 1.42 na raíz quadrada do volume de nível de grãos. A cada uma unidade acrescida na previsão de velocidade do vento para o dia seguinte, aumenta-se cerca de 0.237 na raíz quadrada do volume de nível de grãos, considerando todas as demais variáveis inalteradas. Para a variável temperature, tem-se que, mantendo constante as demais variáveis, a cada 1°F acrescido na temperatura do dia seguinte, aumenta-se aproximadamente 0.1 na raíz quadrada do volume de nível de grãos.

Os parâmetros de $\boldsymbol{\beta}$ do modelo foram submetidos ao teste de significância. Para avaliar a significância dos coeficientes $\boldsymbol{\beta}$ estimados no modelo, foi utilizado o teste t,

que compara cada estimativa com seu erro padrão correspondente. O seguinte teste de hipóteses foi feito para cada parâmetro β_j:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

A estatística t_{calc} para cada coeficiente β_j foi calculada como:

$$t_{calc} = \frac{|\hat{\beta}_j|}{ep(\hat{\beta}_j)},$$

onde $\hat{\beta}_j$ é valor estimado para o j-ésimo coeficiente, e $ep(\hat{\beta}_j)$ o valor do erro-padrão para o j-ésimo coeficiente.

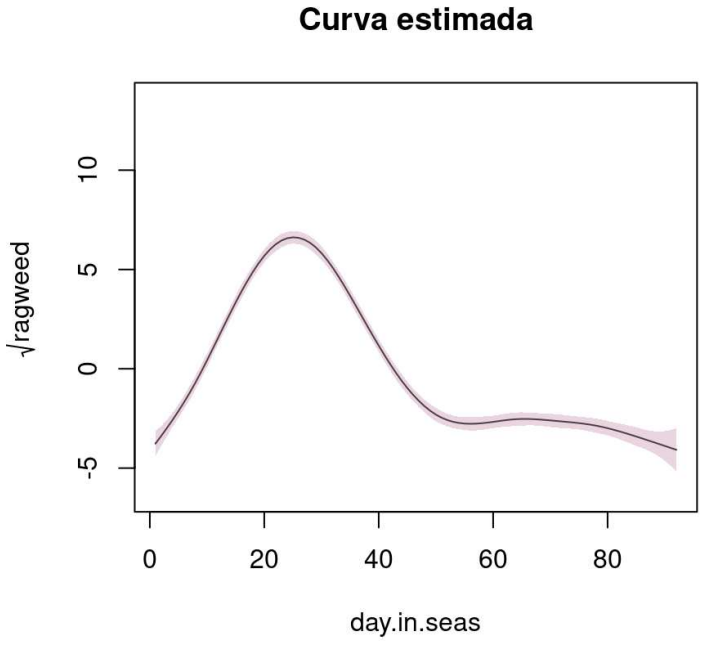
A partir da estatística t_{calc}, o p-valor correspondente foi obtido utilizando a função de distribuição acumulada da distribuição normal padrão:

$$\text{p-valor} = 1 - \Phi(t_{calc}),$$

onde $\Phi(t_{calc})$ é a função de distribuição acumulada da normal padrão avaliada em t_{calc}.

Os p-valores apresentados na Tabela 7 indicam que a um nível de 5% de significância, a hipótese nula foi rejeitada em todos os testes realizados. Logo, os coeficientes $\beta_0, \beta_1, \beta_2$ e β_3 são significativamente diferentes de zero. A Figura 35 apresenta a estimativa não paramétrica, com as bandas de 95% de confiança, para a variável day.in.seas.

Figura 35 – Estimativa não paramétrica para day.in.seas

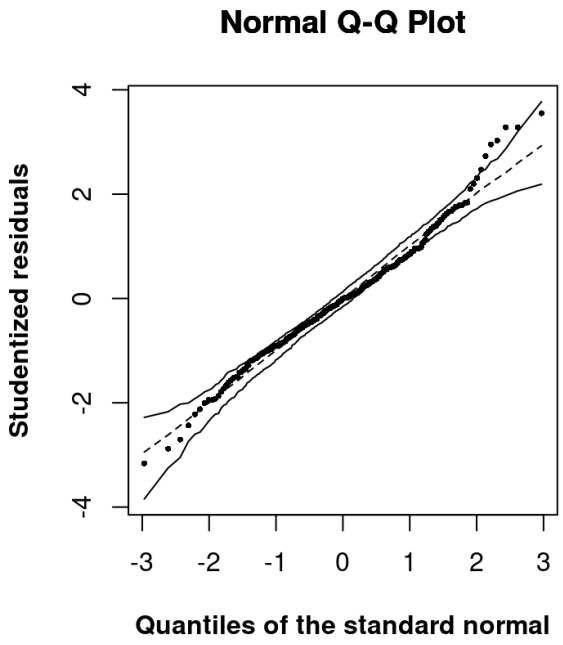


Fonte: Elaboração do autor.

Pela Figura 35, percebe-se que a raíz quadrada do nível de grãos sofre um aumento desde o primeiro dia da temporada até por volta do vigésimo dia. A partir disso, até por volta do quinquagésimo dia, há uma queda na raíz quadrada do nível de grãos. Do quinqugésimo dia até os dias finais, a raíz quadrada fica de certa forma estável, não apresentando grandes aumentos ou quedas.

Para averiguar a qualidade de ajuste, foi feito o gráfico de envelope dos resíduos do modelo ajustado.

Figura 36 – Ragweed: Envelope dos resíduos



Fonte: Elaboração do autor.

Ao analisar o envelope, percebe-se que há alguns pontos fora das bandas de confiança. Entretanto, estes correspondem à aproximadamente 1.5% do total de observações. Como este percentual é menor que 5%, pode-se concluir que os resíduos indicam um bom ajuste do modelo aos dados.

5.3 BOSTON

Uma terceira aplicação é apresentada nesta seção. Os dados aqui modelados estão relacionados a características de moradias em Boston. A ideia é estudar a relação entre o valor médio das casas em diferentes áreas da cidade e várias características socioeconômicas e ambientais.

Os dados utilizados foram originalmente publicados por Harrison e Rubinfeld (1978), com base no censo dos Estados Unidos de 1970. Esse conjunto de dados está disponível no $\{MASS\}$ do R Core Team (2024). Ao todo o conjunto contém 506 observações, cada uma representando uma área de Boston, com 14 variáveis que incluem características das zonas e a variável resposta, que é o valor médio das casas na área.

- **crim**: Taxa de criminalidade per capita na cidade.
- **zn**: Proporção de terrenos residenciais divididos em lotes com mais de 25.000 pés quadrados.
- **indus**: Proporção de acres de negócios não varejistas por cidade.
- **chas**: Variável indicadora de se o local faz fronteira com o rio Charles (1 se sim, 0 se não).
- **nox**: Concentração de óxidos nítricos (em partes por 10 milhões).
- **rm**: Número médio de quartos por habitação.
- **age**: Proporção de unidades ocupadas pelo proprietário construídas antes de 1940.
- **dis**: Distâncias ponderadas para cinco centros de emprego em Boston.
- **rad**: Índice de acessibilidade às rodovias radiais.
- **tax**: Taxa de imposto sobre propriedade por U$10.000.
- **ptratio**: Razão aluno-professor por cidade.
- **b**: $1000(Bk - 0.63)^2$, onde Bk é a proporção de moradores da cidade que são negros.
- **lstat**: Percentual da população de status socioeconômico mais baixo.
- **medv**: Valor mediano das casas ocupadas pelos proprietários (em milhares de dólares).

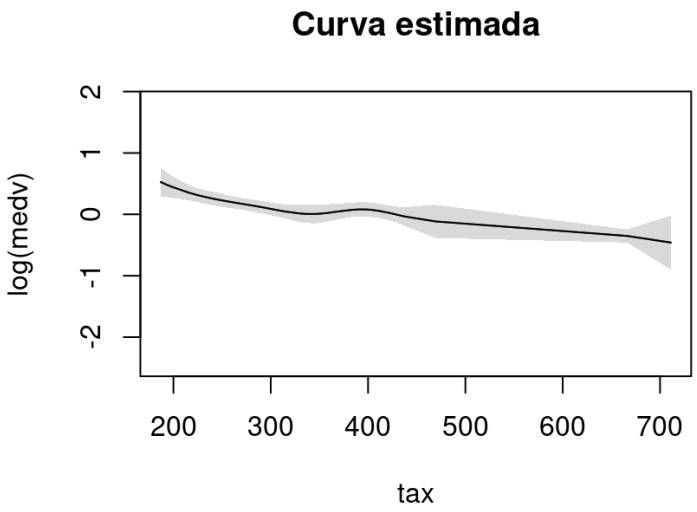
Nos modelos ajustados e analisados, foram consideradas apenas as covariáveis tax, dise $lstat$ para explicarem a variável resposta $\log(medv)$. Além disso, a transformação logarítmica é aplicada pois ajuda a aproximar a distribuição da variável resposta de uma distribuição normal, uma vez que o valor mediano dos preços das casas ($medv$) tende a ter uma distribuição assimétrica à direita.

Inicialmente, ajustou-se um modelo aditivo:

$$\log(medv_i) = g_1(tax_i) + g_2(dis_i) + g_3(lstat_i) + \epsilon_i \tag{5.6}$$

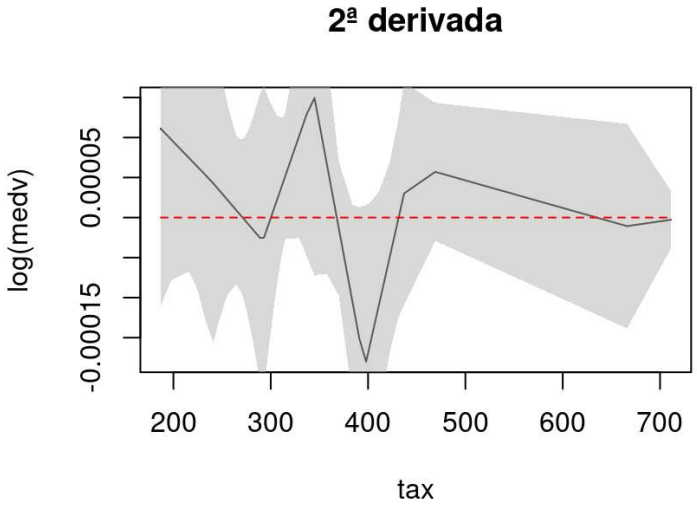
Ajustando o modelo acima, obteve-se as curvas estimadas e suas respectivas derivadas de segunda ordem para cada covariável. As curvas obtidas estão apresentadas nas Figuras 37 a 42.

Figura 37 – Estimativa da curva para tax



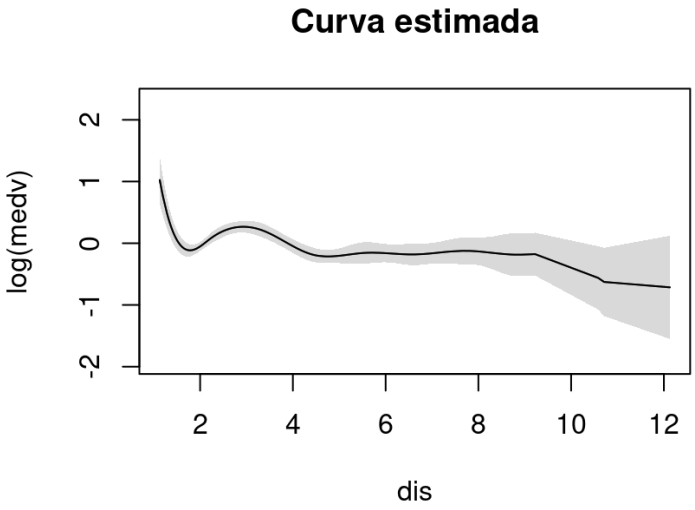
Fonte: Elaboração do autor.

Figura 38 – Estimativa da 2ª derivada para tax



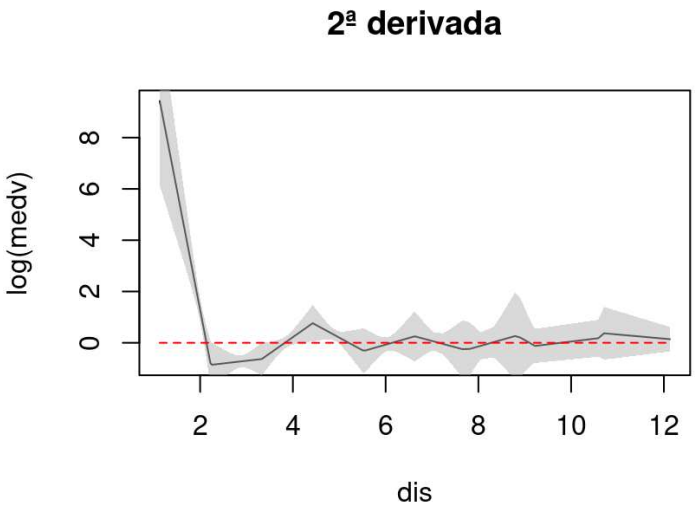
Fonte: Elaboração do autor.

Figura 39 – Estimativa da curva para dis



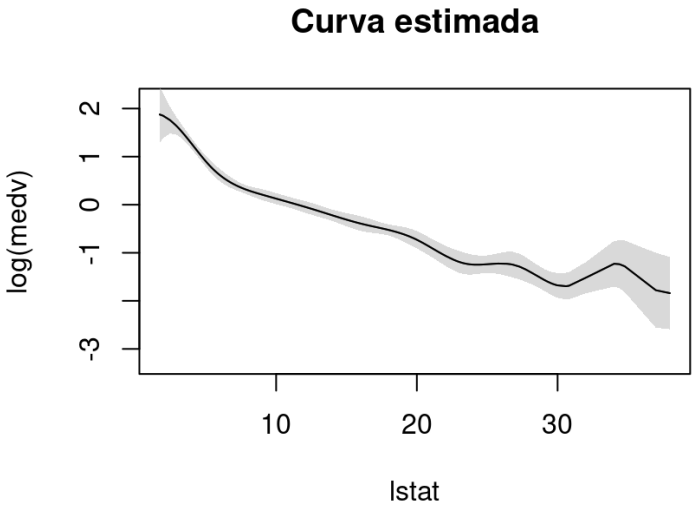
Fonte: Elaboração do autor.

Figura 40 – Estimativa da 2ª derivada para dis



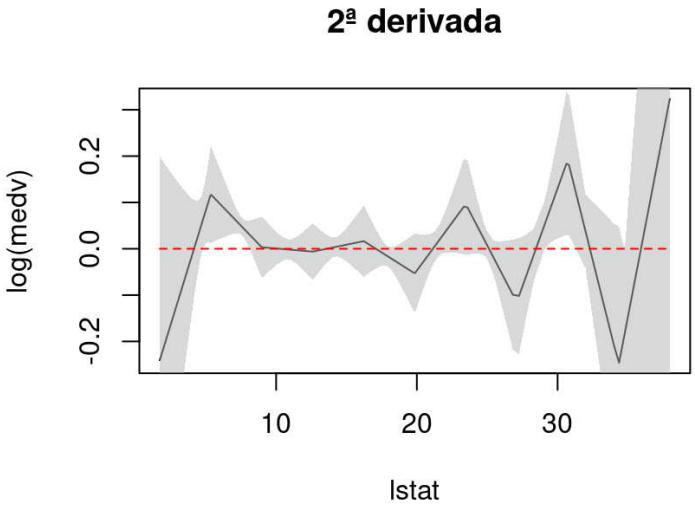
Fonte: Elaboração do autor.

Figura 41 – Estimativa da curva para lstat



Fonte: Elaboração do autor.

Figura 42 – Estimativa da 2ª derivada para lstat



Fonte: Elaboração do autor.

Verificando as bandas de 95% de confiança da segunda derivada, pode-se obter o percentual de pontos das bandas que contêm o valor real zero. Os percentuais foram de 100%, 46.05% e 75.9%, para as variáveis *tax*, *dis* e *lstat*, respectivamente. Sendo assim, segundo a metodologia adotada neste trabalho, a covariável *tax* apresenta indícios de que possui uma relação linear com a variável resposta *medv*. Diante disso, o próximo passo

consiste no ajuste de um novo modelo onde a covariável tax passa a ser linear, formando o seguinte APLM:

$$\log(medv_i) = \beta_0 + \beta_1 tax_i + g_2(dis_i) + g_3(lstat_i) + \epsilon_i \tag{5.7}$$

Após ajustar o segundo modelo, o teste de razão de verossimilhanças foi aplicado com o objetivo de comparar os modelos (5.6) e (5.7).

- Teste 1:

$$H_0 : \text{O modelo (5.7) oferece um ajuste tão bom quanto o modelo (5.6)}$$
$$H_1 : \text{O modelo (5.7) não oferece um ajuste tão bom quanto o modelo (5.6)}$$

Tabela 8 – Boston: p-valores dos testes de razão de verossimilhanças

Teste	Estatística de teste	p-valor
1	0	1

Analisando o p-valor da Tabela 8 pode-se concluir que o p-valor é maior que 0.05, portanto, a hipótese nula não é rejeitada a um nível de 5% de significância no teste. Conclui-se que, pelo teste de razão de verossimilhanças, o modelo (5.7) oferece um bom ajuste aos dados mesmo sendo o modelo mais simples. Para reforçar esta conclusão, pode-se analisar o critério BIC, em que dá-se preferência ao modelo com menor BIC.

Tabela 9 – Boston: BIC dos modelos ajustados

Modelos	BIC
(5.6)	951.495
(5.7)	-26.03

Os valores apesentados na Tabela 9 indicam um menor valor do BIC para o modelo (5.7). Logo, este critério também permite a conclusão de que o modelo ideal para os dados é aquele cujo a variável tax é considerada como tendo uma relação linear:

$$\log(medv_i) = \beta_0 + \beta_1 tax_i + g_2(dis_i) + g_3(lstat_i) + \epsilon_i$$

Como resultado, tem-se para o modelo proposto:

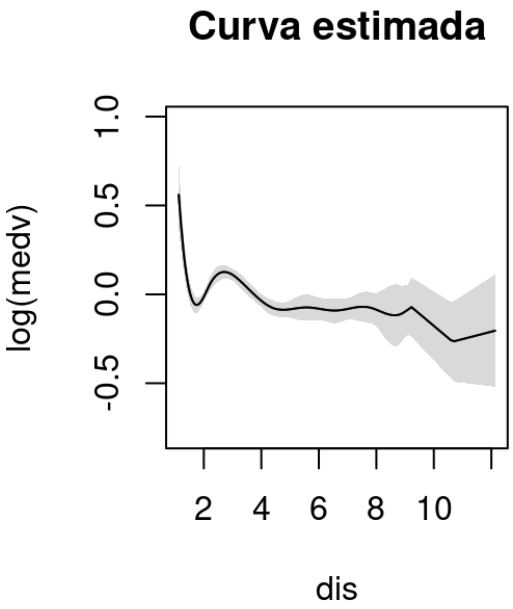
Tabela 10 – Boston: Estimativas de máxima verossimilhança penalizada dos parâmetros, seus erros-padrão e p-valores

Parâmetro	Estimativa	Erro-padrão	p-valor
λ_1	0.001	-	-
λ_2	1.000	-	-
σ^2	0.039	4.565e-06	-
β_0	3.289	1.042e-03	0
β_1	-0.0006	6.116e-09	0

Os valores da Tabela 10 mostram que, mantendo todas as covariáveis com o valor zero, o log da variável do valor médio das casas é de U$ 3.29 (em milhares de dólares), aproximadamente. Logo, o valor médio dos preços corresponde à U$ 26.83 (em milhares de dólares). Em relação a taxa de imposto, tem-se que mantendo as demais variáveis constantes, a cada uma unidade acrescida na taxa de imposto, reduz U$ 0.0006 (em milhares de dólares) no log do valor médio das casas. O teste de significância foi aplicado nos parâmetros $\boldsymbol{\beta}$. A maneira como o teste foi aplicado está detalhada na seção 5.2. Os p-valores apresentados na Tabela 10 são todos menores que 0.05, logo, a um nível de 5% de significância, todos os parâmetros são significativamente diferentes de zero.

Em relação a parte não paramétrica do modelo, obteve-se as curvas para as variáveis dis e lstat, apresentadas nas Figuras 43 e 44.

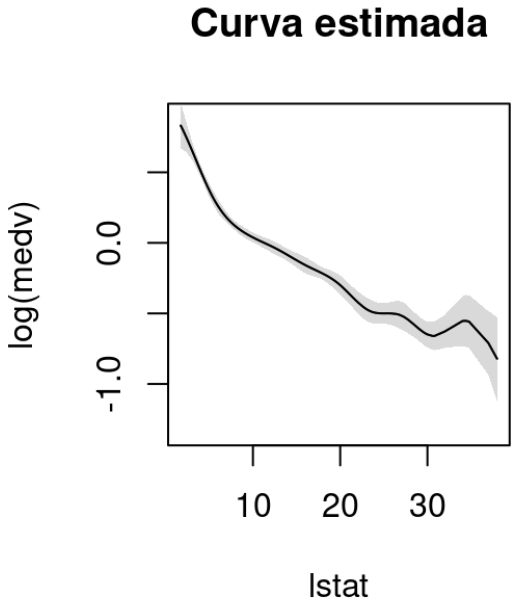
Figura 43 – Estimativa não paramétrica para a variável dis



Fonte: Elaboração do autor.

Analisando a curva estimada para a variável *dis*, pode-se perceber que mantendo as demais variáves inalteradas, as casas mais próximas aos centros, ou seja, com as menores distâncias, são em média mais caras. Para as distâncias maiores que quatro unidades (não é informado a unidade de medida das distâncias), a curva não apresenta grandes oscilações, o que indica que os valores médios das casas não sofrem muitas alterações a partir de uma determinada distância aos centros.

Figura 44 – Estimativa não paramétrica para a variável lstat

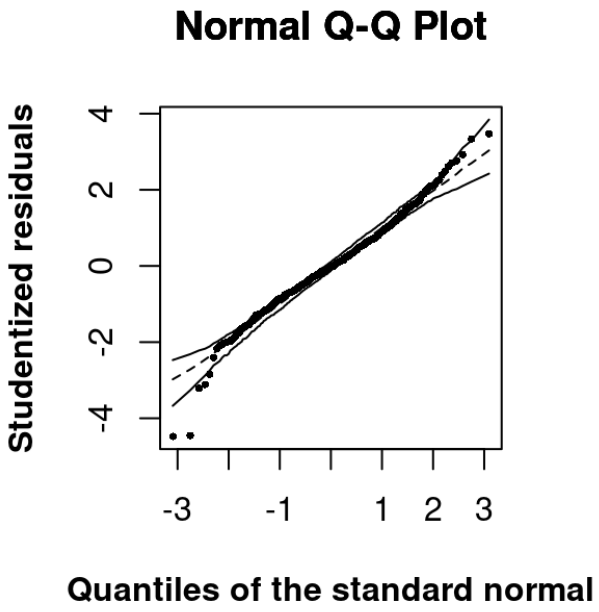


Fonte: Elaboração do autor.

Analisando a curva estimada para a variável *lstat*, pode-se perceber que mantendo as demais variáves inalteradas, as casas com menores percentuais de população de status socioeconômico mais baixo, são em média mais caras. Com isso, as casas cujos os moradores são mais desfavorecidos em termos sociais e econômicos possuem, em média, menores preços.

Para validar o modelo ajustado, foi feito o gráfico de envelope dos resíduos.

Figura 45 – Boston: Envelope dos resíduos



Fonte: Elaboração do autor.

Pelo gráfico envelope, tem-se que cerca de 1.97% dos pontos estão além dos limites das bandas de confiança. Logo, por este percentual ser menor que 5%, pode-se concluir que o modelo (5.7) apresenta um bom ajuste aos dados.

6 CONCLUSÃO

Através deste trabalho foi possível construir uma metodologia, ainda não vista na literatura, que é capaz de selecionar APLMs por meio da análise das derivadas de segunda ordem das curvas estimadas. Os resultados da simulação apresentados na Seção X, demostram que é possível identificar o tipo de relação (linear ou não linear) que as covariáveis de um modelo aditivo estabelecem com a variável resposta. Sendo assim, para verificar quais covariáveis pertecem a parte paramétrica e não paramétrica de um APLM, adotando a metodologia apresentada neste trabalho, deve-se primeiramente ajustar um modelo aditivo, considerando que todas as variáveis assumem uma relação não linear.

Obtidas as derivadas de segunda ordem das curvas estimadas por este modelo, bem como as suas bandas de confiança, testa-se se estas são retas em zero, pois este é o resultado esperado para variáveis com relações lineares. Para verificar isto, são calculadas as proporções dos pontos das bandas de confiança que incluem o valor zero, sendo que em casos de proporções maiores que 85%, pode-se considerar a covariável como sendo linear e ajustar um novo modelo acatando essas mudanças em relação a parte do modelo em que as covariáveis estarão. Vale mencionar aqui a sugestão de que a maneira escolhida para avaliar os pontos das bandas de confiança deve ser melhor explorada. Um ponto que pode ser analisado em trabalhos futuros é a possível necessidade de aplicação da correção de Bonferroni, que é um método estatístico que ajusta o nível de significância de testes de hipótese múltiplos. Por fim, o modelo aditivo e os outros modelos ajustados podem ser comparados via teste de razão de verossimilhanças e BIC, com a finalidade de encontrar o modelo que oferece melhor ajuste aos dados.

Em relação ao teste de razão de verossimilhanças, verificou-se que, devido as penalizações presentes em modelos como os APLMs, os resultados dos testes, apesar de satisfatórios em relação ao que era esperado, são inválidos pois as estatísticas de teste calculadas não convergiram para uma distribuição qui-quadrado. Diante disso, como mencionado, a proposta alternativa passou a ser o uso do Bootstrap para a aplicação destes testes. A aplicação deste método no presente estudo foi inviabilizada devido ao elevado custo computacional associado à reestimação dos parâmetros dos modelos em cada replicação. Entretanto, a aplicação do Bootstrap permanece como uma alternativa viável para estudos futuros, caso se disponha de maior capacidade computacional ou de algoritmos mais eficientes para a reestimação dos parâmetros.

Neste trabalho, também foi possível obter os estimadores de máxima verossimilhança penalizada, construir as curvas referentes à parte não paramétrica por meio de P-splines, calcular a variância dos estimadores através da matriz de informação de Fisher ou por Bootstrap quando a primeira opção torna-se inviável e obter o parâmetro de suavização λ por meio da validação cruzada. Apesar destas não serem abordagens inovadoras, estas

etapas são necessárias para a construção dos modelos estudados e geralmente não são abordadas de forma mais detalhada na literatura. Um exemplo deste diferencial presente neste trabalho são os resultados que demostram por meio de simulações as propriedades dos EMVPs para os APLMs. Sendo assim, além de evidenciarem a qualidade da metodologia apresentada, os resultados destas etapas contribuíram para a exploração mais ampla dos modelos APLMs.

Contudo, o estudo apresenta algumas limitações, como a não inversão da matriz de informação de Fisher em alguns casos, a abordagem apenas para modelos com a distribuição normal, entre outros, o que sugere a necessidade de investigações futuras sobre estes pontos mencionados.

Em suma, o trabalho realizado representa um avanço significativo no tema seleção de APLMs, uma vez que aborda uma metodologia inédita para isso: a seleção por meio da análise das derivadas de segunda ordem das curvas estimadas. O uso da metodologia em aplicações utilizando dados reais, gerou resultados que demonstram a importância destes modelos em diversas áreas do conhecimento, como a socioeconomia e a biologia. Sendo assim, pode-se afirmar que o presente trabalho oferece resultados importantes para estudos subsequentes sobre seleção de APLMs.

REFERÊNCIAS

DAVISON, A.C.; HINKLEY, D.V. **Bootstrap methods and their application**. New York: Cambridge University Press, 1997.

DE BOOR, Carl. **A Practical Guide to Splines**. Berlin: Springer, 1978.

EILERS, P.H.C.; MARX, B.D. Flexible smoothing with B-splines and penalties. **Statistical Science**, 11(2), p.89-121, 1996.

EUBANK,R.L. **Nonparametric Regression and Spline Smoothing**. 2ªedição. Marcel Dekker, 1999.

GREEN, P. J.; SILVERMAN, B. W. **Nonparametric regression and generalized linear models:** a roughness penalty approach. London; New York: Chapman and Hall, 1994.

HARRISON, D.; RUBINFELD, D.L. Hedonic prices and the demand for clean air. J. Environ. **Economics and Management 5**, 81–102, 1978.

HASTIE, T.; TIBSHIRANI, R. **Generalized Additive Models**. London: Chapman and Hall, 1990.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning:** data mining, inference, and prediction. 2nd ed. New York, NY: Springer, 2009. (Springer series in statistics).

IBACACHE-PULGAR, G.; PAULA, G. A.; CYSNEIROS, F. J. A. Semiparametric additive models under symmetric distributions. **Test**, v.22, p.103-121, 3 de out. de 2013.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. [s.l.]: Springer, 2013.

LIU, X.; WANG, L.; LIANG, H. Estimation and Variable Selection for Semiparametric Additive Partial Linear Models. **Stat Sin.**, 21 (3), p.1225-1248, 2011.

R Core Team (2024). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

RUPPERT, D.; WAND, M.P.; CARROL, R.J. **Semiparametric Regression**. New York: Cambridge University Press., 2003.

SASSI, Cecilia P et al. **Modelos de regressão linear múltipla utilizando os softwares R e statistica: uma aplicação a dados de conservação de frutas**. São Carlos: ICMC-USP. Disponível em: http://www.icmc.usp.br/biblio/relatorios_tecnicos.php. Acesso em: 28 fev. 2025. , 2012.

SCHWARTZ, J. Nonparametric smoothing in the analysis of air pollution and respiratory illness. **Canadian Journal of Statistics**, 22(4), p. 471–487, 1994.

SEVERINI, T. A. **Likelihood Methods in Statistics**. New York: Oxford University Press., 2000.

STARK, P. C.; RYAN, L. M.; McDONALD, J. L.; BURGE H. A. Using meteorologic data to model and predict daily ragweed pollen levels. **Aerobiologia**, 13, 177-184, 1997.

TORMAN, V. B. L.; COSTER, R.; RIBOLDI, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. **Revista HCPA**, 2012. Disponível em: https://seer.ufrgs.br/hcpa/article/view/29874i.

VANEGAS, L.H.; PAULA, G.A. An extension of log-symmetric regression models: R codes and applications. **Journal of Statistical Simulation and Computation**, 86, p.1709-1735, 2016.

VENABLES, W. N.; RIPLEY, B. D. Modern Applied Statistics with S. Fourth Edition. Springer, New York, 2002.

VIGOTTI, M.A.; ROSSI, G.; BISANTI, L.; ZANOBETTI, A.; SCHWARTZ, J. Short term effect of urban air pollution on respiratory health in Milan, Italy, 1980-1989. **Journal of Epidemiology and Community Health**, 50, S71-S75, 1996.

WAND M. (2018). SemiPar: Semiparametic Regression. **R package version 1.0-4.2**, <https://CRAN.R-project.org/package=SemiPar>.

WANG, B.; FANG, Y.; LIAN, H.; LIANG, H. Additive partially linear models for massive heterogeneous data. **Electronic Journal of Statistics**, v.13, n.1, p.391-431, 9 de fev. de 2019. Disponível em: <https://doi.org/10.1214/18-EJS1528>. Acesso em: 20 de nov. de 2023.

WOOD, S. N. **Generalized Additive Models:** an introduction with R. Second Edition. Boca Raton: Chapman and Hall, 2017.

WOOLDRIDGE, J. M. **Introductory Econometrics:** A Modern Approach. 5. ed. Mason: South-Western Cengage Learning, 2012.

YU, Y.; RUPPERT, D. Penalized spline estimation for partially linear single-index models. **Journal of the American Statistical Association**, 97, p.1042-1054, dez. de 2002.