UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS
CURSO DE ESTATÍSTICA

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA PREDIÇÃO EM MODELOS ADITIVOS

Marco Aurelio Valles Leal

Maringá
2022

MARCO AURELIO VALLES LEAL

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA PREDIÇÃO EM MODELOS ADITIVOS

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de bacharel em Estatística pela Universidade Estadual de Maringá.

Orientador: Profº Drº George Lucas Moraes Pezzot
Coorientador: Profº Drº Willian Luís de Oliveira

Maringá
2022

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA PREDIÇÃO EM MODELOS ADITIVOS

MARCO AURELIO VALLES LEAL

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de bacharel em Estatística pela Universidade Estadual de Maringá.

Aprovado em: _____/_____/_____.

BANCA EXAMINADORA

_____

Orientador

Profº Drº George Lucas Moraes Pezzot

Universidade Estadual de Maringá

_____

Membro da banca

Nome do professor membro da banca

Instituição do professor membro da banca

_____

Membro da banca

Nome do professor membro da banca

Instituição do professor membro da banca

RESUMO

É comum, nas mais diversas áreas, investigar e modelar a relação entre variáveis. O modelo mais simples é denominado modelo de regressão linear simples e assume que a média da variável resposta é modelada como uma função linear das variáveis explicativas, supondo erros aleatórios com média zero, variância constante e não correlacionados. Entretanto, nem sempre a relação existente é perfeitamente linear. Neste contexto, é possível flexibilizar o modelo de regressão linear modelando a dependência da variável resposta com cada uma das variáveis explicativas em um contexto não paramétrico. Esta nova classe de modelos é dita modelos aditivos e mantêm a característica dos modelos de regressão lineares de serem aditivos nos efeitos preditivos. Portanto, este projeto visa apresentar os modelos aditivos, além de técnicas de suavização utilizadas para ajustar modelos no contexto não paramétrico. Por fim, a metodologia é aplicada em dados artificiais (simulados) e em dados reais, dando enfoque à qualidade das predições.

Palavras-chave : Regressão. Modelo aditivo. Suavizadores.

Sumário

1 Introdução **2**
1.1 Objetivo Geral 2
1.2 Objetivos Específicos 3

2 Referencial Teórico **3**

3 Metodologia **4**
3.1 Regressão Linear 4
3.1.1 Regressão Linear Múltipla 5
3.2 Modelo Aditivo 6
3.3 Suavizadores 7
3.3.1 Técnicas de suavização 8
3.3.2 Loess 9
3.3.3 Kernels 9
3.3.4 Splines 10
3.3.5 Splines de regressão 10
3.4 Seleção de Modelos - Enfoque de Predição 11

4 Resultados e Discussão **13**
4.1 Estudo de simulação 13
4.1.1 Cenário 1 13
4.1.2 Cenário 2 18

5 Referências **20**

1 Introdução

A análise de regressão é uma técnica amplamente utilizada na estatística que visa explorar e modelar a relação entre variáveis (MONTGOMERY ET AL.,2012). Usualmente, é de interesse apenas uma variável, denominada resposta ou dependente, cuja análise deve considerar a dependência de um conjunto de variáveis observáveis, chamadas de variáveis explicativas, independentes ou preditoras. Neste cenário, os modelos de regressão (linear simples ou múltipla) podem ser utilizados. A dependência da variável resposta é constituída pelo somatório dos termos que representam as variáveis preditoras e seus respectivos parâmetros, que devem ser estimados por mínimos quadrados ou máxima verossimilhança. Ainda, considera-se um componente aleatório do erro com média zero e variância constante.

Nota-se, porém, que em muitos casos a relação existente entre a variável resposta (média) e cada uma das variáveis explicativas não é perfeitamente linear. Um solução seria acrescentar uma transformação nas variáveis regressoras adotando, por exemplo, o método de transformação de Box-Cox. Determinar uma transformação que represente a correta relação existente nem sempre é fácil. Outra possibilidade é flexibilizar o modelo de regressão linear, modelando a dependência da variável resposta com cada uma das variáveis explicativas em um contexto não paramétrico (EUBANK, 1999). Esta nova classe de modelos é dita modelos aditivos e mantêm a característica dos modelos de regressão linear de serem aditivos nos efeitos preditivos (HASTIE E TIBSHIRANI,1990).

Este modelo é composto pela soma de funções suavizadas das variáveis preditoras, o que possibilita examinar o efeito de cada uma na variável resposta. Neste contexto, as funções devem ser estimadas por meio de suavizadores, que estima uma tendênica menos variável e descreve sua dependência em relação à resposta (Y). Para análise visual, tais suavizadores podem ser representados em diagramas de dispersão. Alguns dos métodos mais adotados para obter as estimativas suavizadas são: Bin smoother, running mean, running line, Loess ou Lowess, suavizador de Kernels e Splines.

Portanto, o objetivo do presente trabalho é apresentar os modelos aditivos (considerando apenas uma variável preditora ou explicativa). Ademais, visa-se estudar as principais técnicas de suavização existentes utilizadas para ajustar modelos no contexto não paramétrico, em particular os modelos aditivos, apresentando suas principais características e aplicações. Além disso, pretende-se mostrar o ganho na predição, ao se empregar modelos mais flexíveis.

1.1 Objetivo Geral

O objetivo deste trabalho é apresentar os modelos aditivos, uma generalização dos modelos de regressão linear, descrevendo suas principais características e estudar algumas técnicas de estimação do modelo, no contexto não paramétrico.

1.2 Objetivos Específicos

- Introduzir os modelos aditivos, especificando sua principais características e apresentar os principais métodos de estimação dos parâmetros do modelo assim como as técnicas de diagnóstico;
- Apresentar técnicas de suavização utilizadas para estimar funções não paramétricas presentes nos modelos aditivos, identificando suas principais características;
- Introduzir métricas para verificar a qualidade de predição;
- Realizar um estudo de simulação para averiguar a qualidade do ajuste dos modelos em alguns cenários, considerando os modelos aditivos;
- Aplicar a metodologia estudada a um conjunto de dados reais, comparando modelos e técnicas de estimação e predição.

2 Referencial Teórico

Existem distintas abordagens obter estimativas das funçoes em métodos de regressão não-paramétricos. Estas estimativas dependem dos próprios dados e de suas observações vizinhas em torno de um dado ponto. Um dos primeiros e mais utilizados métodos de regressão não-paramétrica foi apresentado por Nadaraya-Watson (1964), denominados estimadores tipo núcleo (Kernel), os quais foram aperfeiçoados com os métodos de regressão polinomial local, conhecidos como loess (CLEVELAND, 1979).

A estimação de f(x) consiste em ajustes locais, realizando vários ajustes paramétricos por meio de regressão polinomial com pesos (Lowess),considerando os dados mais próximos do ponto onde deve ser feita a estimação da função (DELICADO, 2008). Ainda deve-se escolher de forma apropriada os parâmetros da largura da banda (parâmetro de suavização) e os graus de ajuste polinomiais para obter o melhor ajuste da regressão.

Além disso, tem-se o métodos splines (vide, por exemplo, Reinsch 1967 e Eubank 1999) corresponde em encontrar um estimador para f(X) que minimiza a soma de quadrados dos resíduos, adicionando um termo que penaliza a falta de suavidade das funções estimadas, uma solução é utilizar a suavização spline cúbico (GREEN E SILVERMAN, 1994).Estes são generalizações de polinômios cúbicos adotados na regressão paramétrica. Dentre os splines mais conhecidos são os splines cúbicos, os splines cúbicos naturais e os B-splines (vide, por exemplo, Hastie e Tibshirani 1990; Green e Silverman 1994 e Wood 2017).

Green e Yandell (1985), em seus estudos sobre modelos lineares parciais, discorrem sobre modelos lineares generalizados semiparamétricos, valendo-se de splines cúbicos, e apresentam as estimativas de máxima verossimilhança penalizada e métodos para estimar o parâmetro

de suavização. Green e Silverman (1994) expõem em seu trabalho detalhes sobre a regressão não paramétrica, bem como, splines e modelos lineares generalizados. E por fim, Fahrmeir e Tutz (2001) descrevem a teoria de modelos semiparamétricos e não paramétricos e apresentam diversas aplicações.

,

3 Metodologia

3.1 Regressão Linear

Análise de regressão é uma técnica estatística utilizada para investigar e modelar a relação entre variáveis. Aplicações de regressão são numerosas e ocorrem em quase todas as áreas do conhecimento, como engenharia, ciências físicas e químicas, economia, ciência biológicas, etc. Resumidamente a regressão tem como objetivo descrever uma relação entre uma variável de interesse, chamada de variável resposta ou dependente (Y) e um conjunto de variáveis preditoras ou independentes (X), as co-variáveis. Através do modelo é possível estimar parâmetros e fazer inferências sobre os mesmos, como testes de hipóteses e intervalos de confiança.

Além disso, o modelo de regressão pode ser usado para predição, onde se é esperado que grande parte da variação de Y seja explicado pelas variáveis X. Dessa forma, obtem-se valores esperados de Y correspondentes a valores de X que não estavam entre os dados.

O modelo de regressão linear simples, constitui uma tentativa de estabelecer uma equação matemática linear que descreve o relacionamento entre duas variáveis, X (preditora) e Y (resposta). O modelo de regressão linear populacional é definido por:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

onde o intercepto β_0 e a inclinação da reta β_1 são parâmetros desconhecidos e ε é um erro aleatório. Pressupõe-se que os erros têm média zero, variância σ^2 desconhecida e são não correlacionados, assim, as respostas também não têm relação. A média da distribuição é dada por uma função linear de x:

$$E(Y|x) = \beta_0 + \beta_1 x \tag{2}$$

Os parâmetros β_0 e β_1, também chamados de coeficientes da regressão, tÊm uma interpretação simples e muitas vezes útil. A inclinação β_1 é a alteração média da distribuição de Y produzida por uma mudança unitária da variável X, ou seja, o quanto varia a média de Y para o aumento de uma unidade de X. Se os dados de X incluem $x = 0$, então o intercepto β_0 é a

média da distribuição da resposta Y quando $x = 0$. Porém, se a observação no zero não estiver incluída, β_0 não tem interpretação prática e é chamado de intercepto ou coeficiente linear, pois é o ponto onde a reta regressora corta o eixo y.

3.1.1 Regressão Linear Múltipla

Geralmente, faz-se necessário adotar um model que considere as demais variáveis regressoras, como é o caso do modelo de regressão linear múltipla. A resposta y pode estar relacionada a k regressores ou variáveis preditoras. A relação do do modelo é dada por,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Os parâmetros $\beta_j, j = 0, 1, ..., k$ são os coeficientes de regressão. O parâmetro β_j representa a mudança esperada na resposta y a cada mudança unitária de x_j quando quando todas as outras variáveis regressoras x_i, com $(i \neq j)$, são constantes. Por isso, os parâmetros β_j são, frequentemente, chamados de coeficientes parciais de regressão.

Os modelos de regressão linear múltipla são, usualmente, tidos como modelos empíricos ou funções aproximadas. Isto é, a verdadeira função que descreve o relacionamento entre y e $x_1, x_2, ..., x_k$ é desconhecida, mas em certos intervalos das variáveis regressoras, o modelo de regressão linear é uma aproximação adequada para a verdadeira função desconhecida.

O Método do Mínimos Quadrados pode ser utilizado para estimar os coeficientes de regressão da equação

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Suponha que $n > k$ observações estão disponíveis, seja y_i a i-ésima obervação ou nível do regressor x_j. Assumindo que o erro ε do modelo tem $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ e são não correlacionados, temos o modelo de regressão amostral:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, ..., n \end{aligned}$$

A função de mínimos quadrados é

$$\begin{aligned} S(\beta_0, \beta_1, ..., \beta_k) &= \sum_{i=1}^{n} (y_i - f_i(\beta_0, \beta_1, ..., \beta_k))^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 \end{aligned}$$

A função S deve ser minimizada em relação à $\beta_0, \beta_1, ..., \beta_k$. Logo, os estimadores de mínimos quadrados desses parâmetros devem satisfazer

$$\begin{aligned} \left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_k}} &= -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta_0} - \sum_{j=1}^{k} \hat{\beta_j} x_{ij} \right) = 0 \\ \left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_k}} &= -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta_0} - \sum_{j=1}^{k} \hat{\beta_j} x_{ij} \right) x_{ij} = 0, \quad j = 1, 2, ..., k \end{aligned}$$

Simplificando, obtém-se as equações normais de mínimos quadrados:

$$\begin{aligned} n\hat{\beta_0} + \hat{\beta_1} \sum_{i=1}^{n} x_{i1} + \hat{\beta_2} \sum_{i=1}^{n} x_{i2} + ... + \hat{\beta_k} \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\ \hat{\beta_0} \sum_{i=1}^{n} x_{i1} + \hat{\beta_1} \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta_2} \sum_{i=1}^{n} x_{i1} x_{i2} + ... + \hat{\beta_k} \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\ &\vdots \\ \hat{\beta_0} \sum_{i=1}^{n} x_{ik} + \hat{\beta_1} \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta_2} \sum_{i=1}^{n} x_{ik} x_{i2} + ... + \hat{\beta_k} \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i \end{aligned}$$

Nota-se que há $p = k + 1$ equações normais, uma para cada coeficiente de regressão desconhecido. A solução dessas equações resultará nos estimadores de mínimos quadrados $\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_k}$.

3.2 Modelo Aditivo

Uma das mais populares e úteis ferramentas em análise de dados é o modelo de regressão linear. Se a dependência de de Y em X é linear ou quase linear, então o modelo de regressão linear é útil. Caso esta dependência não seja linear, não iremos querer resumi-la em uma linha reta. Poderíamos adicionar um termo quadrático, mas geralmente é dificultoso encontrar a forma mais apropriada. Nesse contexto, tem-se os modelos aditivos, que podem ser vistos como uma flexibilização do modelo de regressão linear, considerando a dependência da variável resposta com cada uma das variáveis explicativas em um contexto não paramétrico.

Para isto, considera-se que cada uma das variáveis explicativas está relacionada à média da variável resposta Y através de uma função univariada desconhecida (função suave) não especificada de uma forma paramétrica, ou seja, o componente sistemático é formado por uma soma de funções suaves não especificadas das variáveis explicativas. Esta nova classe de modelos é dita modelos aditivos e mantêm a característica dos modelos de regressão lineares de serem aditivos nos efeitos preditivos. Os modelos aditivos são um caso particular de uma classe mais geral denominada modelos aditivos generalizados, definido por $y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon$ (HASTIE & TIBSHIRANI, 1990). Neste trabalho sera abordado um modelo com apenas uma covariável neste caso,

$$y = \alpha + f(X) + \epsilon,$$

onde os erros ϵ são independentes, com $E(\epsilon) = 0$ e $var(\epsilon) = \sigma^2$. A $f(X)$ é uma função univariada arbitrária.

Os modelos aditivos mantêm muitas das boas propriedades dos modelos lineares, porém são mais flexíveis. Como visto, Uma das vantagens de modelos lineares é sua simplicidade na interpretação: caso o interesse seja em saber como a previsão muda conforme mudanças em x_j, só é necessário saber o valor de β_j, embora a função de resposta parcial f_j desempenha esse mesmo papel em um modelo aditivo.

3.3 Suavizadores

Essas funções do componente sistemático podem ser estimadas através de um suavizador (*smoother*), uma ferramenta que representa a tendência da variável resposta como função das covariáveis disponíveis. No caso em que apenas uma covariável está disponível para predizer a variável resposta, um suavizador do gráfico de dispersão é frequentemente utilizado.

Um suavizador (*smoother*) pode ser definido como uma ferramenta para resumo da tendência das medidas Y como função de uma ou mais medidas X. É importante destacar que as estimativas das tendências terão menos variabilidade que as variáveis respostas observadas, o que explica o nome de suavizador para a técnica aplicada (HASTIE & TIBSHIRANI, 1990). Chamamos a estimativa produzida por um suavizador (*smoother*) de "*smooth*". O caso de uma variável preditora é chamado de suavizador em diagrama de dispersão.

Os suavizadores possuem dois usos principais, sendo o primeiro uso a descrição. Um gráfico de dispersão suavizador pode ser usado para melhorar a aparência visual de um gráfico de dispersão de Y vs X, para nos ajudar a encontrar uma tendência no gráfico. O segundo uso é de estimar a dependência da esperança de Y com o seus preditores e nos servem como blocos de construção para os modelos aditivos.

O suavizador mais simples é o caso dos preditores categóricos, como sexo (masculino,

feminino). Para suavizar Y podemos simplesmente realizar a médias dos valores de Y para cada categoria. Este processo captura a tendência de Y em X. Pode não parecer que simplesmente realizar as médias seja um processo de suavização, mas este conceito é a base para a configuração mais geral, já que a maioria dos suavizadores tenta imitar a média da categoria através da média local, ou seja, realizar a média dos valores de Y tendo os valores preditores próximos dos valores alvo. Esta média é feita nas vizinhanças em torno do valor alvo.

Nesse caso, tem-se duas decisões a serem tomadas:

• Como realizar a média dos valores da resposta em cada vizinhança;

• O quão grande esta vizinhança deve ser.

A questão de como realizar a média em uma vizinhança é a questão de qual tipo de suavizador utilizar, pois os suavizadores diferem principalmente pelo jeito de realizar as médias. O tamanho da vizinhança a ser tomada é normalmente expressa em forma de um parâmetro. Intuitivamente grandes vizinhanças irão produzir estimativas com variância pequena mas potencialmente com um grande viés e inversamente quando adotado vizinhanças pequenas. Portanto temos uma troca fundamental entre variância e viés estipulada pelo parâmetro suavizador. Este problema é análogo à questão de quantas variáveis preditoras colocar em uma equação de regressão.

3.3.1 Técnicas de suavização

Entre as principais técnicas de suavização estão a regressão paramétrica, vista anteriormente e que consiste em uma linha de regressão estimada por mínimos quadrados. Essa abordagem pode ou não ser apropriada para dado conjunto de dados.

O suavizador bin, também conhecido como regressograma, imita um suavizador categórico, particionando os valores preditores em regiões disjuntas e então realizando a média da resposta em cada região. A estimativa final não tem uma forma bem suavizada, pois é possível ver um salto em cada ponto de corte.

A média móvel (*running mean*) é outra técnica que leva em conta o cálculo da média. É muito comum utilizar uma vizinhança/região de $(2k + 1)$ observações, k para a esquerda e k para a direita de cada observação, onde o valor de k tem um comportamento de troca entre suavidade e qualidade do ajuste.

Um problema comum encontrado na média móvel é o viés. Uma saída é usar pesos para dar mais importância às vizinhanças mais próximas. Uma solução ainda melhor é utilizar a técnica de linha móvel (*running line*), na qual novamente são definidas as vizinhanças para cada ponto, tipicamente os k pontos mais próximos de cada lado. Nesse caso é mais interessante considerar a proporção de pontos em cada vizinhança, ou seja, $w = \frac{(2k+1)}{n}$, denominado *span*. Então ajusta-se uma linha de regressão aos pontos de cada região, que é usada para encontrar o valor predito suavizado para o ponto de interesse.

3.3.2 Loess

Também chamado de *Lowess*, essa técnica pode ser vista como uma linha móvel com pesos locais (*locally weighted running line*). Um suavizador desse tipo, seja denominado $s(x_0)$, usando k vizinhos mais próximos pode ser computada por meio dos seguintes passos:

- Os k vizinhos próximos de x_0 são identificados e denotados por $N(X_0)$;
- É computada a distância do vizinho-próximo mais distante de x_0:

$$\Delta(x_0) = max_{N_{x_0}} |X_0 - x_i|$$

- Pesos w_i são designados para cada ponto $(N_{x_0}$, usando a função de peso tri-cúbica:

$$W\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right)$$

  onde

$$W(u) = \begin{cases} (1-u^3)^3, & 0 \leq u \leq 1 \\ 0, & \text{caso contrário} \end{cases}$$

- $s(x_0)$ é o valor ajustado no ponto x_0 do ajuste de mínimos quadrados ponderados de y para x contidos em $N(X_0)$ usando os pesos computados anteriormente.

As hipóteses em relação ao modelo *Loess* são menos restritivas se comparadas às do modelo de regressão linear, já que assume-se que ao redor de cada ponto x_0 o modelo deve ser aproximadamente **uma função local?**.

Destaca-se que nessa técnica deve-se ter atenção à escolha do valor do *span*. Um valor muito pequeno faz com que a curva seja muito irregular e tenha variância alta. Por outro lado, um valor muito grande fará com que a curva seja sobre-suavizada, podendo não se ajustar bem aos dados e resultando em perda de informações e viés alto. Nos passos mostrados anteriormente o valor do *span* foi escolhido através do método de vizinhos mais próximos.

3.3.3 Kernels

Um suavizador kernel usa pesos que decrescem suavemente enquanto se distancia do ponto de interesse x_0. Vários métodos podem ser chamados de suavizadores kernel através dessa definição. Porém na prática, o suavizador kernel representa a sequência de pesos descrevendo a

forma da função peso através de uma função densidade com um parâmetro de escala que ajusta o tamanho e a forma dos pesos perto de x_0. Um suavizador Kernel pode ser definido da forma

$$\hat{y}_i = \frac{\sum_{j=1}^{n} y_i K\left(\frac{x_i - x_j}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b}\right)}$$

onde b é o tamanho da vizinhança (parâmetro de escala), e K uma função kernel, ou seja, uma função densidade. Existem diferentes escolhas para K, geralmente usa-se a densidade de uma Normal, tendo-se assim um kernel Gaussiano.

3.3.4 Splines

Um *Spline* pode ser visto como uma função definida por um polinômio por partes. Pontos distintos são escolhidos no intervalo das observações (nós) e um polinômio é definido para cada intervalo, dessa forma é possível modelar com polinômios mais simples as curvas mais complexas. Os *splines* dependem principalmente do grau do polinômio e do número e localização dos nós.

Essa técnica é interessante pois tem uma maior flexibilidade para o ajuste dos modelos em comparação com o modelo de regressão polinomial ou linear e, após a determinação da localização e quantidade de nós, o modelo é de fácil ajuste. Além disso, o *spline* permite modelar um comportamento atípico dos dados, o que não seria possível com apenas uma função.

3.3.5 Splines de regressão

Existem várias diferentes configurações para um *spline*, mas uma escolha popular é o *spline* cúbico, contínuo e contendo primeira e segunda derivadas contínuas nos nós. As *splines* cúbicas são as de menor ordem nas quais a descontinuidade nos nós são suficientemente suaves para não serem vistas a olho nu, então a não ser que seja necessário mais derivadas suavizadas, existe pouca justificativa para utilizar *splines* de maior ordem.

Para qualquer grupo de nós, o *spline* de regressão é ajustado a partir de mínimos quadrados em um grupo apropriado de vetores base. Esses vetores são as funções base representando a família do pedaço do polinômio cúbico, com valor dado a partir dos valores observados de X.

Uma variação do *spline* cúbico é o *spline* cúbico natural, que contêm a restrição adicional de que a função é linear além dos nós dos limites. Para impor essa condição, é necessário que, nas regiões dos limites: $f''' = f'' = 0$, o que reduz a dimensão do espaço de $K + 4$ para K, se há K nós. Então com K nós no interior e dois nos limites, a dimensão do espaço do ajuste é de $K + 2$.

Quando trabalha-se com *splines*, existe uma dificuldade em escolher a localização e quantidade ideal dos nós, sendo mais importante o número de nós do que sua localização. Salienta-se que incluir mais nós que o necessário pode resultar em uma piora do ajuste do modelo. Existem algumas maneiras para fazer essas escolhas, como por exemplo colocar os nós nos quantis das variável preditora (três nós interiores nos três quartis).

Outro problema é a escolha de funções base para representar o *spline* para dados nós. Suponha que os nós interiores são denotados por $\xi_1 < ... < \xi_k$ e os nós dos limites são ξ_0 e ξ_{k+1}. Uma escolha simples de funções base para um *spline* cúbico é conhecida como base de séries de potência truncada, que deriva de:

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{K} \theta_j (x - \xi_j)_+^3$$

Onde s tem as propriedades necessárias: é um polinômio cúbico em qualquer subintervalo $[\xi_j, \xi_{j+1})$, possui duas derivadas contínuas e possui uma terceira derivada.

A função s pode ser escrita como uma combinação linear de $K+4$ funções base $P_j(x)$: $P_1(x) = 1, P_2(x) = x$ e assim por diante. Cada função deve satisfazer as três condições e ser linearmente independente para ser considerada uma base. Então fica claro que são necessários $(K + 4)$ parâmetros para representar um *spline* cúbico.

As funções base B-*spline* fornecem uma alternativa numericamente superior para a base de séries de potência truncada. A ideia principal é de que qualquer função base $B_j(x)$ é diferente de zero em um intervalo de no máximo cinco diferentes nós. Fica claro que as funções B_j são *splines* cúbicas, e são necessárias $K + 4$ delas para abranger o espaço.

A partir disso, observa-se que *splines* de regressão podem ser atrativos devido à sua facilidade computacional, quando os nós são dados. Porém a dificuldade em escolher o número e localização dos nós pode ser uma grande desvantagem da técnica.

3.4 Seleção de Modelos - Enfoque de Predição

Em regressão, a fim de elaborar boas funções de predição, cria-se um critério para mensurar o desempenho que determinada função predição $g : \mathbb{R}^d \Rightarrow \mathbb{R}$, valendo-se, por exemplo, do do risco quadrático (IZBICKI E SANTOS, 2020).

$$R_{pred}(g) = E\left[(Y - g(X))^2\right],$$

constata-se que (X,Y) é uma observação nova não utilizada ao se estimar g. Sendo assim, melhor será a função de predição g, quanto menor for o risco. Outras funções de perda pode ser empregadas, porém, a função $L(g(X); Y) = (Y - g(X))^2$ (denominada função de perda quadrática), será utilizada.

Para se medir a performance de um estimador, baseando-se em seu risco quadrático, criar

uma boa função de predição equivale a encontrar um bom estimador para a função de regressão, sendo que a melhor função de prediçao para Y é a função de regressão.

Normalmente, é habitual ajustar distintos modelos para a função de regressão e encontrar qual deles apresenta um melhor poder preditivo, ou seja, aquele que possui o menor risco. Um modelo pode interpolar os dados e, mesmo assim, possuir um poder preditivo (IZBICKI E SANTOS, 2020).

O método de seleção de modelos pretende selecionar uma boa função g. Nesse sentido, usa-se o critério do risco quadrático para averiguar a qualidade da função. Assim, escolhe-se uma função g em uma classe de candidatos G que tenha um bom poder preditivo (baixo risco quadrático). Dessa maneira, visa-se evitar modelos que tenham sub ou super-ajuste.

O risco observado, conhecido como erro quadrático médio em relação aos dados de treinamento, e determinado por,

$$EQM(g) = \frac{1}{n} E \sum_{i=1}^{n} \left[(Y_i - g(X_i))^2 \right],$$

este estimador, se usado para realizar a seleção de modelos poder levar um super-ajuste (ajuste perfeito dos dados). Usualmente e comum dividir os dados em dois conjuntos, um de treinamento e validação. Utiliza-se os dados de treinamento para estimar a regressão e e avalia-se o erro quadrático médio por meio do conjunto de validação. Este procedimento de divisão é chamado de *data splitting*.

Algumas variações podem ser realizadas como o processo *k-fold cross validation* consiste em dividir a base dados em K conjuntos, no qual o modelo será treinado com K-1 conjuntos restantes onde o conjunto que ficou de fora na primeira vez será empregado como conjunto de teste e o algoritmo faz o rodízio entre os K conjuntos até que todos os dados sejam vistos como dados de treino e validação. Alternativamente, pode utilizar o *leave-one-out cross validation* (LOOCV), no qual o modelo é ajustado utilizando todas as observação com exceção da i-ésima delas. Será utilizado este último para o processo de avaliação e escolha do modelo.

4 Resultados e Discussão

4.1 Estudo de simulação

Nesta seção, serão utilizadas simulações de dados para gerar situações nas quais possam ser aplicadas as técnicas estudadas, analisando, assim, suas respectivas performances. Para os resultados obtidos, quatro técnicas de suavização serão empregadas, sendo elas: o suavizador de *kernel*, *Loess*, *splines* de regressão de grau 1 e grau 3. Realizar-se-ão ajustes para o primeiro cenário, considerando distintos parâmetros de suavização para avaliar, visualmente, os comportamentos das curvas em diagramas de dispersão. Em seguida, adotando o método de *data splitting*, *leave one out cross-validation*, encontrar-se-á um parâmetro de suavização que forneça a ocorrência do menor erro quadrático médio possível, comparando, desta forma, os métodos de suavização.

Posteriormente, este procedimento será repetido para cada cenário em mil amostras, contabilizando a quantidade de vezes em que cada técnica apresenta o menor erro quadrático médio. Por exemplo, para o primeiro cenário, será gerado mil amostras aleatórias de tamanho n. Para cada amostra será empregado o procedimento acima, salvando seus repectivos erros quadráticos médio. Ao final da simulação, será contabilizado se a ocorrência do erro quadrático médio em cada técnica foi mínima para, por fim, comparar estes resultados e verificar qual técnica obtém o melhor resultado em uma simulação de mil amostras. Vale ressaltar que serão empregados dois comportamentos, uma proveniente de uma função senoidal e outra de uma função Gamma: Cenário 1 e Cenário 2. Ainda, serão gerados nove sub-cenários, valendo-se da combinação de três tamanhos amostrais (150, 250 e 350), em três valores de desvio padrão distintos.

4.1.1 Cenário 1

Para este cenário, será considerado X uma sequência de 0 a 50 e Y, definido pela função

$$y = 10 + 5sen\pi\frac{x}{24} + \varepsilon,$$

onde ε é um termo aleatório, normalmente, distribuído com média zero e variância constante. Os tamanhos amostrais utilizados serão iguais a $150, 250$ e 350 e valores de desvio padrão $0.5, 1$ e 2. Na Figura 1, tem-se o comportamento dos dados para cada desvio padrão, considerando 350 observações com a curva : $10 + 5sen\pi\frac{x}{24}$.

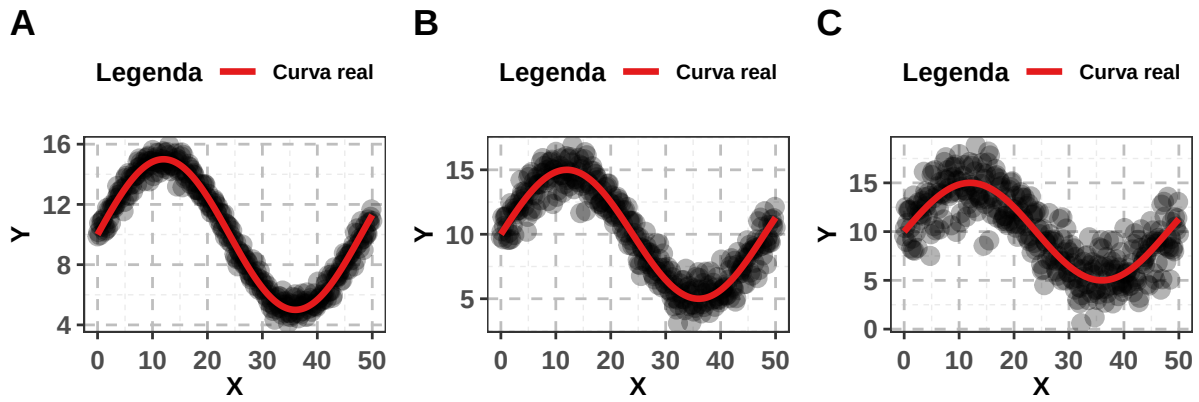

Figura 1: Gráfico de dispersão dos dados simulados e curva real.

Levando em conta a configuração do gráfico C (Figura 1), serão ajustadas curvas distintas para cada método de suavização, adotando parâmetros de suavização arbitrários apresentados na Figura 2.

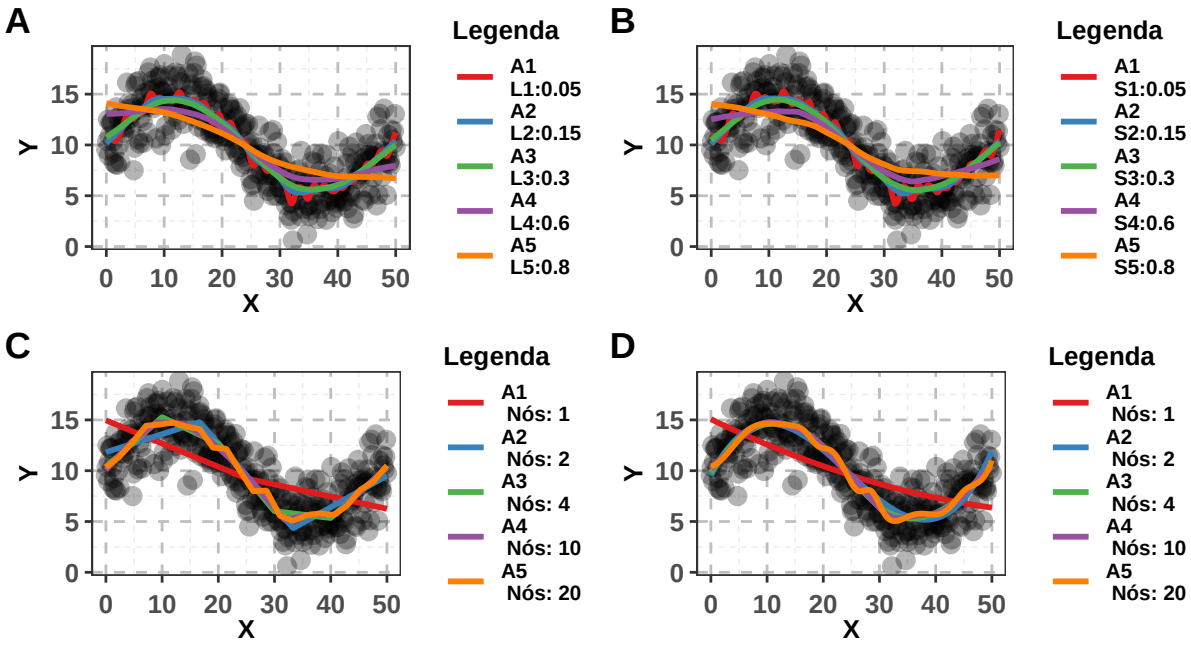

Figura 2: Comparação entre diferentes ajustes, com parâmetros de suavização distintos, considerando os suavizadores (A) Kernel, (B) Loess, (C) Splines de Regressão Grau 1 e (D) Splines de Regressão Grau 3.

Ao avaliar os ajustes dos gráficos A e B, verifica-se que, conforme o parâmetro de suavização aumenta, estes tendem a ser muito suaves. Ou seja, na medida em que o parâmetro se torna suficientemente grande, a curva tenderá a ser uma reta. Porém, quando este parâmetro tende a ser muito pequeno, o ajuste realizado interpolará os dados. Para os gráficos C e D, na proporção em que o parâmetro de suavização aumenta, a curva ajustada apresenta rugosidade em sua forma. Conforme este parâmetro se torna pequeno, a curva tenderá a ser uma reta.

Os suavizadores em diagramas de dispersão remetem a uma ideia visual de como o

ajuste está se comportando em relação aos dados. Na Figura 3, são apresentados os gráficos contendo os resultados do erro quadrático médio mínimo obtidos, realizando o *leave one out cross-validation*. Neste gráfico, verifica-se o comportamento dos erros quadrático médio em relação a seus repectivos parâmetros de suavização. Ainda, nota-se para qual parâmetro de suavização haverá o melhor ajuste, levando em consideração o erro menor erro quadrático possível dentre todos os candidatos.Em outras palavras, ao realizar ajustes controlando o valor do parâmetro de suavização, obter-se-á um ajuste no qual o Erro Quadrático Médio (EQM) será o menor de todos, logo, este será o melhor candidato para representar os dados. Observa-se, então, os parâmetros que, supostamente, induzirão um melhor ajuste para cada suavizador.

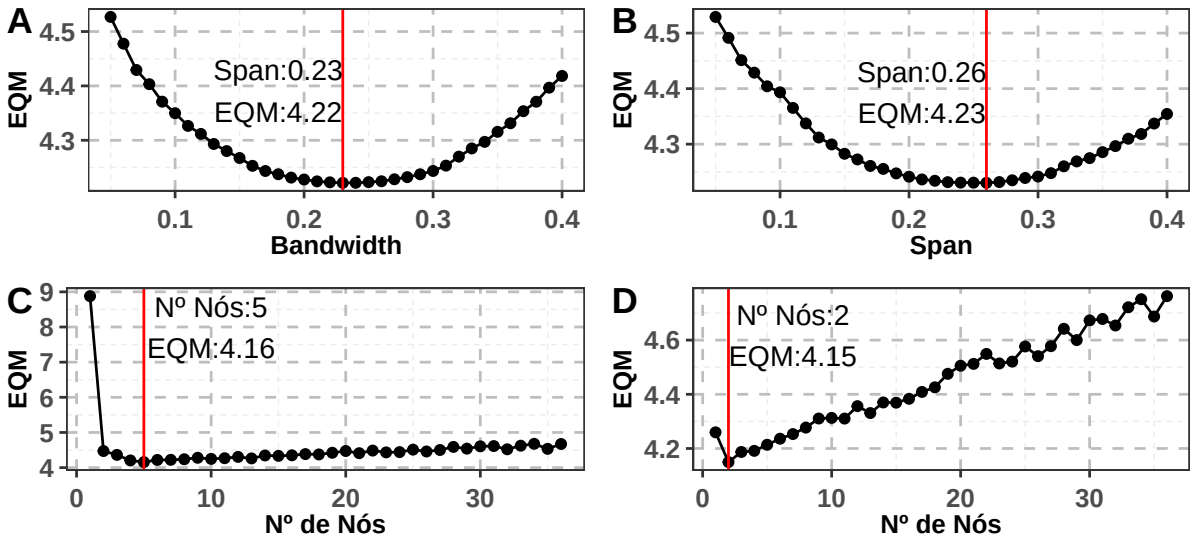

Figura 3: Erro quadrático médio versus parâmetro de suavização pós aplicação do Leave One Out Cross-Validation. (A) Kernel, (B) Loess, (C) Splines de Regressão Grau 1 e (D) Splines de Regressão Grau 3.

Na Figura 4, são demonstrados os ajustes, levando em consideração os melhores parâmetros obtidos por meio do processo de validação cruzada.

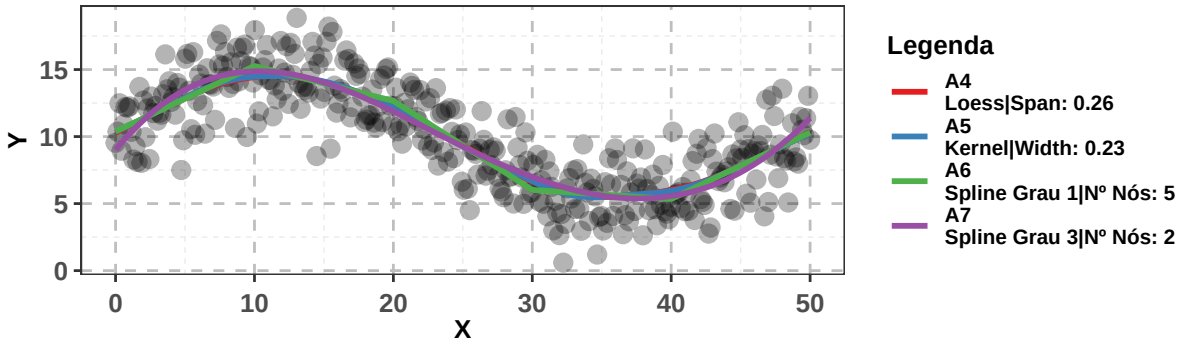

Figura 4: Comparação dos ajustes entre os métodos de suavização

Com o auxilio da Tabela 1, concluir-se-á que o melhor método de suavização, considerando o erro quadrático médio mínimo para esta amostra simulada é o *splines* de regressão cúbico. Porém, observando e analisando os resultados obtidos apenas em uma amostra, levaria à decisões equivocadas.

Tabela 1: Erro Quadrático Médio para os suavizadores Loess, Kernel e Spline Cúbico

Smoother	EQM
Kernel	4.22
Loess	4.23
Splines Grau 1	4.16
Splines Grau 3	4.15

Até então, o procedimento para escolha do melhor parâmetro de suavização para os métodos foi descrito, sendo a técnica com o melhor desempenho aquela que obtiver o menor erro quadrático médio. Neste momento, será reanalisado este procedimento em um processo de geração de amostras aleatórias. Serão considerados os cenários apresentados no começo deste capítulo: três tamanhos amostrais, para três variabilidades, totalizando nove cenários distintos. Para cada cenário serão geradas mil amostras aleatórias, aplicando o procedimento *leave one out cross-validation* para cada método de suavização. Em seguida, contabilizar-se-á a quantidade de vezes em que cada técnica obteve erro quadrático mínimo. Por fim, avaliar-se-á qual técnica obteve o melhor desempenho.

Na Tabela 2, estão esquematizados os resultados obtidos para cada cenário, considerando mil amostras simuladas para os suavizadores *Kernel*, *Loess*, *Splines* de regressão grau 1 e *Splines* de regressão grau 3.

Tabela 2: Percentual do Erro quadrático mínimo para cada suavizador em relação a 1000 amostras.

TAMANHO	VAR	Kernel	Loess	Sp. Reg. 1	Sp. Reg. 3
150	0.5	0.9%	5.3%	19.7%	74.1%
150	1.0	1.4%	8.4%	31.1%	59.1%
150	2.0	1.4%	10.8%	48%	39.8%
250	0.5	0.7%	3.8%	16.9%	78.6%
250	1.0	1.1%	7%	23.5%	68.4%
250	2.0	1.7%	13.1%	35.8%	49.4%
350	0.5	0.8%	3.7%	15.8%	79.7%
350	1.0	1.3%	7%	20.5%	71.2%
350	2.0	1.8%	11.7%	33.5%	53%

Destaca-se o suavizador *splines* cúbico por obter o melhor desempenho em quase todos os cenários, exceto o terceiro (tamanho de amostra igual a 150 e DP = 2), no qual o *splines* de

grau 1 obteve erro quadrático médio mínimo em cerca de 48% das amostras simuladas. Ainda quando fixamos o valor do tamanho amostral, constata-se que, conforme a variabilidade dos dados aumenta, há indicios de que os resultados obtidos para a técnica *splines* cúbico estejam se dispersando para as demais técnicas. Observa-se ainda que conforme o tamanho amostral aumenta, há indícios de que os percentuais estejam convergindo e estabilizando, envidenciando que os *splines* cúbico tendem a obter um desempenho melhor em relação aos demais métodos. Além disso, ressalta-se que os percentuais para o suavizador de *kernel* foram os piores, obtendo EQM's mínimo, de até no máximo 1,8% das amostras em relação aos cenários simulados.

Na Figura 5, são apresentados os comportamentos dos erros quadráticos médios obtidos do processo de simulação das amostras. De forma sucinta, verifica-se uma tendência decrescente conforme o tamanho amostral aumenta, assim como a sua aplitude tende a ficar menor. Ademais, visualmente, não é observada diferença significativa entre as técnicas *Kernel* e *Loess*. Ainda, percebe-se que os *splines* (grau 1 e grau 3) apresentam um comportamento mediano relativamente inferior quando comparado aos demais.

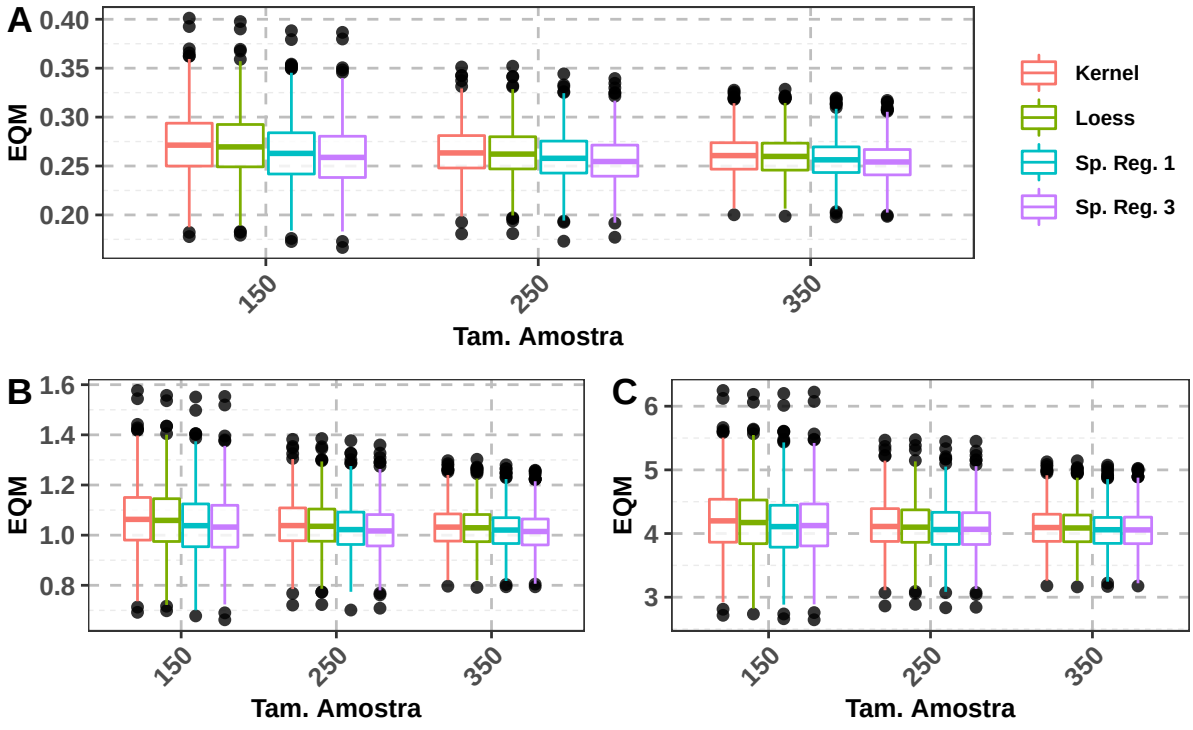

Figura 5: Comparação do erro quadrático para as 1000 amostras para cada suavizador por (A) DP = 0.5, (B) DP = 1 e (C) DP = 2

Portanto, levando em consideração a Tabela 2 e a Figura 5 pode-se concluir que, para o Cenário 1, o suavizador *splines* de regressão cúbico tende apresentar um melhor desempenho em relação às demais técnicas, sendo o método mais indicado para representar e captar o comportamento.

4.1.2 Cenário 2

Para este cenário, os valores para x serão uma sequência de 0.1 à 2. Ainda, temos que $y = f(x) + \varepsilon$, com $f(x) \sim Gamma(6, 10)$ e $\varepsilon \sim N(0, \sigma^2)$. Serão considerados tamanhos amostrais iguais a $150, 250$ e 350 e valores de desvio padrão $0.05, 0.1$ e 0.15. Na Figura 6, tem-se o comportamento dos dados para cada desvio padrão, considerando 350 observações.

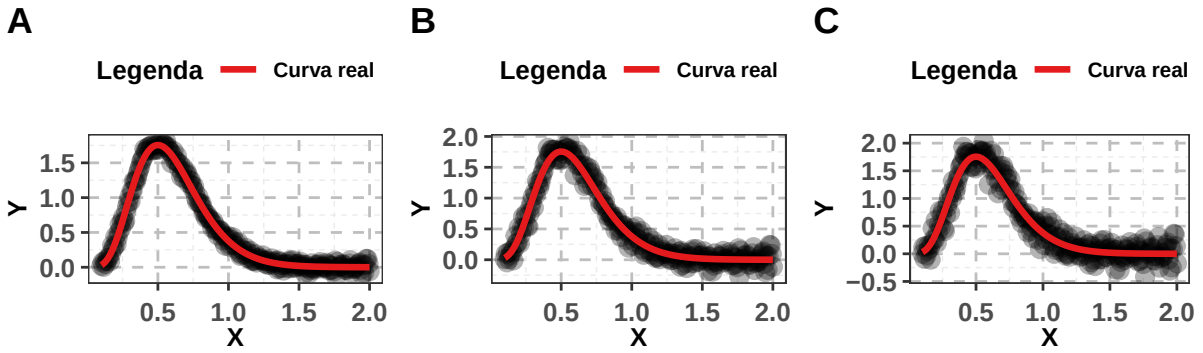

Figura 6: Gráfico de dispersão dos dados gerados para o estudo de simulação.

Analogamente, ao que foi proposto no Cenário 1, será avaliado qual suavizador apresenta um melhor desempenho de predição. A Tabela 3 apresenta os valores do erro quadrático médio proveniente do processo de *leave one out cross-validation*, considerando apenas uma amostra. Ressalta-se que os valores obtidos entre as técnicas estão bem próximo entre si, porém o *splines* de regressão de grau 1 sendo que o de menor valor.

Tabela 3: Erro Quadrático Médio para os suavizadores Loess, Kernel, Splines de Regressão Linear e Splines de Regressão Cúbico

Smoother	EQM
Kernel	0.0243
Loess	0.0241
Splines Grau 1	0.0234
Splines Grau 3	0.0246

Quando realizamos a análise aplicando o procedimento de simulação de amostras, que se encontra sumarizado na Tabela 4, observa-se que o suavizador *splines* de regressão cúbico apresentou o melhor desempenho em todos cenários, obtendo o erro quadrático médio mínimo indo de aproximadamente 70% à 92% das amostras entre os cenários simulados. Ainda, ressalta-se que, conforme o tamanho amostral aumenta, o percentual apresenta uma tênden cia de aumento e indícios de estabilização. Quando fixado um tamanho amostral, o desempenho entre as amostras simuladas com variabilidade distinta apresenta uma tendência decrescente.

De forma análoga ao Cenário 1, o comportamento para os EQM's (vide Figura 7), apresenta uma tendência decrescente conforme o tamanho amostral aumenta, assim, percebe-se que

Tabela 4: Percentual do Erro quadrático mínimo para cada suavizador em relação a 1000 amostras, para o Cenário 2

TAMANHO	VAR	Kernel	Loess	Sp. Reg. 1	Sp. Reg. 3
150	0.05	0.6%	2.6%	7.4%	89.4%
150	0.10	0.7%	6.3%	13.5%	79.5%
150	0.15	0.7%	9.1%	20.6%	69.6%
250	0.05	0.4%	1.7%	8.2%	89.7%
250	0.10	0.7%	3.4%	10.3%	85.6%
250	0.15	0.8%	6.5%	12.7%	80%
350	0.05	0.2%	2%	5.9%	91.9%
350	0.10	0.6%	2.5%	8.9%	88%
350	0.15	1%	4.7%	11.4%	82.9%

a amplitude tende a ficar menor. Ainda, percebe-se que não há diferença signficativa entre os suavizadores *kernel* e *loess*. Os *splines* apresentam ter um comportamento mediano inferior, com destaque ao *splines* cúbico que, visualmente, aparenta ser menor quando comparados com as demais técnicas.

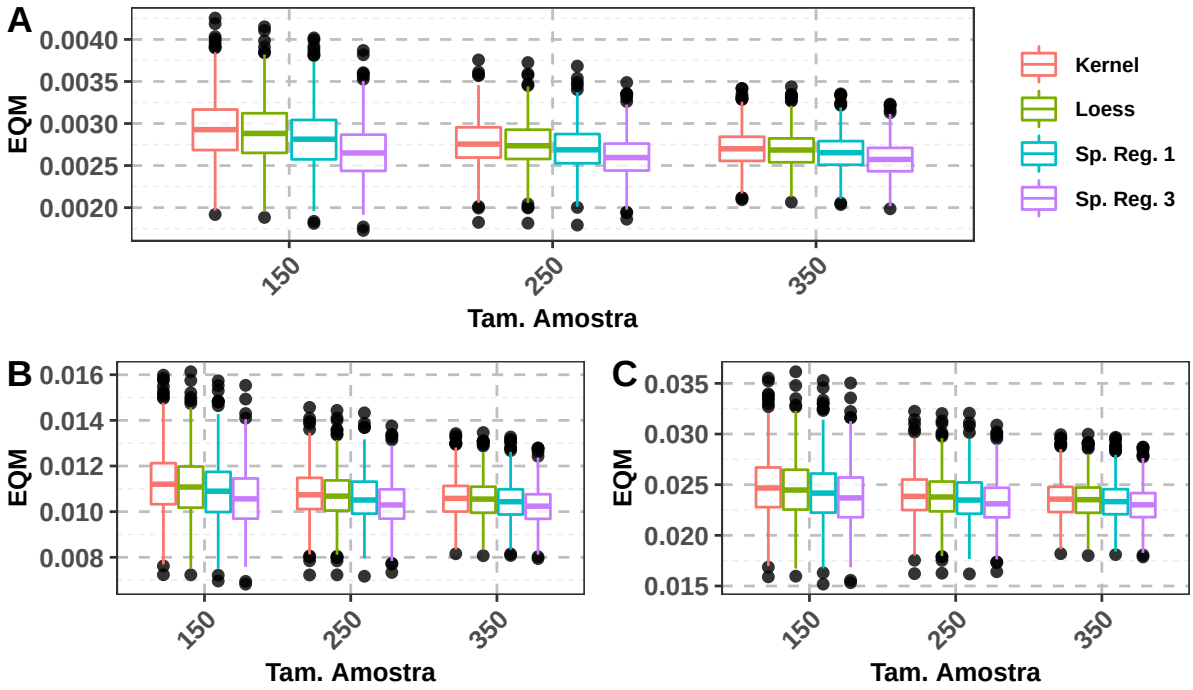

Figura 7: Comparação do erro quadrático para as 1000 amostras para cada suavizador por (A) DP = 0.05, (B) DP = 0.1 e (C) DP = 0.15

Portanto, baseado na Tabela 4 e na Figura 7, os dados evidenciam que o suavizador *splines* cúbico possui um desempenho melhor em relação aos outros suavizadores.

5 Referências

BUJA, A., HASTIE, T. & TIBSHIRANI, R. (1989). **Linear smoothers and additive models**. The Annals of Statistics, 17, 453-510.

CLEVELAND, W. S. (1979). **Robust locally weighted regression and smoothing scatterplots**. Journal of the American Statistical Association, 74, 829-836.

DELICADO, P., 2008 **Curso de Modelos no Paramétricos** p. 200.

EUBANK, R. L(1999) **Nonparametric Regression and Spline Smoothing**. Marcel Dekker, 2o edição. Citado na pág. 1, 2, 29

FAHRMEIR, L. & TUTZ, G. (2001) **Multivariate Statistical Modelling Based on Generalized Linear Models**. Springer, 2o edição. Citado na pág. 15

GREEN, P. J. & YANDELL, B. S. (1985) **Semi-parametric generalized linear models**. Lecture Notes in Statistics, 32:4455. Citado na pág. 15

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

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). **Generalized additive models**, volume 43. Chapman and Hall, Ltd., London. ISBN 0-412-34390-8.

MONTGOMERY, D. C. & PECK, E. A. & VINING, G. G. **Introduction to Linear Regression Analysis**. 5th Edition. John Wiley & Sons, 2012.

IZBICK, r & SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. ISBN 978-65-00-02410-4.

TEAM, R. CORE. R: **A language and environment for statistical computing**. (2013).