**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Regressão spline de nós livres para modelagem de curvas de
crescimento multifásica**

Alex Rocha Soares

Dissertação apresentada para obtenção do título de Mestre
em Ciências, Área de concentração: Estatística e Experi-
mentação Agronômica

**Piracicaba
2019**

Alex Rocha Soares
Bacharelado em Matemática

**Regressão spline de nós livres para modelagem de curvas de
crescimento multifásica**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Prof. Dr. **CRISTIAN MARCELO VILLEGAS LOBOS**

Dissertação apresentada para obtenção do título de Mestre
em Ciências, Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Soares, Alex Rocha

Regressão spline de nós livres para modelagem de curvas de crescimento multifásica / versão revisada de acordo com a resolução CoPGr 6018 de 2011
Alex Rocha Soares. Piracicaba, 2019 .

63 p.

Dissertação (Mestrado) — USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1.Regressão spline 2.Crescimento Multifásicos 3.Localidade dos nós 4.Número de nós . I. Título.

DEDICATÓRIA

Dedico esse trabalho a duas pessoas de extrema importância no meu caminhar na pós-graduação: Patricia Iwagaki e Felipe Felix, graças a vocês eu iniciei meu caminho e conclui ele. Minha eterna gratidão.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer aos meus amigos Daniela, Gisele e Wesley, afinal, vocês estavam ao meu lado antes de entrar na graduação e continuam ao meu lado após meu mestrado. Vocês são fundamentais.

Gostaria de agradecer aos meus amigos: Cristiano, Donizetti, Denner Felipe, Juliana, Lucas Pamela, Patricia e Tais, por estarem comigo durante a minha jornada na graduação. Amo vocês! Agradeço também aos amigos que fiz na ESALQ (Raquel, Cayan, Bruna, Gustavo), vocês são pessoas extraordinárias!

Aos professores da UNESP Rio Claro e aos professores da ESALQ, por mostrarem sempre que ser professor vai além de uma sala de aula. Vocês são luz e exemplo!

Aos meus amigos da B2w (em especial a equipe do TI Financeiro) que me aguentaram nesse último ano de 2018, e fizeram com que cada dia fosse um dia melhor. Vocês estão bem alocados em meu coração.

Aos funcionários da ESALQ, em especial a minha amiga Solange, por diversas vezes ter me salvado! A Capes pelo apoio Financeiro.

E por fim, ao meu professor, orientador e querido colega da ESALQ, Cristian. Ter a paciência que o senhor teve não é para qualquer um. A admiração que tenho pela sua pessoa e pelo seu profissional é algo extremamente grande. Muito obrigado pela gigantesca ajuda, professor!

SUMÁRIO

RESUMO	7
ABSTRACT	8
LISTA DE FIGURAS	10
LISTA DE TABELAS	11
1 INTRODUÇÃO	13
2 REVISÃO BIBLIOGRÁFICA	15
2.1 Raça Hereford	15
2.2 Crescimento Multifásico	15
2.3 Regressão não linear	17
2.3.1 Estimação de parâmetros de regressão	18
2.4 <i>Spline</i>	19
2.4.1 Spline Linear	21
2.4.2 Spline Quadrático	21
2.4.3 Spline Cúbico	22
2.4.4 B-spline	23
2.4.5 P-spline	25
2.4.6 Critério de Validação Cruzada Generalizado	26
2.4.7 Critério de informação de Akaike	26
2.5 Suavização <i>spline</i>	27
3 MATERIAIS E MÉTODOS	29
3.1 Banco de dados	29
3.2 Quantidade e posicionamento dos nós	29
3.2.1 Método da busca cega através da seção dourada	30
3.3 Estudo de simulação	30
3.4 Aspectos computacionais	31
3.5 Numero máximo de nós com busca cega	31
4 RESULTADOS	33
4.1 Estudo de Simulação	33
4.1.1 Tamanho amostral 20	33
4.1.2 Tamanho amostral 50	35
4.1.3 Tamanho amostral 100	36
4.1.4 Numero fixo de nós com busca cega	38
4.1.5 Tamanho amostral 20	39
4.1.6 Tamanho amostral 50	41
4.1.7 Tamanho amostral 100	44
4.1.8 Comentários finais do estudo de simulação	46
4.2 Estudo de dados completos	47

4.2.1	Ajuste Linear	47
4.2.2	Ajuste Quadrático	48
4.2.3	Ajuste Cúbico	49
4.3	Estudo de dados médios	50
4.3.1	Ajuste Linear	50
4.3.2	Ajuste Quadrático	51
4.3.3	Ajuste Cúbico	52
4.4	Discussão	53
5	CONCLUSÃO	55
5.1	Trabalhos Futuros	55
	REFERÊNCIAS	57
	APÊNDICES	61

RESUMO

Regressão spline de nós livres para modelagem de curvas de crescimento multifásica

Neste trabalho, apresentamos os modelos de regressão *spline* de nós-livres como uma alternativa aos modelos não lineares utilizados em curvas de crescimento multifásico. Estudaremos o algoritmo de busca cega a través da seção dourada para escolher a melhor quantidade de nós e suas respectivas localidades. O pacote **freeknotspline** do software livre R foi utilizado para ajustar os modelos propostos. O critério de informação de Akaike foi usado para escolher o melhor modelo para diferentes graus do polinômio associados ao *spline*. Estudos de simulação foram realizados para entender melhor a posição dos nós, localidade e grau do polinômio relacionado aos modelos de regressão *spline* de nós-livres e como isto pode afetar a qualidade de ajuste do modelo. Com base no nosso estudo de simulação, propomos uma forma empírica de determinar o numero de nós, deixando que o algoritmo de busca escolha a posição dos nós. A metodologia é aplicada aos dados de crescimento multifásico de vacas da raça Hereford.

Palavras-chave: Regressão spline; Crescimento Multifásico; Localidade de nós; Número de nós

ABSTRACT

Free-Knot Spline Regression for Modeling Multiphase Growth Curve

In this work, we present the free-knot spline regression models as an alternative to the nonlinear models used in multiphase growth curves. We will study the blind search algorithm through the gold section to choose the best number of knot and their respective locations. The package `freeknotspline` of the free software R was used to fit the proposed models. The Akaike information criterion was used to choose the best model for different degrees of the polynomial associated with the spline. Simulation studies were performed to better understand the position of the knot, location and degree of the polynomial related to spline regression models of free-knots, and how this may affect the goodness of fit of the model. Based on our simulation study, we propose an empirical way of determining the number of knots, letting the search algorithm choose the position of the knots. The methodology is applied to the multiphase growth data of Hereford breed females.

Keywords: Spline Regression; Multiphasic Growth; Knot location; Number of Knot;

LISTA DE FIGURAS

2.1	Animal da raça Hereford	15
2.2	Exemplo de comportamento logístico difásico	17
2.3	Gráfico de uma função de grau um descontínua com nó em $x=5$	20
2.4	Exemplo de spline linear com 2 nós em $b_1 = 5$ e $b_2 = 7$	22
2.5	Exemplo de spline quadrático com 2 nós em $b_1 = 4$ e $b_2 = 7$	23
2.6	Bases spline para polinômios de grau 1 e 2 com nós equidistantes	24
2.7	Bases spline para polinômios de grau 3 e 4 com nós equidistantes	25
3.1	Gráfico de dispersão dos dados de crescimento de fêmeas da raça Hereford	29
3.2	Gráfico de dispersão dos dados médios de crescimento de fêmeas da raça Hereford	30
4.1	Ajuste com spline linear a dados simulados de tamanho amostral 20 e 15 nós	33
4.2	Ajuste com spline quadrático a dados simulados de tamanho amostral 20 e 15 nós	34
4.3	Ajuste com spline cúbico a dados simulados de tamanho amostral 20 e 15 nós	34
4.4	Ajuste com spline linear a dados simulados de tamanho amostral 50 e 46 nós	35
4.5	Ajuste com spline quadrático a dados simulados de tamanho amostral 50 e 46 nós	35
4.6	Ajuste com spline cubico a dados simulados de tamanho amostral 50 e 46 nós	36
4.7	Ajuste com spline linear a dados simulados de tamanho amostral 100 e 80 nós	37
4.8	Ajuste com spline quadrático a dados simulados de tamanho amostral 100 e 80 nós	37
4.9	Ajuste com spline cubico a dados simulados de tamanho amostral 100 e 80 nós	38
4.10	Gráfico de dispersão para tamanho amostral 20	39
4.11	Ajuste com spline linear para tamanho amostral 20 e 10 nós	40
4.12	Ajuste com spline quadrático para tamanho amostral 20 e 10 nós	40
4.13	Ajuste com spline cúbico para tamanho amostral 20 e 10 nós	41
4.14	Gráfico de dispersão de tamanho amostral 50	42
4.15	Ajuste com spline linear para tamanho amostral 50 e 10 nós	42
4.16	Ajuste com spline quadrático para tamanho amostral 50 e 10 nós	43
4.17	Ajuste com spline cúbico para tamanho amostral 50 e 10 nós	43
4.18	Gráfico de dispersão de tamanho amostral 100	44
4.19	Ajuste com spline linear para tamanho amostral 100 e 10 nós	45
4.20	Ajuste quadrático com tamanho amostral 100 e 10 nós	45
4.21	Ajuste com spline cúbico para tamanho amostral 100 e 10 nós	46
4.22	Ajuste Linear a dados completos utilizando 4 nós	47

4.23	Ajuste quadrático a dados completos utilizando 4 nós	48
4.24	Ajuste cúbico a dados completos utilizando 4 nós	49
4.25	Ajuste Linear a dados medios utilizando 4 nós	50
4.26	Ajuste quadrático a dados médios utilizando 4 nós	51
4.27	Ajuste cúbico a dados médios utilizando 4 nós	52

LISTA DE TABELAS

4.1	Localização dos nós para spline linear a dados completos com valores de pesos preditos	48
4.2	Localização dos nós para spline quadrático a dados completos com valores de pesos preditos	49
4.3	Localização dos nós para spline cúbico a dados completos com valores de pesos preditos	49
4.4	Valores de AIC e λ para ajustes utilizando dados completos	50
4.5	Localização dos nós para <i>spline</i> linear a dados médios com valores preditos	51
4.6	Localização dos nós para ajuste quadrático a dados médios com valores preditos	51
4.7	Localização dos nós para ajuste cúbico a dados médios com valores preditos	52
4.8	Valores de AIC e λ para ajustes utilizando dados médios	52

1 INTRODUÇÃO

As curvas de crescimento possuem grande aplicabilidade nas áreas agronômicas e biológicas podendo representar diversos fatores, como por exemplo, relacionar uma sequência de medidas de peso e idade de animais ou vegetais. Essas curvas, em geral, possuem forma sigmoide com uma assíntota superior e, às vezes, apresentam uma assíntota inferior. Segundo RAO (1958), dois problemas de interesse podem ser abordados: um é a construção de um modelo caracterizando o crescimento de um organismo individual em um determinado período de tempo e o outro, é a comparação das características de crescimento em diferentes condições (sexo, raça, tratamentos, entre outros).

Em alguns casos, o crescimento do objeto em estudo apresenta um comportamento característico de forma a identificar ciclos, no qual é possível dividir a curva em fases. Então, torna-se necessário utilizar modelos de crescimento multifásico, por contemplarem parâmetros exclusivos para cada fase. O que na realidade se faz, é considerar em um único modelo, dois ou mais submodelos para explicar o crescimento animal/vegetal como um todo.

Uma forma comumente utilizada, em diversos trabalhos, de se ajustar curvas de crescimento multifásica é por meio da soma de modelos não lineares, como: Modelo Gompertz, logístico, Richards, Von Bertalanffy, entre outros. Todavia, o estudo de outros métodos não usuais nessa área, podem ser interessantes e às vezes mais eficientes. No caso desse trabalho, será abordado estudo de funções *spline* para modelagem de curvas de crescimento, no qual a ênfase em questão esta no posicionamento e na localidade dos nós.

A aplicação dos *splines* como forma de representar crescimento animal ainda é recente, entretanto, devido à sua grande utilidade em situações práticas, vem contribuindo de maneira mais assídua para representar tais fatores. resultando um grande crescimento em sua utilização. Dentre os benefícios na utilização das funções *splines* destaca-se a possibilidade de ajustar polinômios de baixo grau em segmentos curtos da trajetória de crescimento (uma vez selecionada o local e a quantidade de nós) e, desta forma, ser mais flexível no ajuste das flutuações estacionárias de crescimento animal. Outro ponto relevante é sua fácil implementação computacional.

Diante do exposto, o objetivo desse trabalho é ajustar funções *spline* a dados de crescimento de animais da raça Hereford, tanto para dados completos quanto para dados médios, determinar a melhor localização e o melhor posicionamento dos nós, e verificar a qualidade e benefício desse tipo de ajuste.

2 REVISÃO BIBLIOGRÁFICA

2.1 Raça Hereford

De acordo com A.B.H.B. (Associação Brasileira de Hereford e Braford, 2017), a raça Hereford é originária do condado inglês de Herefordshire, zona com condições climáticas favoráveis para a produção de pastos superiores. O primeiro exemplar chegou ao Brasil no início do século XX, e por possuir alta adaptabilidade aos mais diversos tipos de ambiente, encontra-se distribuído em diversos outros países.

Segundo os mesmos autores, a pelagem da raça Hereford caracteriza-se por ser vermelha e de cara, ventre, extremidade da cauda e partes inferiores totalmente brancas, como representado na Figura 2.1, em que essa pelagem recebe o nome de "pampa". Essa raça é famosa também, devido ao seu tamanho, resistência e aptidão para a produção de carne. Devido a sua conformação e capacidade de engordar é considerada razoavelmente rústica e prolífica. Sua vida útil gira em torno de 12 e 15 anos aproximadamente.

Figura 2.1. Animal da raça Hereford

Fonte: A.B.H.B. (Associação Brasileira de Hereford e Braford, 2017)

2.2 Crescimento Multifásico

O termo curvas de crescimento, segundo FITZHUGH (1976), evoca a imagem de uma curva sigmoide, que descreve uma sequência de medidas de tamanho ao longo do tempo de vida. O estudo de curvas de crescimento, segundo OLIVEIRA *et al.* (2000) geralmente é feito através de ajuste de funções não lineares pois, dessa maneira, é possível sintetizar informações de todo o período de vida dos animais, ou seja, um conjunto de informações em série de peso por idade, em um pequeno conjunto de parâmetros interpretáveis biologicamente, facilitando o entendimento do fenômeno. De acordo com MAZZINI (2001), as funções mais utilizadas para descrever o crescimento de animais são:

$$\text{Gompertz } Y = ae^{-be^{kt}};$$

$$\text{Brody } Y = a - be^{-kt};$$

$$\text{logístico } Y = a(1 + e^{b-kt})^{-1};$$

$$\text{Von Bertalanfly } Y = a(1 - be^{-kt})^3.$$

em que Y é a variável dependente (por exemplo: peso de animais, comprimento de planta, teor de água em sementes), a é a assíntota do modelo que representa o valor de estabilização da variável dependente em relação ao tempo t , b corresponde ao parâmetro de locação (parâmetro não possui interpretação biológica), k representa a taxas de crescimento, t designa o tempo.

VIEIRA e MISCHAN (1976) compararam curvas de crescimento e ressaltaram a importância dos modelos logístico e Gompertz em aplicações biológicas.

BERGAMASCO *et al.* (2001) fizeram a comparação dos modelos logísticos, Gompertz e Brody para dados de crescimento de fêmeas da raça Holandesa do nascimento até os dois anos de idade em que, todos os modelos se ajustaram bem aos dados, entretanto, o modelo logístico apresentou melhor facilidade computacional, um bom coeficiente de determinação R^2 .

No âmbito de comparação de curvas, LOBO *et al.* (2006) ajustaram os modelos logístico, Gompertz, Von Bertalanffy, Brody e Richards a dados de peso-idade de ovinos da raça Santa-Inês em que o modelo Richards foi o que melhor descreveu o conjunto de dados.

Em alguns casos, o crescimento do objeto em estudo apresenta um comportamento característico de forma a identificar ciclos, em que é possível dividir a curva em fases. Então torna-se necessário utilizar modelos de crescimento multifásico, por contemplarem parâmetros exclusivos para cada fase. O que na realidade se faz, é considerar em um único modelo, dois ou mais submodelos para explicar o crescimento animal/vegetal como um todo.

Uma forma comumente utilizada, em diversos trabalhos, de se ajustar curvas de crescimento multifásica é por meio da soma de modelos não lineares, como: Modelo Gompertz, Logístico, Richards, Von Bertalanffy, entre outros.

Na figura 2.2 esta representado o comportamento de um modelo logístico difásico. Note que, a curva inicia um crescimento na vizinhança do número dois, em seguida, aparenta um comportamento estável, voltando a crescer na vizinhança do número cinco, ou seja, esse crescimento se apresenta em fases. A equação que descreve a curva é dada por:

$$Y(t) = \frac{1}{1 + e^{10-6t}} + \frac{2}{1 + e^{15-3t}}. \quad (2.1)$$

Note que a formação final é a soma de duas equações logísticas, separadas em duas fases. A primeira fase é dada pela expressão:

$$f_1(t_i) = \frac{1}{1 + e^{10-6t_i}}$$

e a segunda fase é dada por:

$$f_2(t_i) = \frac{2}{1 + e^{15-3t_i}}.$$

Figura 2.2. Exemplo de comportamento logístico difásico

Os numeradores representam os valores máximos adquiridos em cada fase e sua soma representa o valor máximo atingido pela curva.

Trabalhando com dados de crescimento multifásico, MENDES *et al.* (2008), fizeram um estudo com dados de fêmeas da raça Hereford, em que ajustaram o modelo logístico com diferentes estruturas de erro. Como conclusão, afirmou-se que o modelo logístico difásico com estrutura de erro auto regressivo de segunda ordem AR(2) apresentou boa qualidade em relação aos aspectos estatísticos e biológicos, sendo então, recomendados para descrever a curva de crescimento do rebanho em questão, já o modelo logístico difásico com estrutura de erro auto regressivo de primeira ordem AR(1) apresentou boa qualidade de ajuste ao se considerar o conjunto de dados individuais, todavia, não produziu estimativas para os parâmetros condizentes com a realidade. Em todos esses casos, a análise de dados foi feita considerando uma função que é combinação não linear dos parâmetros do modelo, e por este motivo são denominados de regressão não linear.

2.3 Regressão não linear

O termo regressão foi introduzido na linguagem estatística por Pearson, para designar a expressão de uma variável como função de uma ou mais variáveis (CECON *et al.*, 2012). Na estatística uma das matérias mais comuns é avaliar uma possível relação entre uma variável dependente com uma ou mais variáveis independentes (MAZUCHELI e ACHCAR, 2008). Para estudar essa relação, são conhecidas duas classes de modelos: os lineares e os não lineares.

De acordo com RATKOWSKI (1993) modelos de regressão que descrevem processos naturais são, geralmente, modelos não lineares nos parâmetros e, segundo BATES e WATTS (1988), um modelo é não linear nos parâmetros se ao menos uma das derivadas parciais da função em relação aos parâmetros do modelo, depender de pelo menos um dos parâmetros.

Segundo SMYTH (2006), um modelo de regressão não linear pode ser definido como:

$$Y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i \quad i = 1, \dots, n \quad (2.2)$$

em que Y_i representa a variável resposta da i -ésima observação; f é uma função conhecida que depende das variáveis explicativas x_i e $\boldsymbol{\theta}$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ é um vetor de parâmetros desconhecidos; ε_i é o erro aleatório não observado com média zero e variância constante σ^2 .

Muitos são os modelos de regressão não linear utilizados frequentemente para descrever fenômenos, alguns são: logístico, Gompertz, Von Bertalanffy entre outros.

2.3.1 Estimação de parâmetros de regressão

Consideremos o modelo (2.2), a estimação dos parâmetros da função é feita utilizando o processo de mínimos quadrados ordinários minimizando a soma de quadrados de resíduo, dada por:

$$\min_{f: \mathbb{R} \rightarrow \mathbb{R}} \left\{ \sum_{i=1}^n (y_i - f(x_i, \boldsymbol{\theta}))^2 \right\}. \quad (2.3)$$

Rescrevendo a equação (2.3) com uma penalização, obtém-se:

$$S(f) = \min_{f: \mathbb{R} \rightarrow \mathbb{R}} \left\{ \sum_{i=1}^n (y_i - f(x_i, \boldsymbol{\theta}))^2 + \phi(f) \right\}, \quad (2.4)$$

em que ϕ é uma função definida no espaço de Sobolev com objetivo de penalizar a função a fim de deixá-la mais suave MENDONZA (2017).

Para casos especiais, em que o objetivo é estimar uma classe específica de funções é necessário determinar um espaço restrito, para isso, considere f uma função definida em um conjunto aberto $\Omega \in \mathbb{R}^n$, sendo representada por $L^p(\Omega)$. Quando $p = 2$ e o estudo é feito em um intervalo fechado $[a, b]$ temos o chamado Espaço de Hilbert, da qual a norma da função em um intervalo $[a, b]$ é representada por:

$$S(f) = \|f\|_{L^2([a,b])} = \left(\int_a^b |f(x)|^2 dx \right)^{\frac{1}{2}}. \quad (2.5)$$

Dado o espaço de Hilbert, é possível definir o espaço de Sobolev. Esse espaço é formado por todas as funções contínuas $f^m \in L^2[a, b]$ em que f^m é a m -ésima derivada da função f , ou seja, o espaço de Sobolev, representado por W_2^m é definido por:

$$W_2^m[a, b] = \{f : f^m \text{ é absolutamente contínua, e } f^m \in L^2[a, b]\}.$$

Portanto, pode se reescrever (2.4) da seguinte forma:

$$S(f) = \min_{f \in W_2^2[a,b]} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx \right\}. \quad (2.6)$$

Para RUPPERT *et al.* (2003b), essa forma apresentada é formada por duas partes, a primeira é formada pela qualidade do ajuste (mínimos quadrados) e a segunda é a suavidade do ajuste. O parâmetro λ indicado, é de extrema importância, uma vez que, dependendo do seu valor, pode-se ter uma curva muito próxima aos dados quanto muito longe dos mesmos. O critério de validação cruzada e o critério de validação cruzada generalizado é uma das formas de se determinar o melhor valor de λ (GOLUB *et al.*, 2012).

2.4 Spline

A origem do nome spline vem de uma régua elástica, utilizada para desenhos de engenharia. A formulação matemática foi desenvolvida na década de 60, todavia, há registros que comprovem que era utilizada desde o século passado (ROGERS e ADAMS, 1990). Segundo THOMSON e EMERY (2014), as funções splines são muito utilizadas em geociências, pois este método gera curvas mais suaves ao mesmo tempo tentando honrar ao máximo de dados, mas mesmo assim, não é um interpolador exato. Uma outra vantagem da utilização de spline, é a fácil implementação computacional.

A idéia geral, consiste em interpolar um polinômio que passe pelos pontos do plano. Todavia, por esses pontos, pode-se interpolar infinitos polinômios de graus diferentes. Com o intuito de ter um melhor ajuste polinomial, essa interpolação é feita por partes, nas quais os pontos de ruptura (ou quebra) são chamados de nós.

A abordagem do número de nós é algo de extrema importância para o estudo de spline, uma vez que, se o número de nós for muito grande estará se aproximando muito de um interpolador exato, e se esse número for muito pequeno a curva ajustada pode não ser a melhor. A localização dos nós também é algo a ser pensado a fim de determinar o melhor ajuste. O estudo sobre a quantidade e a localidade dos nós ainda é algo que desperta diferentes pensamentos no assunto.

Segundo BURCHARD (1974) uma das formas para análise do problema é tratar a localização dos nós como parâmetros. Segundo LINDSTROM (1999) essa abordagem apresenta duas grandes desvantagens: a primeira é que encontrar o local ótimo para os nós é algo muito difícil e segundo, que esta relacionado com a primeira, é que o vetor de nós, inclui valores replicados que podem gerar um comportamento não suave da curva prevista.

SPIRITI *et al.* (2018) discutem a relação entre o número e a posição dos nós. Os autores apresentam uma técnica de busca cega (pesquisa estocástica) para determinar a quantidade, apresenta também uma forma de separar o plano em grades e determinar a

melhor localização dos nós, eles apresentam essa técnica como técnica de seção dourada. Por fim, é apresentado uma equação para determinar a localidade e o posicionamento dos nós.

Considere a seguinte função $h(x)$, em que D representa o grau do polinômio, K o número de nós e b_k o valor do nó:

$$h(x) = \sum_{j=0}^D \beta_{0j} x^j + \sum_{k=1}^K \sum_{j=0}^D \beta_{kj} (x - b_k)_+^j,$$

em que

$$(x - b_k)_+^j = \begin{cases} (x - b_k)^j & x \geq b_k \\ 0, & x < b_k. \end{cases}$$

Percebe-se que para cada intervalo entre os valores dos nós, tem-se um polinômio de grau D . Como exemplo, tome uma função de grau um ($D = 1$) e com apenas um nó. Logo, $h(x)$ pode ser representada por:

$$h(x) = \begin{cases} \beta_{00} + \beta_{01}x + \beta_{10} + \beta_{11}(x - b_1) & x \geq b_1 \\ \beta_{00} + \beta_{01}x & x < b_1. \end{cases}$$

Em ambas as partes separadas pelos nós, tem-se uma função polinomial, e como não possui nenhuma restrição, essa função é contínua. Todavia, ao se analisar o valor dessa função no nó, obtém-se:

$$h(b_1) = \begin{cases} \beta_{00} + \beta_{01}b_1 + \beta_{10} & x \geq b_1 \\ \beta_{00} + \beta_{01}b_1 & x < b_1, \end{cases}$$

ou seja, a função não é contínua no nó.

Graficamente, esta representado na figura 2.3 um polinômio de grau um, com nó em $x = 5$. Perceba o salto apresentado no valor do nó. A não continuidade nos nós é uma das diferenças entre polinômio por partes e splines. A equação geral da função spline, segundo OLIVEIRA (2011), é dada por:

$$h(x) = \sum_{j=0}^D \beta_{0j} x^j + \sum_{k=1}^K \beta_k (x - b_k)_+^D,$$

em que K representa o número de nós, D o grau do polinômio, b_k o valor do nó e

$$(x - b_k)_+^D = \begin{cases} (x - b_k)^D, & x \geq b_k \\ 0, & x < b_k. \end{cases}$$

Quando D assume alguns valores particulares, temos alguns splines polinomiais mais conhecidos, como é o caso do linear ($D=1$), quadrático ($D=2$) e cúbico ($D=3$). A seguir apresentamos mais detalhes de um spline linear, quadrático e cúbico.

Figura 2.3. Gráfico de uma função de grau um descontínua com nó em $x=5$

2.4.1 Spline Linear

De acordo com HARRELL (2001), a forma mais simples de spline é o spline linear, que é caracterizada como um conjunto de funções lineares com diferentes inclinações em cada intervalo definido pelos nós. A equação geral de um spline linear, no qual K representa o número de nós é dada por:

$$h(x) = \beta_{00} + \beta_{01}x + \sum_{k=1}^K \beta_k(x - b_k)_+ \quad (2.7)$$

em que:

$$(x - b_k)_+ = \begin{cases} (x - b_k) & x \geq b_k \\ 0, & x < b_k. \end{cases}$$

Como forma de exemplificar um spline linear (veja figura 2.4), considere um número de nós iguais a 2. A função será dada por:

$$h(x) = \begin{cases} \beta_{00} + \beta_{01}x & x < 5 \\ \beta_{00} + \beta_{01}x + \beta_1(x - 5) & 5 \leq x \leq 7 \\ \beta_{00} + \beta_{01}x + \beta_1(x - 5) + \beta_2(x - 7) & x > 7. \end{cases}$$

Nota-se que na conexão dessas funções, encontradas nos nós, é apresentado um comportamento não suavizado, em que acontecem mudanças abruptas na direção do segmento de reta. Com o intuito de obter polinômios que possuem um comportamento mais suave uma forma é olhar para graus maiores do que um.

Figura 2.4. Exemplo de spline linear com 2 nós em $b_1 = 5$ e $b_2 = 7$

2.4.2 Spline Quadrático

Equacionalmente, os splines possuem a mesma base de funções, o que diferencia o spline linear para o quadrático é grau do polinômio. A ideia do spline quadrático é ser mais suave do que o spline linear, uma vez que a função quadrática não possui pontos de mudança abrupta de sinal. Segue a ideia da equação geral do spline, restrita ao fato do polinômio possuir grau 2:

$$h(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \sum_{k=1}^K \beta_k(x - b_k)_+^2, \quad (2.8)$$

em que

$$(x - b_k)_+^2 = \begin{cases} (x - b_k)^2, & x \geq b_k \\ 0 & x < b_k. \end{cases}$$

Segue, na figura 2.5, uma ilustração do comportamento de um spline quadrático, em que a lei da função é dada por:

$$h(x) = \begin{cases} -\frac{1}{2}x^2 + 2x + 2, & 1 \leq x \leq 4 \\ -\frac{1}{2}x^2 + 2x + 2 + (x - 4)^2 & 4 < x \leq 7 \\ -\frac{1}{2}x^2 + 2x + 2 + (x - 4)^2 + (x - 7)^2 & x > 7. \end{cases} \quad (2.9)$$

Figura 2.5. Exemplo de spline quadrático com 2 nós em $b_1 = 4$ e $b_2 = 7$

2.4.3 Spline Cúbico

Na literatura, HARRELL (2001) e THOMSON e EMERY (2014) entre outras, afirmam que um dos splines com maior aplicabilidade é o spline cúbico, que consiste no ajuste de polinômios de ordem três para cada intervalo determinado pelos nós. Esse spline pode ser dividido em duas categorias: spline cúbico restrito (se o polinômio determinado antes do primeiro nó e depois do último, forem funções lineares) e spline cúbico não restrito (a função determinada antes e depois do primeiro nó são polinômios de grau maior do que um). Nota-se ainda que, como o spline cúbico possui grau três, sua forma geral é:

$$h(x) = \sum_{j=0}^3 \beta_{0j} x^j + \sum_{k=1}^K \beta_k (x - b_k)_+^3,$$

em que a primeira derivada em relação a x é dada por:

$$h'(x) = \beta_{01} + 2\beta_{02}x + 3\beta_{03}x^2 + \sum_{k=1}^K 3\beta_k (x - b_k)_+^2$$

e ainda, a segunda derivada é dada por:

$$h''(x) = 2\beta_{02} + 6\beta_{03}x + \sum_{k=1}^K 6\beta_k (x - b_k)_+.$$

Como $h(x)$ possui a segunda derivada contínua, tem-se que ela é uma função suave. Por ser a função de grau menor que possui a propriedade de segunda derivada contínua, tem-se uma maior aplicabilidade, uma vez que, com esse grau menor, consegue-se ajustar a dados sem ter que elevar tanto o grau do polinômio.

2.4.4 B-spline

B-spline é uma abreviação de splines básicos (em inglês, *basis* ou *basics splines*). O termo *basics* refere-se à aplicação de uma transformação feita na variável x antes de iniciar o ajuste do modelo (DE BOOR, 1978). Considere $\Omega_\infty = \{b_j\}_{j \in \mathbb{Z}}$ uma sequência não decrescente de nós. Em EILERS e MARX (1996), o i -ésimo B-spline de ordem k para a sequência de nós Ω_∞ é definida por:

$$B_j^k(t) = \frac{t - b_j}{b_{j+k} - b_j} B_j^{k-1}(t) + \frac{b_{j+k+1} - t}{b_{j+k+1} - b_{j+1}} B_{j+1}^{k-1}(t)$$

Como corolário dessa definição, tem-se que:

- $B_j^k(t) = 0$ para todo $t \notin [x_j, x_{j+k}]$
- $B_j^k(t) > 0$ para todo $t \in (x_j, x_{j+k})$
- $\sum_{j \in \mathbb{Z}} B_j^k = 1$, isto é, a sequência B-spline consiste em funções que somam até 1 e fornece uma partição da unidade.

Pela definição e resultados seguintes tem-se que: Uma função spline pode ser escrita como combinação linear de B-splines RUPPERT *et al.* (2003b), ou seja:

$$s(t) = \sum_{j \in \mathbb{Z}} \beta_j B_j^k(t).$$

O comportamento dessa base varia de acordo com o grau do polinômio. As figuras 2.6 e 2.7 representam possíveis bases para polinômios e seus respectivos graus.

Figura 2.6. Bases spline para polinômios de grau 1 e 2 com nós equidistantes

Figura 2.7. Bases spline para polinômios de grau 3 e 4 com nós equidistantes

Importante notar que essas representações baseiam-se em nós equiespaçados. ROBBINS *et al.* (2005) compararam avaliações genéticas de dados de crescimento em bovinos da raça Gelbvich utilizando funções splines lineares com três nós com o intuito de determinar a praticidade de implementação de modelos longitudinais. De acordo com os autores, os modelos longitudinais forneceram uma melhor modelagem das curvas em estudo. Outra aplicação de spline vem por HUISMAN *et al.* (2002), em que realizam estudos com polinômio ortogonal de Legendre cúbico e spline cúbica com o propósito de avaliar qual modelo melhor se ajusta aos dados de peso-idade de suínos. Como conclusão, os autores não identificaram qual modelo foi melhor uma vez que o critério Bayesiano foi superior para o polinômio ortogonal, entretanto, o critério de informação de Akaike foi superior para o spline. OLIVEIRA (2011) fez a utilização de regressão spline para descrever curvas de crescimento em ovinos cruzados, em que comparou com modelos não lineares usuais. Como conclusão, obteve que funções splines mostraram-se mais adequadas do que os modelos não lineares. De acordo com OLIVEIRA (2011), spline, além de curvas de crescimento, possuem aplicações em ajuste de curvas de lactação, modelagem de efeitos não genéticos, avaliações genéticas, avaliações de tendência genética e também podem ser utilizadas em outras áreas, tais como engenharia, ciências agrárias, computação e arquitetura. A fim de evitar um superajuste, uma nova função, chamada de função de penalização, pode ser utilizada sobre os B-spline.

2.4.5 P-spline

Notou-se que, ao se modelar uma função como soma de B-splines, a integral da segunda derivada ao quadrado pode ser expressa como uma função quadrática nos coeficientes que representam o somatório (EILERS e MARX, 1996). A idéia do autor foi

tomar uma base como muitos B-spline e utilizar uma penalidade discreta, em que, a forma de estabelecer a rugosidade da curva é:

$$R = \int_a^b [f''(x)]^2 dx, \quad (2.10)$$

em que $[a, b]$ é o intervalo em que esta se estudando os dados. Uma vez que esta se referindo a $f(x)$ como soma de B-splines, tem-se que $f(x) = \sum_{j \in \mathbb{Z}} a_j B_j(x)$. A idéia em si do P-spline é descartar a segunda derivada utilizando B-splines igualmente espaçados. EILERS e MARX (1996) utilizam uma forma discreta de penalização:

$$S(f) = \sum_{i=1}^n \left(f(x_i) - \sum_{j \in \mathbb{Z}} a_j B_j(x_i) \right)^2 + \lambda \sum_{j \in \mathbb{Z}} (\Delta^i a_j)^2, \quad (2.11)$$

em que i é a i -ésima diferença dos coeficientes das bases B-spline e Δ^i é o operador de diferenças. Δ^i pode ser escrita como uma matriz $[(m+1+p-d) \times (m+1+p)]$, em que $m \in \mathbb{Z}$ e $d = p-1$ no qual d representa a ordem do polinômio.

Então, $S(f)$ pode ser escrita na sua forma matricial da seguinte forma:

$$S(f) = (y - \mathbf{B}a)'(y - \mathbf{B}a) + \lambda a' \mathbf{P}_d a, \quad (2.12)$$

em que $\mathbf{P}_d = (\Delta^d)'(\Delta^d)$.

2.4.6 Critério de Validação Cruzada Generalizado

Uma definição se torna necessária antes de descrever o Critério de Validação Cruzada Generalizado. Essa definição pode ser encontrada em RUPPERT *et al.* (2003a).

Considere o modelo de regressão linear em termos de funções bases B-spline

$$\mathbf{y} = \mathbf{B}a + \boldsymbol{\varepsilon} \quad (2.13)$$

Então, o ajuste dos valores de um P-spline utilizando o critério de mínimos quadrados penalizados é dado por:

$$\hat{y} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda \mathbf{P}_d)^{-1} \mathbf{B}' = H_\lambda y, \quad (2.14)$$

em que a matriz $\mathbf{H}_\lambda$ é chamada de matriz de suavização. O conceito de validação cruzada parte da idéia de retirar uma observação e ajustar o modelo aos demais dados, e então calcular o quadrado da diferença entre o valor perdido e a sua predição, ou seja, $\sum_{i=1}^n (y_i - \hat{y}_{i-1})^2$. Com o intuito de não fazer esse processo “ n ” vezes, HASTIE e TIBSHIRANI (1986) propõe a seguinte alternativa:

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{y}_i}{1 - \mathbf{H}_\lambda^{ii}} \right\}^2, \quad (2.15)$$

em que $\mathbf{H}_\lambda^{ii}$ são os elementos da matriz de suavização, $\mathbf{H}_\lambda$. CRAVEN e WAHBA (1978) propôs uma apresenta essa fórmula modificada e se refere a ela como validação cruzada Generalizada, em que $\sum_{i=1}^n \mathbf{H}_\lambda^{ii}$ é substituído por $\text{tr}(\mathbf{H}_\lambda)$ e a fórmula é dada por:

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{y}_i}{1 - \frac{\text{tr}(\mathbf{H}_\lambda)}{n}} \right\}^2. \quad (2.16)$$

O melhor valor de λ será aquele que minimiza $CV(\lambda)$ e $GCV(\lambda)$.

2.4.7 Critério de informação de Akaike

O critério de informação de Akaike (AIC) é utilizado quando necessita-se comparar modelos não aninhados ou quando queira comparar três ou mais modelos e pode ser representado pela seguinte expressão:

$$AIC = -2\ell(\hat{\theta}) + 2p,$$

em que p representa o número de parâmetros do modelo, $\ell(\hat{\theta})$ representa o logaritmo da função de verossimilhança avaliado no estimador de máxima verossimilhança de θ ou seja, $\hat{\theta}$. O modelo terá um melhor ajuste quanto menor for o valor de AIC (PAZEL *et al.*, 2006).

2.5 Suavização *spline*

A suavização *spline* é dada através de funções estimadas $f(x_i)$ obtidas através do conjunto de observações y_i com a finalidade de equilibrar uma melhor forma o ajuste. Considere o conjunto $\{x_i, Y_i\}$ das observações. A função $\hat{f}$ que suaviza a função f é definida como:

$$\sum_{i=1}^n \{Y_i - \hat{f}(x_i)\}^2 + \lambda \int \hat{f}''(x)^2 dx, \quad (2.17)$$

em que λ é um parâmetro de suavização com as seguintes propriedades:

- $\lambda > 0$
- Quando $\lambda \rightarrow 0$, o segundo termo da equação (2.17) tende a zerar, portanto tem-se uma convergência para a interpolação *spline*.
- Quando $\lambda \rightarrow \infty$ tem-se uma interpolação exata dos dados, o que foge a idéia de uma curva suave (MENDONZA, 2017).

Todavia, λ é um parâmetro que varia em todo conjunto dos números reais positivos. Para cada λ escolhido, uma curva suavizada diferente é ajustada. Agora, dentre todas as curvas, qual é a que melhor se ajusta aos dados? Uma forma de escolher esse λ é através do Critério de Validação Cruzada Generalizada (ou "generalized cross-validation").

3 MATERIAIS E MÉTODOS

3.1 Banco de dados

Os dados utilizados neste trabalho referem-se ao peso de 55 animais da Raça Hereford, fêmeas, nascidas entre os anos 1999 a 2001, na Agropecuária Recreio, situada no município de Bagé no Rio Grande do Sul. Foram feitas as pesagens de 15 em 15 dias, desde o nascimento até 675 dias de idade (MENDES, 2007).

O gráfico de dispersão desses dados pode ser observado na figura 3.1

Figura 3.1. Gráfico de dispersão dos dados de crescimento de fêmeas da raça Hereford

Note que, o gráfico apresenta o crescimento do peso do animal de forma multifásica, uma vez que, entre os dias 267 e 407 a próximo dos dias 547, a curva apresenta um decaimento, voltando a crescer posterior a esses números, ou seja, o crescimento acontece em ciclos.

Uma forma alternativa de fazer o estudo do crescimento desses animais, é através dos dados médios. Todavia, a qualidade do spline ajustado ou representação do comportamento desses animais pode ter interferências, uma vez que a quantidade de dados analisadas é diferenciada, então como forma de analisar todas as possíveis vantagens e desvantagens, será feito nesse trabalho o estudo utilizando ambas as formas e fazendo as devidas comparações.

Na figura 3.2, tem-se a representação dos dados médios de crescimento.

3.2 Quantidade e posicionamento dos nós

Diante da importância da localização e quantidade nós, muitos autores apresentam formas alternativas de se encontrar a melhor quantidade e melhor posicionamento,

Figura 3.2. Gráfico de dispersão dos dados médios de crescimento de fêmeas da raça Hereford

todavia nenhum deles ainda é aceita pela comunidade científica como melhor maneira. Nesse trabalho, o algoritmo empregado será o da busca cega através da seção dourada.

3.2.1 Método da busca cega através da seção dourada

Dentre todos os métodos, o algoritmo de busca estocástica mais simples é uma pesquisa cega. O método da busca cega através da seção dourada, consiste em escolher “ n ” posições aleatórias no intervalo $[0, 1]$ de maneira a se obter o menor valor do GCV . Segundo SPIRITI *et al.* (2012) a vantagem dessa abordagem é que não se ficará preso ao mínimo local, todavia o método é lento e a função não se adapta a domínios dimensionais mais elevados.

Em aspectos computacionais, o pacote `freeknotspline` possui algumas funções que utilizam esse método para determinar a melhor localidade e quantidade de nós.

A função utilizada nesse trabalho foi a `freepsgold` (mínimos quadrados penalizados), uma vez que ela abrange um parâmetro de penalização λ que representa a suavização da curva ajustada.

3.3 Estudo de simulação

Diversos cenários de simulação foram estudados, com diferentes funções $f(x_i)$. Em todos eles, o comportamento dos dados era de maneira multifásica. Nesse trabalho será apresentado um deles.

Analisou-se, então, por simulação o comportamento dos nós, tanto em quantidade como em localidade. O estudo foi feito fixando uma função e mudando o tamanho

amostral. Assumimos que $Y_i \sim N(f(x_i), \sigma^2 = 1)$, em que

$$f(x_i) = 10 \sin(10\pi x_i) + 3 \cos(2\pi x_i) \quad (3.1)$$

com $x_i \sim U(0, 1)$.

3.4 Aspectos computacionais

O pacote *freeknotspline* do R e as funções *freepsgold* e *free.search.numknot* foram utilizadas para os ajustes desse trabalho, tanto para os ajustes finais quanto para o estudo de simulação. Ambas as funções utilizam o método da seção dourada e da busca cega para representação do ajuste. Ainda é utilizado o método do algoritmo genético mas para demonstração nesse trabalho, será utilizado apenas o método da seção dourada. Para determinação dos melhor posicionamento e quantidade, é necessário pré determinar um critério de comparação, no caso apresentado, foi utilizado o critério de informação de Akaike (AIC).

3.5 Numero máximo de nós com busca cega

Os tamanhos amostrais escolhidos foram: 20, 50 e 100. No primeiro momento, será utilizado o método da busca cega para determinar a melhor quantidade e melhor localidade de nós. Para uso da função `free.search.numknot` é necessário determinar a quantidade mínima e a máxima de nós. A fim de verificar o melhor entre todas as possíveis situações, será utilizada uma quantidade mínima igual a 1 (quantidade mínima que um spline pode aderir) e quantidade máxima igual a “ $n - 2$ ”, em que “ n ” é o tamanho amostral. O primeiro e o último valor da amostra não pode ser um nó, uma vez que a função é constituída por bases splines, ou seja, a função é nula fora do domínio da amostra.

4 RESULTADOS

4.1 Estudo de Simulação

4.1.1 Tamanho amostral 20

Inicialmente, toma-se o tamanho amostral igual a 20, ou seja, coloca-se como valor máximo de nós 18, para o ajuste linear, quadrático e cúbico. Os pontos pontilhados do gráfico estarão representando a localização dos nós.

Figura 4.1. Ajuste com spline linear a dados simulados de tamanho amostral 20 e 15 nós

Figura 4.2. Ajuste com spline quadrático a dados simulados de tamanho amostral 20 e 15 nós

Figura 4.3. Ajuste com spline cúbico a dados simulados de tamanho amostral 20 e 15 nós

A primeiro momento, pode-se reparar que, independente do ajuste em questão, a quantidade de nós determinada pela busca cega é igual a 15 (Figuras 4.1, 4.2 e 4.3). Outra observação importante é que as curvas se assemelham muito, independente do

ajuste. Tanto o ajuste o quadrático quanto o ajuste cúbico, não parecem possuir mudanças suaves nos pontos de quebra. Os nós não apresentam uma lógica quanto a localidade.

4.1.2 Tamanho amostral 50

O tamanho amostral em estudo agora é 50, então o valor máximo de nós poderá ser de no máximo 48.

Figura 4.4. Ajuste com spline linear a dados simulados de tamanho amostral 50 e 46 nós

Figura 4.5. Ajuste com spline quadrático a dados simulados de tamanho amostral 50 e 46 nós

Figura 4.6. Ajuste com spline cubico a dados simulados de tamanho amostral 50 e 46 nós

Tanto o ajuste linear (Figura 4.4), quadrático (Figura 4.5) e cúbico (Figura 4.6) utilizam 46 nós para chegar ao ajuste ideal. Assim como para o tamanho amostral 50, tem-se que todas as curvas se assemelham muito. Os ajustes não parecem ser suaves, isso se da pelo fato da quantidade de nós estar muito próxima ao tamanho amostral.

4.1.3 Tamanho amostral 100

Para um tamanho amostral igual a 100, tem-se que o valor máximo de nós pode ser 98.

Figura 4.7. Ajuste com spline linear a dados simulados de tamanho amostral 100 e 80 nós

Figura 4.8. Ajuste com spline quadrático a dados simulados de tamanho amostral 100 e 80 nós

Figura 4.9. Ajuste com spline cubico a dados simulados de tamanho amostral 100 e 80 nós

Dado os resultados obtidos até aqui, pode-se perceber que: a função nos da como resultado uma quantidade de nós que se aproxima do valor de dados observados o que aproxima muito a curva final com uma interpolação exata, o que foge do ideal de um spline. Tanto a função quadrática quanto a cúbica, se assemelham muito com a linear, uma vez que, pela quantidade grande nós, não se tem muito espaços para variar o comportamento da curva. Como alternativa para esse problema de interpolação, sera proposto na segunda parte do estudo de simulação, uma forma diferente de analisar a quantidade de nós.

4.1.4 Numero fixo de nós com busca cega

O intuito dessa parte da simulação é verificar uma alternativa para diminuir a quantidade de nós, ter um bom ajuste e por fim, encontrar um padrão para a localização dos nós. A quantidade fixada será igual a quantidade de mudanças de sinal que a curva real, dada pela função 3.1, possui. O tamanho amostral vai de encontro com o feito na primeira parte.

4.1.5 Tamanho amostral 20

O primeiro tamanho amostral analisado é igual a 20. Pelo gráfico de dispersão 4.10 nota-se que a função possui 10 mudanças de sinal no domínio $[0, 1]$. Sendo assim, a proposta é ajustar o spline utilizando 10 nós, tanto para o spline linear, quadrático e cúbico.

Figura 4.10. Gráfico de dispersão para tamanho amostral 20

Figura 4.11. Ajuste com spline linear para tamanho amostral 20 e 10 nós

Figura 4.12. Ajuste com spline quadrático para tamanho amostral 20 e 10 nós

Figura 4.13. Ajuste com spline cúbico para tamanho amostral 20 e 10 nós

Ainda que, na figura 4.12 e 4.13 não demonstrem uma suavidade nos pontos de quebra, tem-se que esses ajustes não estão interpolando. O ajuste linear, nos primeiros 15 pontos da amostra, estão muito próximos da curva, todavia, vale ressaltar que o tamanho amostral é muito pequeno. Em nenhum dos casos consegue-se perceber uma lógica quanto a localização dos nós.

4.1.6 Tamanho amostral 50

Com o tamanho amostral igual 50, pode-se notar pela figura 4.14 que a curva possui 10 mudanças de sinal, sendo assim, será utilizado 10 nós no domínio da função.

Figura 4.14. Gráfico de dispersão de tamanho amostral 50

Figura 4.15. Ajuste com spline linear para tamanho amostral 50 e 10 nós

Figura 4.16. Ajuste com spline quadrático para tamanho amostral 50 e 10 nós

Figura 4.17. Ajuste com spline cúbico para tamanho amostral 50 e 10 nós

Trabalhando com um tamanho amostral maior, conseguimos notar algumas coisas importantes, como por exemplo o fato do ajuste linear possuir os nós nos pontos de mudança de sinal. Já para o ajuste quadrático, temos que seus respectivos nós estão nos pontos de inflexão da curva. Os nós do ajuste cúbico não possuem um lugar fixo, mas

conseguimos notar que varia entre estar nos pontos de inflexão e de mudança. Os ajustes, mesmo que com uma quantidade de nós inferior ao da primeira parte, se aproxima muito dos valores observados (note figura 4.15, 4.16 e 4.17). Tanto o ajuste quadrático quanto o cúbico já demonstram uma suavidade maior em todo o seu domínio.

4.1.7 Tamanho amostral 100

Com o tamanho amostral 100, poderá ser verificado a tendência da localidade dos nós. Observe inicialmente as 10 mudanças de sinal presentes na figura 4.18

Figura 4.18. Gráfico de dispersão de tamanho amostral 100

Figura 4.19. Ajuste com spline linear para tamanho amostral 100 e 10 nós

Figura 4.20. Ajuste quadrático com tamanho amostral 100 e 10 nós

Figura 4.21. Ajuste com spline cúbico para tamanho amostral 100 e 10 nós

Com uma número maior de dados, pode-se notar que o comportamento da curva ainda se mantém próximo aos dados. Os nós mantêm as propriedades já mencionadas. Tanto o ajuste cúbico quanto o quadrático se mostram extremamente flexíveis e suaves, já o linear, apresenta nos pontos de quebra, uma mudança brusca mas ainda sim, consegue ser fiél ao comportamento dos dados.

4.1.8 Comentários finais do estudo de simulação

Diante da simulação apresentada, conseguimos notar uma lógica quanto ao posicionamento dos nós. Quando o ajuste em questão for o linear, eles estão alocados nos pontos em que há uma mudança de sinal. Uma utilização menor do número de nós não é suficiente para que as funções *spline* linear se tornem suaves, todavia, podem ser importantes para que elas não interpoem os dados, e sim se aproximem bastante dos dados observados, gerando um ajuste de grande qualidade.

No caso do ajuste quadrático, temos uma função um pouco mais maleável quando comparado com a linear, sendo assim, ele se adapta melhor aos dados e possui como localidade dos nós os pontos de inflexão. O ajuste se mostrou satisfatório, uma vez que a curva se aproxima dos valores observados, mas não interpola, o que garante mais uma vantagem do método proposto.

O ajuste cúbico não possui uma localidade fixa, ele varia entre os pontos de inflexão e ponto de mudança de sinal, todavia, a função cúbica é flexível, então, com uma quantidade de nós menor, ela ainda assim consegue se ajustar bem aos dados.

Vale ressaltar que, quando o número amostral é muito pequeno, essas conclusões podem não ser dadas como verdadeiras, uma vez que os nós não apresentam uma real localização.

Dado o resultados das simulações, podemos ajustar modelos *splines* aos dados do trabalho, partindo do pressuposto que o número total de nós sera igual ao total de mudanças de sinal que a curva apresentar.

Com base no estudos feitos até então associado com o estudo de simulação, podemos utilizar uma maneira para ajuste de dados de crescimento multifásica, fixando uma quantidade de nós. Nessa parte do trabalho, serão aplicadas as técnicas vistas até então para o conjunto de dados apresentados. Para todos os ajustes será utilizado os dados em sua forma completa e com os dados médios, ou seja, os dados completos são aqueles que possuem mais de uma medida durante o dia (vários animais medidos no mesmo dia), e a outra versão se dá com a média de peso diária.

4.2 Estudo de dados completos

4.2.1 Ajuste Linear

Será apresentado o ajuste final para dados completos com spline linear. Como na simulação, a quantidade de nós será igual à quantidade de mudanças de sinal da curva (4 neste caso), e a dos nós localizados nos pontos em que há essa mudança, veja Figura 4.22.

Figura 4.22. Ajuste Linear a dados completos utilizando 4 nós

Os valores dos pesos preditos para cada nó específico são apresentados na Tabela 4.1.

Tabela 4.1. Localização dos nós para spline linear a dados completos com valores de pesos preditos

Localização dos nós	Valores de pesos preditos
194.6063	151.4354
372.3218	140.1436
483.1723	233.4217
555.8052	235. 2723

4.2.2 Ajuste Quadrático

A utilização de spline quadrático para ajuste, sera provinda da regra de: nós no pontos de inflexão de função e quantidade ótima obtida usando o pacote `free.search.numknot`. A idéia é fornecr uma curva mais suave quando comparada com o spline linear em decorrência da flexibilidade de sua função, veja Figura 4.23.

Figura 4.23. Ajuste quadrático a dados completos utilizando 4 nós

Os valores dos pesos preditos para um nó específico é apresentado na Tabela 4.2.

Tabela 4.2. Localização dos nós para spline quadrático a dados completos com valores de pesos preditos

Localização dos nós	Valores de pesos preditos
141.5629	118.5828
284.5851	142.2086
452.6665	206.9978
509.3134	237.1348

4.2.3 Ajuste Cúbico

Os ajustes cúbicos, de acordo com as simulações, não apresentam localidade específica, todavia, sabemos que a localização se dá nos pontos de mudança de sinal e/ou nos pontos de inflexão da função. O método da seção dourada define a melhor localização para os 4 nós, uma vez que a quantidade foi determinada inicialmente, veja Figura 4.24.

Figura 4.24. Ajuste cúbico a dados completos utilizando 4 nós

Os valores dos pesos preditos para um nó específico é apresentado na Tabela 4.3.

Tabela 4.3. Localização dos nós para spline cúbico a dados completos com valores de pesos preditos

Localização dos nós	Valores de pesos preditos
212.0084	149.6364
401.3698	163.7232
496.7815	238.2437
535.0283	230.6452

Com base na Tabela 4.4, o modelo com o menor valor de AIC é o spline cúbico, portanto para um ponto de vista teórico, esse modelos seria o mais adequado para o representar os dados.

Tabela 4.4. Valores de AIC e λ para ajustes utilizando dados completos

Tipo de Ajuste	Valores de AIC	Valor de λ
<i>spline</i> Linear	14613.52	149.7653
<i>spline</i> Quadrático	14611.67	479212.5
<i>spline</i> Cúbico	14611.61	1982273879

4.3 Estudo de dados médios

4.3.1 Ajuste Linear

Uma forma alternativa de analisar os dados é através de seu comportamento médio. Diferente da primeira parte que observamos os dados como foram observados, nesse momento será utilizado a média das observações diariamente. Inicialmente para o caso linear, temos a representação do ajuste na figura 4.25.

Figura 4.25. Ajuste Linear a dados medios utilizando 4 nós

Os valores dos pesos preditos para um nó específico é apresentado na Tabela 4.5.

Tabela 4.5. Localização dos nós para *spline* linear a dados médios com valores preditos

Localização dos nós	Valores preditos
194.7024	151.5159
372.1739	140.2598
483.1779	206.7480
556.7170	230.9006

4.3.2 Ajuste Quadrático

A utilização de *spline* quadrático para ajuste, sera provinda da regra de: nós no pontos de inflexão de função e quantidade ótima obtida usando o pacote `free.search.numknot`, veja Figura 4.26.

Figura 4.26. Ajuste quadrático a dados médios utilizando 4 nós

Os valores dos pesos preditos para um nó específico é apresentado na Tabela 4.6.

Tabela 4.6. Localização dos nós para ajuste quadrático a dados médios com valores preditos

Localização dos nós	Valores preditos
141.2002	118.6181
284.9342	142.1571
452.4933	207.1670
508.2249	2367968

4.3.3 Ajuste Cúbico

Assim como discutido nos ajuste cúbico de dados completos, o ajuste cúbico não possui localização lógica, o que pode se notar através das simulações é que a localização dos nós varia entre os pontos de mudança de sinal da curva e os pontos de inflexão, veja Figura 4.27.

Figura 4.27. Ajuste cúbico a dados médios utilizando 4 nós

Os valores dos pesos preditos para um nó específico é apresentado na Tabela 4.7.

Tabela 4.7. Localização dos nós para ajuste cúbico a dados médios com valores preditos

Localização dos nós	Valores preditos
211.5944	149.6226
402.4336	163.6061
495.2043	238.2279
534.7929	230.7197

Com base na Tabela 4.8, o melhor modelo para dados médios é o *spline* cúbico.

Tabela 4.8. Valores de AIC e λ para ajustes utilizando dados médios

Tipo de Ajuste	Valores de AIC	Valor de λ
<i>spline</i> Linear	80.5730	1.7990
<i>spline</i> Quadrático	73.1807	5056.592
<i>spline</i> Cúbico	-69.4414	17336323

4.4 Discussão

Quando se está trabalhando com *spline* linear, nota-se uma diferença muito grande quanto a questão do AIC, uma vez que os dados médios possuem um valor muito menor quando comparado aos ajustes de dados completos. Esse fato se dá pelo efeito de quase interpolação da função *spline*, pois trabalhando com dados médios quase não se tem variância dos dados (o mesmo podemos dizer para os ajustes quadrático e cúbico).

Outra coisa importante a se frisar, é que: quando comparado o ajuste utilizando uma quantidade de nós igual a mudança de sinal e a quantidade dada pela `free.search.numknot`, tem-se que o valor de AIC é maior para o primeiro caso, todavia, o ajuste se assemelha muito. Isso se dá pelo fato que uma quantidade maior de nós, pode aproximar ainda mais o ajuste, mas a quantidade menor de nós pode exigir um valor maior de λ , fazendo com que a função se assemelhe em ambos os casos.

Quando se trabalha com dados completos, a qualidade de ajuste medida por AIC é quase igual, isso porque ambas as funções tendem a manter o ajuste de forma a não interpolação e sim, de comportamento dos dados. Como a função quadrática e cúbica são suaves, tem-se que a qualidade é muito próxima.

5 CONCLUSÃO

A seguir algumas conclusões do trabalho:

- Tantos para dados completos quanto para dados médios, as funções splines são excelentes alternativas para o ajuste, seguindo muito bem a tendência dos dados, quando escolhido o local e quantidade correta de dados;
- Se a quantidade de nós e localidade forem bem escolhidas, pode-se entender que o ajuste quadrático se ajuste tão bem quanto o cúbico, sendo assim, por depender de menos parâmetros e conseguir determinar de maneira mais simples a localidade dos nós (nos pontos de inflexão) pode-se usar o spline quadrático para ajustes de dados. Todavia, para um sentido prático, vale ressaltar que: é necessário compreender a real necessidade do pesquisador. No caso dos dados apresentados nesse trabalho, os nós do ajuste linear, conseguem apresentar os pontos em que acontece uma mudança no peso do animal, já os posicionamento dos nós do ajuste quadrático, evidencia o momento em que o animal começa a mudar o peso.
- Utilizar uma quantidade muito grande de nós, pode tender o ajuste a uma interpolação, o que não é o fator essencial do *spline* (flexível), sendo assim, deve-se atentar a quantidade que será usada. Se for o *spline* quadrático ou linear, uma alternativa é verificar a quantidade de mudança de sinal de função;
- Computacionalmente, os *splines* são excelentes facilitadores de ajuste, uma vez que de maneira simples, o pesquisador consegue ajustar dados que se comportem de diferentes formas.

5.1 Trabalhos Futuros

- Comparar o algoritmo utilizado com outros algoritmos do pacote **freeknot** *spline*;
- Estudar critérios para avaliar a capacidade preditiva do modelo e poder comparar com modelos multifásicos.

REFERÊNCIAS

- A.B.H.B., Associação Brasileira de Hereford e Braford, 2017 Hereford. <<http://www.abhb.com.br/hereford/hereford/>>. Acessado em 10 de julho de 2015.
- BATES, D. M. e D. G. WATTS, 1988 *Nonlinear Regression Analysis and its applications*. John Wiley & Sons, Inc., first edition.
- BERGAMASCO, A., L. AQUINO, J. MUNIZ, e F. SILVA, 2001 Ajuste de modelos não lineares a dados de crescimento de fêmeas da raça holandesa. *Ciência e Tecnologia, Lavras* **25**: 235–241.
- BURCHARD, H., 1974 Splines (with optimal knots) are better. *Journal of Applicable Analysis: An International Journal* **3**: 309–319.
- CECON, P., A. SILVA, M. NASCIMENTO, e A. FERREIRA, 2012 *Métodos Estatísticos*. Editora UFV.
- CRAVEN, P. e G. WAHBA, 1978 Smoothing noisy data with spline functions. *Numerische Mathematik* pp. 377–403.
- DE BOOR, C., 1978 *A Pratical Guide to Spline*. First edition.
- EILERS, P. e B. MARX, 1996 Flexible smoothing with b-spline and penalties. *Statistical Science* **11**: 89–102.
- FITZHUGH, H., 1976 Analysis of growth curves and strategies for altering their shape. *Journal of animal Science* **42**: 1036–1051.
- GOLUB, H., G. M. HEATH, e G. WAHBA, 2012 Generalized cross-validation as a method for choosing a good ridge parameter. *Journal Technometrics* pp. 215–223.
- HARRELL, F., 2001 *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag.
- HASTIE, T. e J. TIBSHIRANI, R, 1986 Generalized additive models. *Statistical Science* **1**: 297–310.
- HUISMAN, A., R. VEERKAMP, e J. ARENDONK, 2002 Genetics parameters for various random regression models to describe the weight data of pigs. *Journal of Animal Science* **80**: 575–582.
- LINDSTROM, M. J., 1999 Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics* **2**: 333–352.

- LOBO, R., L. VILLELA, J. PASSO, e A. OLIVEIRA, 2006 Parâmetros genéticos de características estimadas da curva de crescimento de ovinos da raça santa inês. *Revista Brasileira de Zootecnia* pp. 1012–1019.
- MAZUCHELI, J. e J. A. ACHCAR, 2008 Algumas considerações em regressão não linear. *Acta Scientiarum. Technology* **24**: 1761–1770.
- MAZZINI, A. R. D. A., 2001 *Análise de curva de crescimento de machos Hereford considerando heterogeneidade de variâncias e autocorrelações dos erros*. Dissertação de mestrado, Universidade Federal de Lavras, Lavras.
- MENDES, P., J. MUNIZ, F. SILVA, e A. MAZZINI, 2008 Modelo logístico difásico no estudo de crescimento de fêmeas da raça hereford. *Ciência Rural JCR* .
- MENDES, P. N., 2007 Curvas de crescimento difásicas de fêmeas hereford com erros auto-regressivos e heterogeneidade de variâncias.
- MENDONZA, N. V. R., 2017 *Estruturas unidimensionais e bidimensionais utilizando P-splines nos modelos mistos generalizados com aplicação na produção de cana de açúcar..* Tese de doutorado (estatística e experimentação agrônômica).
- OLIVEIRA, D., 2011 *Funções spline para estudo de curvas de crescimento em ovinos cruzados*. Dissertação (mestrado em estatística aplicada a biometria).
- OLIVEIRA, H., R. LÔBO, e C. PEREIRA, 2000 Comparação de modelos não-lineares para descrever o crescimento de fêmeas da raça guzerá. *Pesquisa Agropecuária Brasileira* **35**: 1843–1851.
- PAZEL, E. F., I. MULLER, C. FINGER, e P. SCHNEIDER, 2006 Ajuste e seleção de modelos tradicionais para série temporal de dados de altura de árvores. *Ciência Florestal*, Santa Maria pp. 177–199.
- RAO, C. R., 1958 Some statistical methods for comparison of growth curves. *Biometrics* **14**: 17.
- RATKOWSKI, D., 1993 Principles os nonlinear regression modeling. *Journal of Industrial Microbiology* **12**: 195–199.
- ROBBINS, K., I. MISZTAL, e J. BERTRAND, 2005 A pratical longitudinal model for evaluting growth in gelbvieh cattle. *J.Anim. SCI* **83**: 29–33.
- RUPPERT, D., M. VAND, e R. CARROLI, 2003a *Semiparametric regression*. Cambridge University Press.

- RUPPERT, D., M. WAND, e R. J. CARROLL, 2003b *Semiparametric Regression*. Universidade Estadual de Campinas - Departamento de Estatística., Universidade Estadual de Campinas - Departamento de Estatística.
- SMYTH, G. K., 2006 *Nonlinear Regression*, volume 3. John Wiley & Sons, Ltd, New York.
- SPIRITI, S., R. EUBANK, P. W. SMITH, e D. YOUNG, 2012 Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation* pp. 1020 –1036.
- SPIRITI, S., P. SMITH, e P. LECUYER, 2018 *freeknotsplines: Algorithms for Implementing Free-Knot Splines*. R package version 1.0.1.
- THOMSON, R. e W. EMERY, 2014 *Data Analysis Methods In Physical Oceanography*. Elsevier.
- VIEIRA, K. e M. MISCHAN, 1976 A Logística e a Gompertz: duas funções alternativas no estudo de dados de crescimento. *Ciência e Cultura* pp. 952–959.

APÊNDICES

Apêndice A

Através de indução finita, pode-se provar que a função geral do spline é contínua para um número finito de nós da seguinte maneira. Considere um conjunto de nós $\{b_1, b_2, \dots, b_K\}$. Para $K = 1$ tem-se:

$$h(x) = \begin{cases} \sum_{j=0}^D \beta_{0j} x^j, & x < b_1 \\ \sum_{j=0}^D \beta_{0j} x^j + \beta_1 (x - b_1)^D & x \geq b_1. \end{cases}$$

Analisando os limites laterais tem-se que:

$$\lim_{x \rightarrow b_1^-} h(x) = \lim_{x \rightarrow b_1^-} \sum_{j=0}^D \beta_{0j} x^j = \sum_{j=0}^D \beta_{0j} b_1^j = h(b_1)$$

e

$$\lim_{x \rightarrow b_1^+} h(x) = \lim_{x \rightarrow b_1^+} \sum_{j=0}^D \beta_{0j} x^j + \beta_1 (x - b_1)^D = \sum_{j=0}^D \beta_{0j} b_1^j = h(b_1),$$

ou seja, $h(x)$ é contínua em b_1 . Além disso, fora de b_1 , $h(x)$ é soma de polinômio, e portanto, contínua. Suponha-se então, que $h(x)$ é contínua em $K - 1$, logo:

$$\lim_{x \rightarrow b_K^-} h(x) = \lim_{x \rightarrow b_K^-} \sum_{j=0}^D \beta_{0j} x^j + \sum_{k=1}^{K-1} \beta_k (x - b_k)^D = \sum_{j=0}^D \beta_{0j} b_K^j + \sum_{k=1}^{K-1} \beta_k (b_K - b_k)^D = h(b_K)$$

e

$$\lim_{x \rightarrow b_K^+} h(x) = \lim_{x \rightarrow b_K^+} \sum_{j=0}^D \beta_{0j} x^j + \sum_{k=1}^K \beta_k (x - b_k)^D = \sum_{j=0}^D \beta_{0j} b_K^j + \sum_{k=1}^{K-1} \beta_k (b_K - b_k)^D = h(b_K).$$

Portanto, $h(x)$ é contínua em b_K . Pela hipótese de indução, $h(x)$ era contínua nos outros nós, e fora desses nós é uma soma de polinômio, ou seja, $h(x)$ é contínua.

Apêndice B

A seguir apresentamos o código R utilizado no Capítulo 4.

```
#-----Ajuste Dados Completos-----
rm(list = ls())
library(freeknotsplines)
dados<- read.csv2('dadoscor.csv', head=T, dec=",")
dadosNA<- na.omit(dados)

#Otimização para spline linear
fit1<-fit.search.numknots(x=dadosNA$TEMPO, y=dadosNA$PESO,
degree =1,minknot = 1, maxknot = 44, alg = "PS",
search = "golden", knotnumcrit = "AIC")

#Otimização para spline Quadrático
fit2<-fit.search.numknots(x=dadosNA$TEMPO, y=dadosNA$PESO,
degree =2,minknot = 1, maxknot = 44, alg = "PS",
search = "golden", knotnumcrit = "AIC")

#Otimização para spline Cúbico
fit3<-fit.search.numknots(x=dadosNA$TEMPO, y=dadosNA$PESO,
degree =3,minknot = 1, maxknot = 44, alg = "PS",
search = "golden", knotnumcrit = "AIC")

#Verificação da quantidade e posicionamento dos nos ótimos
fit1@optknot
fit2@optknot
fit3@optknot

#Cálculo AIC
AIC(fit1)
AIC(fit2)
AIC(fit3)

#Ajuste a dados médios

rm(list = ls())
dados<- read.csv2('dadoscor.csv', head=T, dec=",")
dadosNA<- na.omit(dados)
```

```
library(freeknotsplines)
library(plyr)
DADOS.MEDIOS<-ddply(dadosNA, "TEMPO", summarise, PESO = mean (PESO))
dadosNA <-DADOS.MEDIOS
plot(DADOS.MEDIOS)

modelo1 <- freepsgold(x=dadosNA$TEMPO,y= dadosNA$PESO, degree = 1,
numknot = 4, seed = 555)

modelo2 <- freepsgold(x=dadosNA$TEMPO,y= dadosNA$PESO, degree = 2,
numknot = 4, seed = 555)

modelo3 <- freepsgold(x=dadosNA$TEMPO,y= dadosNA$PESO, degree = 3,
numknot = 4, seed = 555)
```