UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS
CURSO DE ESTATÍSTICA

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA ESTIMAÇÃO DE MODELOS ADITIVOS

Marco Aurelio Valles Leal

Maringá
2021

MARCO AURELIO VALLES LEAL

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA ESTIMAÇÃO DE 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 ESTIMAÇÃO DE 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

Intituição do professor membro da banca

_____

Membro da banca

Nome do professor membro da banca

Intituição do professor membro da banca

rre

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 . . . 2

2 Referencial Teórico **3**

3 Metodologia **3**
3.1 Regressão Linear . . . 3
3.1.1 Estimação dos parâmetros pelo Métodos dos Mínimos Quadrados . 4
3.1.2 Estimação de Máxima Verossimilhança . . . 5
3.1.3 Estimação de σ^2 . . . 6
3.1.4 Teste de Hipóteses para β_0 e β_1 . . . 7
3.1.5 Análise de Variância . . . 8
3.1.6 Regressão Linear Múltipla . . . 9
3.1.7 Adequação do modelo - enfoque inferencial . . . 10
3.2 Modelo Aditivo . . . 11
3.3 Suavizadores . . . 12
3.3.1 Técnicas de suavização . . . 13
3.3.2 Loess . . . 14
3.3.3 Kernels . . . 14
3.3.4 Splines . . . 15
3.3.5 Splines de regressão . . . 15
3.3.6 Backfitting . . . 16
3.4 Seleção de Modelos - Enfoque de Predição . . . 17

4 Resultados e Discussão **17**
4.1 Estudo de simulação . . . 17
4.1.1 Cenário 1 . . . 17
4.1.2 Cenário 2 . . . 21
4.1.3 Cenário 3 . . . 27
4.2 Aplicações . . . 29
4.2.1 Aplicação 1 . . . 29
4.2.2 Aplicação 2 . . . 32

5 Conclusão **37**

6 Referências **38**

1 Introdução

Análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis, sendo essa uma técnica amplamente utilizada na estatística. Usualmente, é de interesse apenas uma variável, chamada de variável resposta ou dependente, e desejamos estudar como esta variável depende de um conjunto de variáveis observáveis, chamadas de variáveis explicativas ou independentes. Nesse contexto, os modelos de regressão linear simples e múltipla podem ser utilizados.

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 e determinar uma função que estima a relação correta existente nem sempre é fácil. Uma alternativa é 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.

As funções suaves do componente sistemático do modelo podem ser estimadas através de um suavizador (*smoother*). Entretanto, algumas técnicas de suavização podem não ser viáveis em alguns problemas.

Portanto, o objetivo do presente trabalho é apresentar os modelos aditivos e 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 utilizar modelos mais flexiveis, em determinadas situações. O *software* estatístico R (*R Team Core*) será utilizado para a realização de todas as análises do estudo.

1.1 Objetivo Geral

O objetivo deste projeto 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 de regressão linear e apresentar os principais métodos de estimação dos parâmetros do modelo assim como as técnicas de diagnóstico e qualidade do ajuste;
- Apresentar técnicas de suavização utilizadas para estimar funções não paramétricas presentes nos modelos aditivos, identificando suas principais características;

- Introduzir os modelos aditivos, especificando sua principais características;
- Apresentar métricas para verificar a qualidade de predição;
- Realizar um estudo de simulação para verificar 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.

2 Referencial Teórico

3 Metodologia

3.1 Regressão Linear

Análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis. Aplicações de regressão são numerosas e ocorrem em quase todos os campos, incluindo engenharia, as ciências físicas e químicas, economia, ciência biológicas, etc. 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.

Um 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 de Y correspondentes a valores de X que não estavam entre os dados. Além disso, através do modelo é possível estimar parâmetros e fazer inferências sobre os mesmos, como testes de hipóteses e intervalos de confiança.

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:

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

O intercepto β_0 e a inclinação da reta β_1 são constantes desconhecidas 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. Os erros possuem distribuição normal: $\varepsilon_i \sim N(0, \sigma^2)$. A média da distribuição é dada por uma função linear de x:

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

Trata-se a variável regressora x controlada pelos dados e mensurada com um erro, enquanto a variável resposta Y é aleatória, ou seja, há uma distribuição de probabilidade para y a cada possível valor de x.

Os parâmetros β_0 e β_1 também são chamados de coeficientes da regressão, tendo 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 Estimação dos parâmetros pelo Métodos dos Mínimos Quadrados

Os parâmetros β_0 e β_1 são desconhecidos e devem ser estimados usando dados de uma amostra. Suponha que tem-se n pares de dados, $(x_1, y_1), ..., (x_n, y_n)$. Esses dados podem ter sido resultado de um experimento controlado feito especificamente para coletá-los, de um estudo observacional, etc. Uma maneira de estimar esses parâmetros, é utilizando o Método dos Mínimos Quadrados, onde não é necessário conhecer a distribuição dos erros. Esse método tem como objetivo encontrar os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros ou desvios do modelo. A equação de regressão linear amostral é definida como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ para i=1,2...n} \tag{3}$$

A partir disso tem-se:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \Rightarrow \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Rightarrow S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Denomina-se os estimadores de mínimos quadrados de β_0 e β_1, $\hat{\beta}_0$ e $\hat{\beta}_1$, respectivamente. Para obtê-los deve-se encontrar os valores para $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam a equação acima, ou seja, deve-se derivar a equação, igualar a zero, isolar os parâmetros e, se a segunda derivada em relação a cada parâmetro for positiva, os valores encontrados são os que minimizam a equação. Sendo assim é fácil verificar que:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Note que a soma dos quadrados dos desvios da média de x denotamos por

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

E a soma dos produtos cruzados dos desvios de x e y

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Dessa forma, é conveniente escrever que

$$\hat{\beta}_1 = \frac{S_{xy}}{Sxx}$$

Então, $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores de mínimos quadrados do intercepto e da inclinação respectivamente. Um resultado importante a respeito da qualidade dos estimadores de mínimos quadrados é o **Teorema Gauss-Markov**, dizendo que para um modelo de regressão com as suposições $E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$ e erros independentes, esses estimadores são não viesados e têm variância mínima entre todos os estimadores não viesados que são combinações lineares dos y_i, com $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$ e $Var(\hat{\beta}_0) = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)$.

E portanto podemos definir, o modelo de regressão linear simples estimado como sendo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4}$$

A diferença entre o valor observado y_i e o correspondente valor estimado é o resíduo, que é importante para verificar a adequação do modelo. Matematicamente, o i-ésimo resíduo é

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x), \quad i = 1, ..., n$$

3.1.2 Estimação de Máxima Verossimilhança

É possível mostrar que os estimadores de máxima verossimilhança para os parâmetros coincidem com os estimadores de mínimos quadrados quando os erros são independentes e normalmente distribuídos.

O modelo é $y_i = \beta_0 + \beta_1 x_i + \varepsilon$ e $\varepsilon \sim N(0, \sigma^2)$, em que ε_i é independente de ε_j para $i \neq j$. E a densidade dos erros é dada por:

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)$$

Tem-se que a função de verossimilhança é

$$L(\beta_0, \beta_1, \sigma^2 | \varepsilon_1, ..., \varepsilon_n) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i}^{n}\varepsilon_i^2\right)$$

Como $\varepsilon_i = y_i - \beta_0 - \beta_1 x_1$, tem-se

$$L(\beta_0, \beta_1, \sigma^2|\varepsilon_1, ..., \varepsilon_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

A função log-verossimilhança pode ser vista como

$$lnL(y, X, \beta, \sigma^2) = -\frac{n}{2}ln(2\pi) - nln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Para um valor fixo de σ,a função acima é maximizada quando o termo $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ é minimizado, ou seja, os estimadores para o β_0 e β_1 são iguais aos estimadores de mínimos quadrados. Esses estimadores são não viesados, têm variância mínima comparado com todos os outros estimadores não viesados, são consistentes e são um conjunto de estatísticas suficientes.

3.1.3 Estimação de σ^2

Além de estimar β_0 e β_1, um estimador de σ^2 é necessário para testar hipóteses e construir intervalos pertinentes ao modelo de regressão. O ideal seria que o estimador não dependesse da adequação do modelo estimado. Porém isso só é possível quando existem várias observações de y para pelo menos um valor de x, ou quando uma informação a respeito de σ^2 está disponível.Quando essa abordagem não pode ser usada, o estimador de σ^2 é obtido através da soma do quadrado dos erros (ou resíduos).

$$\begin{aligned} SQE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\ = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 S_{xy} = Syy - \hat{\beta}_1 S_{xy} \end{aligned}$$

SQE é um estimador viesado para σ^2, pois $E(SQE) = \sigma^2(n-2)$. Então, um estimador não viesado para σ^2 é dado pelo quadrado médio dos erros

$$QME = \hat{\sigma}^2 = \frac{SQE}{n-2}$$

A raiz quadrada de $\hat{\sigma}^2$ é chamada de erro padrão da regressão, tendo as mesmas unidades da variável resposta Y. Devido à $\hat{\sigma}^2$ depender da soma do quadrado dos erros, qualquer violação das suposições dos erros do modelo ou qualquer especificação errada da forma do modelo pode tornar $\hat{\sigma}^2$ inutilizável como estimador de σ^2.

O estimador de máxima verossimilhança de σ^2 é dado por,

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{n}$$

Note que apesar de ser um estimador viesado, conforme n se torna grande o suficiente, este se torna assintóticamente não viesado.

3.1.4 Teste de Hipóteses para β_0 e β_1

Seja, $\varepsilon_i \sim N^{iid}(0, \sigma^2)$, logo temos que $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Queremos testar a hipótese de β_1 ser igual a uma constante qualquer, que chamaremos de β_{10}, já que por meio desse teste temos um indicativo de associação linear entre as variáveis Y e X. Então, definimos as hipóteses $H_0 : \beta_1 = \beta_{10} \times H_1 : \beta_1 \neq \beta_{10}$. Temos que $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x}^2)}\right)$. Assim,

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1)$$

Como normalmente σ^2 é desconhecido, não podemos usar Z_0, como sabemos que QME é um estimador não viciado para σ^2, conseguimos definir a seguinte a estatística:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{QME/\sum (x_i - \bar{x})^2}} \sim t_{n-2}$$

Logo calculamos a estatística t_0 e a comparamos com ponto $t_{\alpha,n-2}$ da distribuição T de Student, ou seja, a regra de decisão pode ser definida e rejeitamos a hipótese nula quando $|t_0| > t_{\alpha/2,n-2}$.

Quando estamos interessados em testar o parâmetro β_0, usamos o mesmo processo, definimos as hipóteses a respeito de β_0 e a estatística do teste e, por fim, comparamos se esta estatística pertence a uma região rejeição ou de não rejeição para hipótese nula H_0.

Quando estamos interessados em avaliar se existe uma boa "correlação" entre a resposta e a variável explicativa, o teste de significância da regressão é um caso específico, onde definimos a hipótese nula testando β_1 igual a zero, ou seja:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Logo se não rejeitamos a hipótese nula, $H_0 : \beta_1 = 0$, implica dizer que não há relação linear entre x e y. Uma outra interpretação seria dizer que, o valor de x pode ser suficientemente pequeno e sendo assim, afeta muito pouco a variância de y, e que o melhor estimador de y para qualquer x seria, $\hat{y} = \bar{y}$. Contudo se rejeitarmos a hipótese nula, podemos dizer que x e um valor suficientemente grande e que exerce influência significativa sob a variabilidade em y. Logo mais, a estatística do teste, quando $\beta_{10} = 0$, é definida :

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{QME/\sum (x_i - \bar{x})^2}} \sim t_{n-2}$$

E rejeitamos a hipótese nula quando $|t_0| > t_{\alpha/2,n-2}$.

3.1.5 Análise de Variância

A análise de variância é baseada no particionamento da variância total da variável resposta y, ou seja, a variação total é decomposta na soma de duas fontes de variação distintas, uma fonte de variação proveniente das observações em relação ao valores ajustados $(Y_i - \hat{Y}_i)$, e a outra dos valores ajustados em relação à média $(\hat{Y}_i - \bar{Y}_i)$. Portanto a tabela da análise de variância é derivada a partir deste particionamento assim como a estatística F utilizada no teste.

$$SQT = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 \quad (5)$$

$$SQT = SQR + SQE \quad (6)$$

Na equação (6), a soma de quadrado total (SQT) mensura a variabilidade total da variável dependente, a soma de quadrado da regressão (SQR) mensura a variabilidade da variável dependente explicada pelo modelo e a soma de quadrados dos erros (SQE) mensura a variabilidade residual, não explicada pelo modelo. Esta equação é chamada de identidade fundamental da análise de variância para o modelo de regressão. Ainda podemos definir a soma de quadrados da regressão como $SQR = \hat{\beta}_1 S_{xy}$, onde $S_{xy} = \sum y_i(x_i - \bar{x})$. Ainda os quadrados médios podem ser definidas provenientes da divisão da soma de quadrados pelos seus respectivos graus de liberdade. Portanto temos que os quadrados médios da regressão (QMR) e os quadrados médios dos erros (QME) podem ser definidos como:

- $QMR = \dfrac{SQR}{1} = SQR = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2$, é o quadrado médio da regressão.
- $QME = \dfrac{SQE}{n-2} = \dfrac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2}$, é quadrado médio dos erros (Resíduos).

E por fim, para definirmos a estatística do teste precisamos de alguns resultados sabemos que $\frac{(n-2)SQE}{\sigma^2} \sim \chi^2_{n-2}$ e $\frac{SQR}{\sigma^2} \sim \chi^2_1$. Ainda para que estatística definida a seguir seja válida, assumimos que a soma de quadrados dos erros (SQE) e a soma de quadrados da regressão SQR são independentes. Portanto sabemos que, o resultado da divisão de duas qui-quadrados nos fornece uma distribuição F, ou seja :

$$F_0 = \frac{\frac{SQR}{\sigma^2}}{\frac{(n-2)SQE}{\sigma^2}} = \frac{QMR}{QME} \sim F_{1,n-2}$$

Ainda podemos verificar que $E(SQE) = \sigma^2$ e $E(SQR) = \sigma^2 + \beta_1^2 S_{xx}$. Sob H_0 verdadeira, se o valor observado de F_0 é significantemente grande então provavelmente a inclinação $\beta 1 \neq 0$, logo teremos indícios para rejeitar a hipótese nula. Logo rejeitamos a hipótese nula com um nível de significância α quando $F_0 > F_{1-\alpha,1,n-2}$, ou seja, quando o valor calculado da estatística F_0 for maior do que o quantil 1-α da distribuição F com 1 grau de liberdade para o numerador e n-2 graus de liberdade para o denominador. A seguir a tabela 1, nos mostra um resumo do procedimento para análise de variância:

Tabela 1: Análise de Variância para o teste de significância da regressão linear simples

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F_0
Regressão	$SQR = \hat{\beta}_1 S_{xy}$	1	QMR	$\frac{QMR}{QME}$
Resíduos	$SQE = SST - \hat{\beta}_1 S_{xy}$	n-2	QME	
Total	SST	n-1		

3.1.6 Regressão Linear Múltipla

Como pode ser visto anteriormente, o modelo de regressão linear simples, com uma variável explicativa, aplica-se a várias situações. Entretanto, diversos problemas envolvem dois ou mais regressores influenciando o comportamento da variável resposta (dependente) y.

Suponha que o rendimento, em libras, de conversão em um processo químico dependa da temperatura e da concentração do catalisador. Um modelo de regressão múltipla que talvez descreva esse relacionamento é dado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

onde y é o rendimento, x_1 a temperatura e x_2 a concentração do catalisador. Esse é um modelo de regressão linear múltipla com duas variáveis regressoras. O termo linear é usado porque a equação é uma função linear dos parâmetros desconhecidos β_1, β_0 e β_2. Outro ponto a ser destacado é que esse modelo forma um plano no espaço tridimensional de y, x_1 e x_2. O parâmetro β_0 é o intercepto do plano; se os dados incluem $x_1 = x_2 = 0$, então β_0 é a média de y quando $x_1 = x_2 = 0$. Caso contrário, β_0 não tem interpretação prática. Já β_1 indica a mudança na resposta média y a cada mudança unitária de x_1 quando x_2 é constante. O parâmetro β_2 indica a mudança na resposta média a cada unidade de mudança em x_2, quando x_1 é constante. Em geral, a resposta y pode estar relacionada a k regressores ou variáveis preditoras. O modelo

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

é chamado de modelo de regressão linear múltipla com k regressores. Os parâmetros $\beta_j, j = 0, 1, ..., k$ são os coeficientes de regressão. Esse modelo descreve um hiperplano no espaço k-dimensional das variáveis regressoras x_j. 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 frequentemente usados como modelos empíricos ou funções aproximadas. Isso 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.

Modelos que tem uma estrutura mais complexa ainda podem ser analisados a partir de técnicas de regressão linear múltipla. Como exemplo, consideremos o modelo polinomial cúbico:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

Se considerarmos $x_1 = x, x_2 = x^2$ e $x_3 = x^3$ temos

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

que é um modelo de regressão linear múltipla com 3 variáveis regressoras.

Isso também é válido para modelos que incluem efeitos de interação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

Considere $x_3 = x_1 x_2$ e $\beta_3 = \beta_{12}$, temos

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

que é um modelo de regressão linear.

Pode ser que o gráfico do modelo (superfície gerada) não seja linear. Mas, em geral, qualquer modelo de regressão linear que é linear nos parâmetros (β's) é um modelo de regressão linear, independente da superfície gerada. Vale ressaltar que os modelos de regressão linear múltipla são frequentemente usados como modelos empíricos ou funções aproximadas.

3.1.7 Adequação do modelo - enfoque inferencial

As suposições mais importantes que foram vistas neste estudo sobre análise de regressão foram:

- A relação entre a resposta y e os regressores é linear;
- A média do termo do erro ϵ é zero;

- O termo do erro ϵ tem variâcia constante σ^2;
- Os erros não são correlacionados;
- Os erros são normalmente distribuídos.

Juntas, as duas última suposições implicam que os erros são variáveis aleatórias independentes. A última suposição é requisito para testar hipóteses e construir intervalos de confiança.No enfoque inferencial, deve-se sempre considerar essas suposições como duvidosas e fazer análises para examinar a adequação do modelo. Violações graves das suposições podem gerar um modelo instável, ou seja, uma amostra diferente pode levar a um modelo completamente diferente com conclusões opostas.Neste momento serão apresentados métodos para diagnosticar as violações das suposições básicas da regressão. Esses métodos são baseados principalmente no estudo dos resíduos do modelo. Os resíduos foram definidos anteriormente como, $e_i = y_i - \hat{y}_i, \qquad i = 1, 2, ..., n$ onde y_i é a observação e $\hat{y}_i$ é o valor estimado correspondente. É uma medida de variabilidade na variável resposta não explicada pelo modelo de regressão. Também é conveniente pensar nos resíduos como os valores observados ou realizados dos erros do modelo. Assim, quaisquer desvios das suposições sobre os erros devem aparecer nos resíduos.

A plotagem dos resíduos é uma ferramenta muito eficaz para investigar até que ponto o modelo de regressão se encaixa nos dados e verificar as suposições para o modelo de regressaõ. Às vezes, é útil trabalhar com resíduos dimensionados, que ajudam a encontrar observações que são outliers ou valores extremos, isto é, observações que estão separadas de alguma forma do resto dos dados.

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 seja de longe 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 (HASTIE & TIBSHIRANI, 1990). Um modelo aditivo é definido por

$$y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

onde os erros ϵ são independentes, com $E(\epsilon) = 0$ e $var(\epsilon) = \sigma^2$. As $f_j s$ são funções univariadas arbitrárias, uma para cada preditor. 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.

Os modelos aditivos mantêm muitas das boas propriedades dos modelos lineares, porém são mais flexíveis. 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. A função de resposta parcial f_j desempenha esse mesmo papel em um modelo aditivo.

3.3 Suavizadores

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). Não assume forma rígida entre a dependência de Y e de suas medidas preditoras $(X_1, X_2, ..., X_p)$. 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 se 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 reposta 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, $w = \dfrac{(2k+1)}{n}$, 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.

3.3.6 Backfitting

No contexto dos modelos aditivos, quando mais de uma covariável está disponível para predizer a resposta, frequentemente utiliza-se o algoritmo retroajuste (HASTIE & TIBSHIRANI, 1987; BUJA et. al., 1989; HASTIE & TIBSHIRANI, 1990), para estimar cada função suave em um cenário não paramétrico, além do intercepto. A ideia do geral do algoritmo pode ser dado pelos seguintes passos:

1. Adotam-se os valores iniciais $\beta_0^{(0)} = \sum_{i=1}^{n}$ e $s_1^0(.) = s_2^0(.) = ... = s_p^0(.) = 0$;
2. Aplica-se um ciclo retroajuste, ou seja, para cada $j_Y = 1, 2, ..., p$, as funções $s_{jY}(.)$ são atualizadas, suavizando $y - \beta_0 - \sum_{jj \neq jY} s_{jj}^{(0)} | (z_{jj})$ por meio de algum suavizador do gráfico de dispersão, o que resulta em novas funções suaves $s_1^1(.), s_2^1(.), ..., s_p^1(.)$.

Para acelerar a convergência, as funções suaves atualizadas podem ser utilizadas, por exemplo, $s_1^1(.)$ ao invés de $s_1^0(.)$ no cálculo de $s_2^1(.)$;

3. Repetem-se os passos 1 e 2 até que se obtenha a convergência.

Essa ideia é genérica, já que os detalhes diferem dependendo da técnica de suavização usada e do contexto no qual o algoritmo será utilizado.

3.4 Seleção de Modelos - Enfoque de Predição

Inserir aqui a parte do livro do Rafael sobre seleção de modelos, Data splitting, Validação cruzada.. a partir da pagina 12 do livro do Rafael - Resumir o conteúdo adequando a notação que você usou acima

4 Resultados e Discussão

4.1 Estudo de simulação

Neste momento será utilizada a simulação de dados, para gerar situações nas quais possam ser aplicadas algumas das técnicas estudadas, analisando assim suas respectivas performances. A partir disso, iremos avaliar sob estes dados, aspectos visuais dos ajustes de algumas das técnicas vistas até momento. Realizaremos comparações , adotando diferentes tamanhos de janelas (*span*) para cada técnica e então para qual valor de *span*, temos o melhor ajuste. Em um segundo momento, classificaremos para qual técnica obtemos o melhor ajuste. Para esta etapa iremos utilizar as técnicas: Suavizadores com Kernel, *LOWESS* (Suavizador em diagrama de dispersão com pesos locais) e *Splines* de Regressão. Por fim, compararemos as performances dos ajustes do modelo linear e modelo aditivo, em relação a qualidade da predição. Para esta finalidade, utilizaremos as métricas apresentadas na Seção 3.4.

4.1.1 Cenário 1

Foram geradas 200 observações, sendo X uma sequência de 0 a 50 e Y definido pela função

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

onde ϵ é um termo aleatório. Para ter uma ideia da variabilidade dos dados, foi calculado o coeficiente de variação, obtendo-se que $CV_x = 0.5817$ e $CV_y = 0.3645$. Na Figura 1 temos o comportamento dos dados juntamente com a curva real.

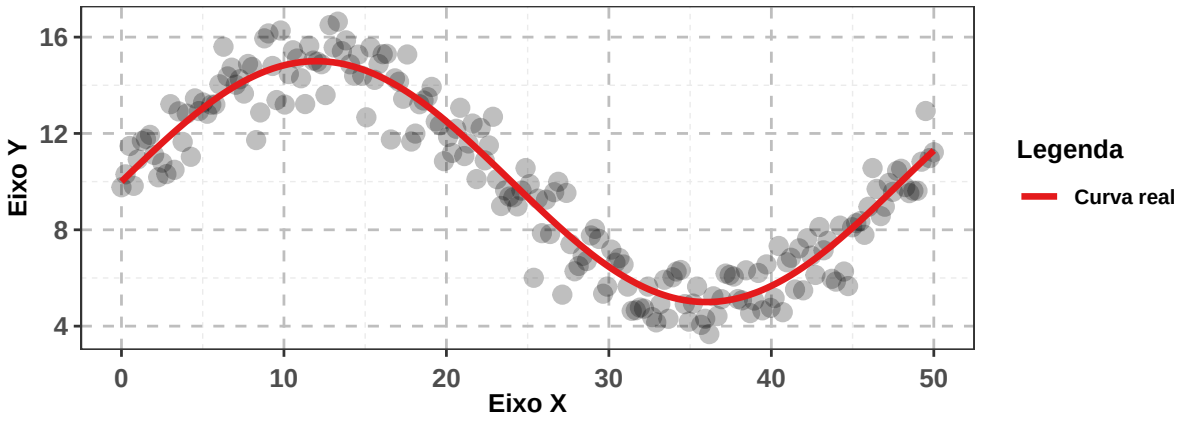

Figura 1: Gráfico de dispersão dos dados gerados e curva real

Bin Smoother

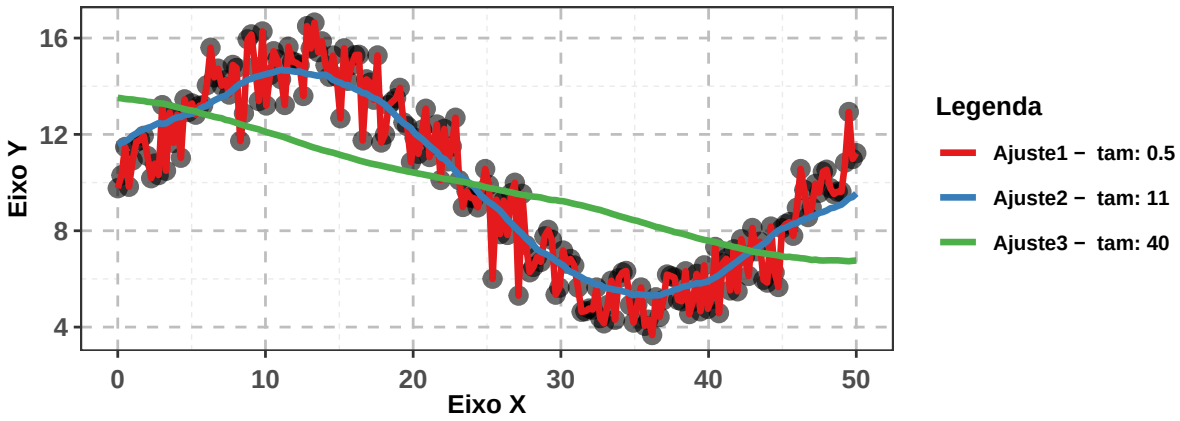

Figura 2: Alguns ajustes utilizando a técnica Bin Smoother

Para o bin, tem-se o tamanho da vizinhança, ou seja, quantos pontos são considerados para fazer o ajuste em cada região, sendo indicado no gráfico por *tam*. Na Figura 2, observa-se que com o tamanho da vizinhança menor (0.5) a técnica interpola os dados, já com o valor maior nota-se que o suavizador tende a uma reta. Com o valor 11 tem-se um equilíbrio da curva, que capta a tendência dos dados. Logo, dentre os valores observados, pode-se concluir que nesse cenário os valores próximos de 11 tendem a apresentar um resultado melhor.

Loess

Nesse caso tem-se o valor do *span*, ou seja, a proporção de observações em cada vizinhança. Na Figura 3, com $span = 0.02$ é possível observar que a técnica fica bastante irregular, enquanto que com $span = 0.2$ a curva já fica mais suave e reflete bem o comportamento das observações. Nesse cenário, a técnica fica suave demais (tendendo a uma reta) quando o valor do *span* é igual a 1. Dessa forma, nota-se que o resultado é mais

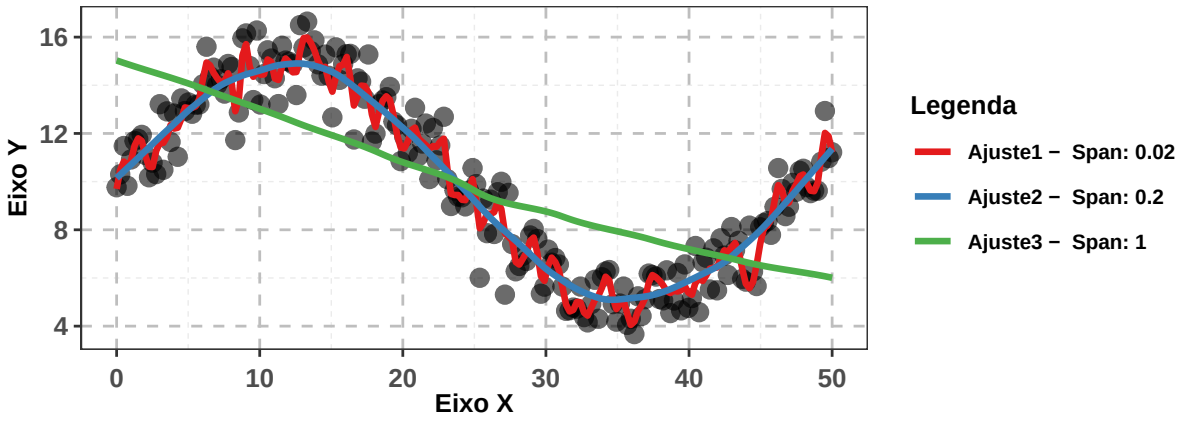

Figura 3: Alguns ajustes utilizando a ténica LOESS

satisfatório para valores de *span* próximos de 0.2.

Kernel

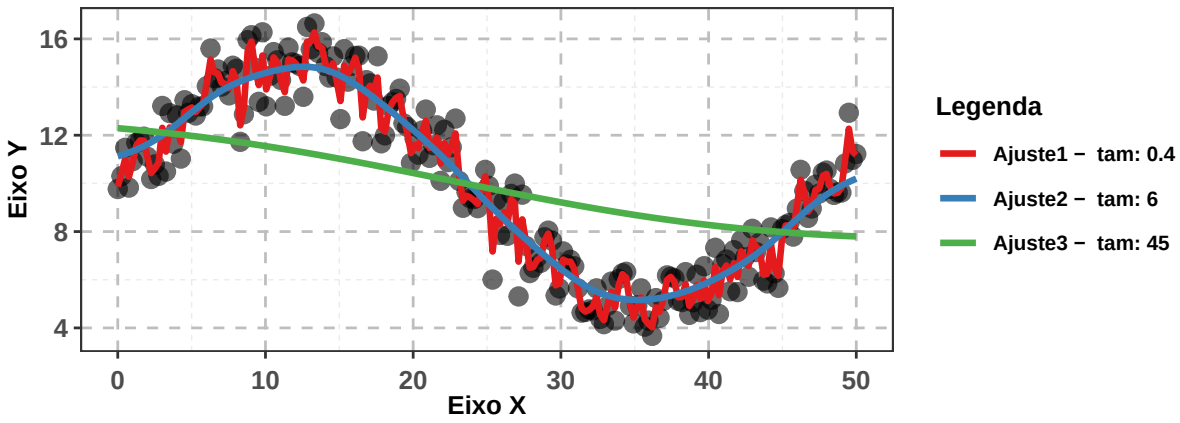

Figura 4: Alguns ajustes utilizando os Suavizadores com Kernel

Nesse método tem-se a largura da região (*bandwidth*), indicado no gráfico por *tam*. Observando a Figura 4, como nas demais comparações até o momento, a técnica Kernel se comporta de maneira semelhante em relação à dimensão da largura nesse cenário: o menor valor resulta em uma curva extremamente irregular, enquanto o maior valor faz a curva ficar muito suave, tendendo a uma reta. Então dentro dos valores observados, percebe-se que valores próximos de 6 podem ser uma escolha interessante, resultando em uma curva suave que acompanha os dados.

Splines de regressão

Através da Figura 5, nota-se que o *spline* de regressão de grau um é extremamente desapropriado, pois a descontinuidade é muito grande. Já o *spline* cúbico segue a tendência dos pontos de uma forma suave, não sendo possível notar qualquer descontinuidade da curva nos nós. Logo, nesse cenário, o *spline* cúbico teve um resultado visualmente melhor.

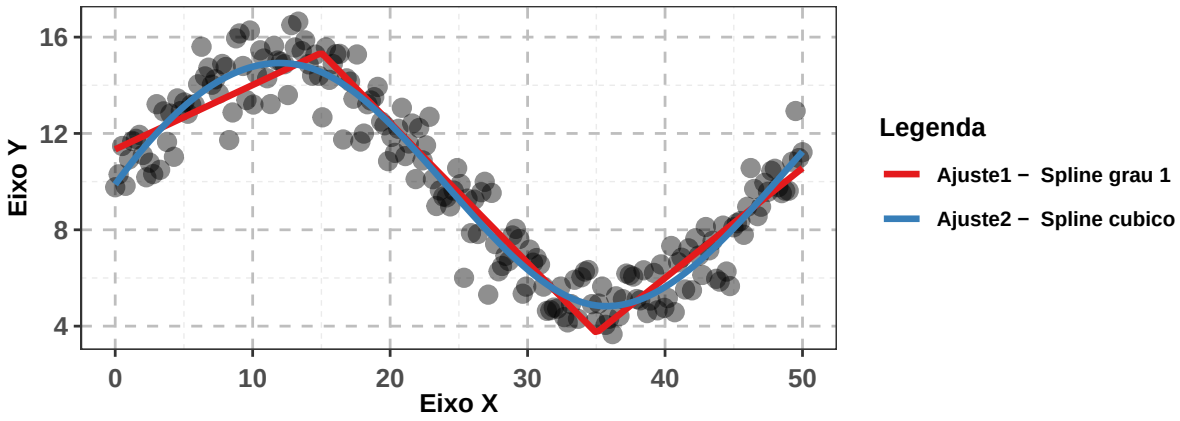

Figura 5: Alguns ajustes utilizando a técnica Spliness de Regressão

Vale ressaltar que em todos os métodos vistos é possível notar a relação de troca entre a variância e o viés. Para valores menores de *span* e largura/tamanho da vizinhança os pontos são interpolados (ou praticamente interpolados), o que resulta em uma curva com grande variância, porém viés pequeno, já que a curva está muito próxima de todas as observações. Por outro lado, quando esses valores aumentam, tem-se a situação inversa: a variância em relação à curva diminui, mas o viés aumenta.

Comparação entre os metódos

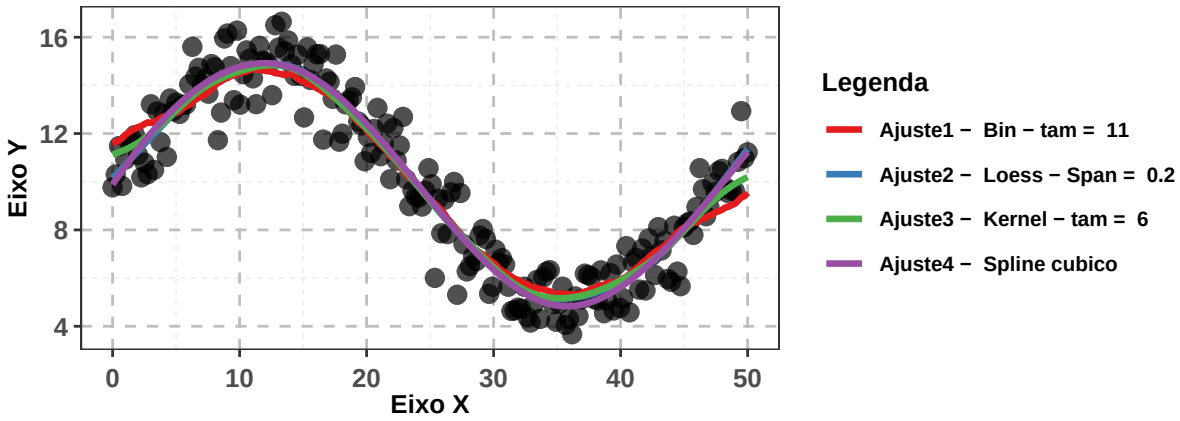

Figura 6: Comparação dos ajustes entre os métodos

Pela figura 6, percebemos que o ajuste bin é extremamente irregular, sendo possível ver a descontinuidade da curva. A técnica kernel não tem um comportamento muito bom nas extremidades. É possivel notar que os ajustes das técnicas *loess* e *spline* cúbico ficaram muito próximas, se adequando muito bem aos dados, mas o *spline* ficou melhor, pois está captando melhor o comportamento dos dados. Portanto, para os dados gerados e dentre os métodos vistos, o *spline* cúbico mostrou um resultado mais satisfatório.

É importante notar que, para a escolha dos valores de *span* ou tamanho/largura da

vizinhança e do método mais adequado para esse cenário, a análise feita foi estritamente gráfica, ou seja, visual. Existem medidas numéricas adequadas para fazer essa análise.

4.1.2 Cenário 2

Para este cenário, foram geradas 201 observações, sendo x uma sequência de 0 à 2 com intervalos de 0.01. Ainda, temos que $y = f(x) + e$, com $f(x) \sim Gamma(6, 10)$ e e $N(0, 0.25)$. O gráfico de dispersão para estes dados pode ser verificado na Figura 7, onde podemos observar o seu comportamento.

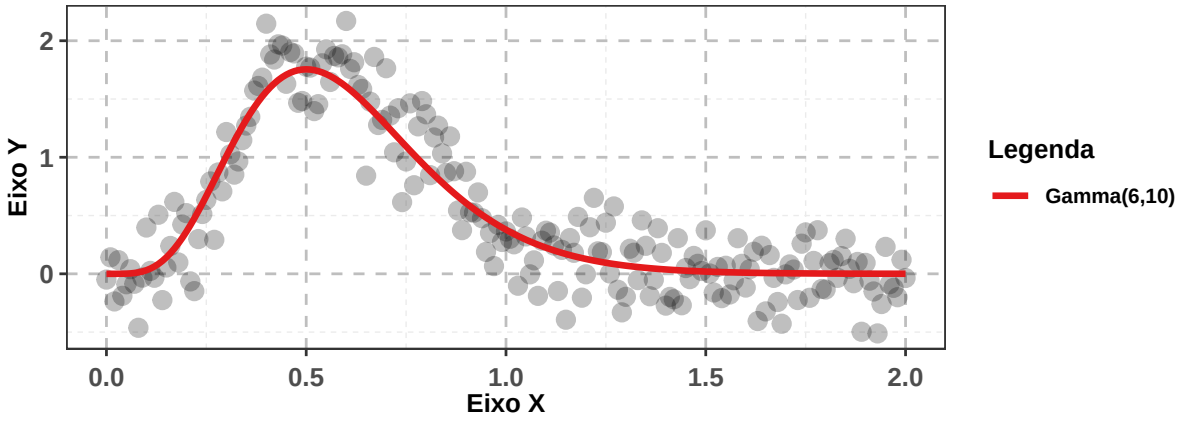

Figura 7: Gráfico de dispersão dos dados gerados para o estudo de simulação

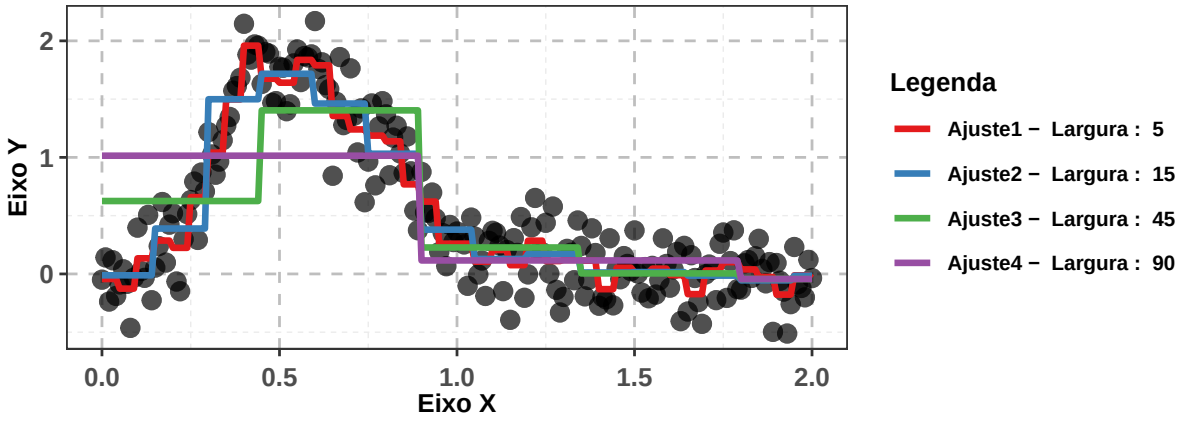

Figura 8: Suavizador bin para diferentes larguras de janela

O suavizador Bin (Figura 8) realiza um ajuste dividindo os dados em k conjuntos, e para cada conjunto disjunto realiza-se a média. Podemos observar que este suavizador tem uma forma altamente irregular, onde basicamente identificamos uma tendência de aumento ou decaimento brusco para as estimativas. Observamos ainda, quando adotamos valores

pequenos para a largura de cada subconjunto, possuímos vários intervalos disjuntos de estimação, sendo assim várias estimativas são calculadas nos dando uma curva irregular no formato escada. Conforme aumentamos a largura de cada subconjunto consequentemente teremos menos intervalos de estimação, e quanto maior for este valor, as nossas estimativas tenderão a uma constante. A média móvel (Figura 9) é uma técnica de suavização onde fixado um ponto x_0 realiza-se a média para k ponto próximos a x_0, repetindo este processo até que todos os pontos possíveis em X tenha uma estimativa. Assim como o suavizador Bin, a média móvel, nos fornece um ajuste irregular.

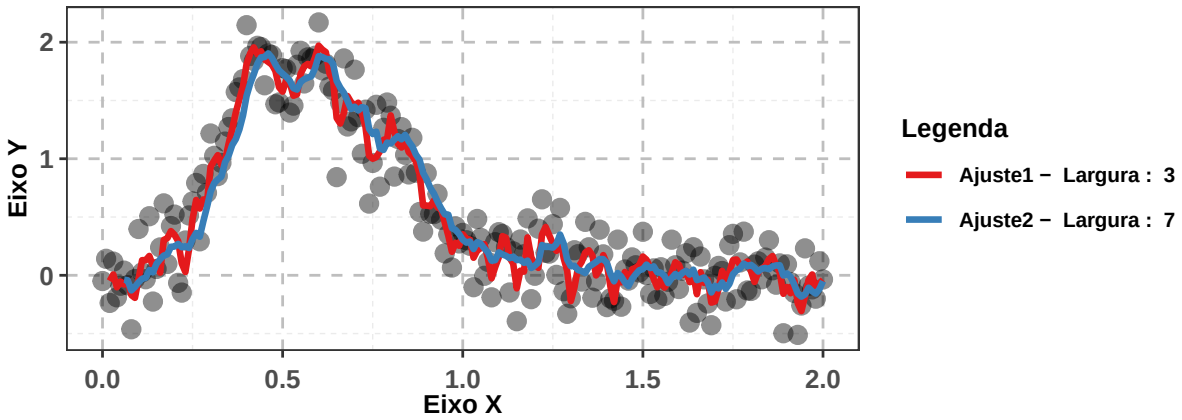

Figura 9: Média Móvel para diferentes larguras de janela

A linha móvel (Figura 10) segue basicamente o mesmo principio da média móvel, mas ao invés da média, ajusta-se um regressão linear com k vizinhos próximos a x_0 e obtem-se a estimativa para este ponto. Como podemos observar quando escolhemos uma proporção de pontos baixa (0.05,0.1), temos um ajuste mais suavizado quando comparado com as técnicas anteriores, porém visualizamos irregularidade em sua forma. Quando aumentamos a proporção de pontos para o ajuste, teremos uma forma altamente suavizada, e quando adotamos um número grande o suficiente de pontos próximos, o ajuste tende a ser uma reta.

Para a técnica *Lowess*, o parâmetro *span* controla a proporção de pontos que influência a suavização para cada valor. Observamos então que para valores de *span* pequenos, ou seja, para os ajustes que utilizam um proporção de pontos pequena (Conforme Figura 11, Ajuste1 - Span : 0.05), nos fornece formas mais irregulares. Logo, quando adotamos poucos pontos, o ajuste tende a interpolar uma quantidade maior de observações. A medida que aumentamos este valor percebemos que o ajuste fica mais suave. Porém quando este valor se torna grande, o ajuste se torna demasiadamente suavizado, e tende a forma linear. Para elegermos o melhor ajuste iremos adotar o critério de troca entre variância e viés que consiga captar a maior quantidade de observações possíveis. Logo para vizinhanças grandes por adotarmos valores altos para o *span*, estas estimativas tendem a ter variância pequena, porém um viés alto. Ainda estes ajustes não captam de forma

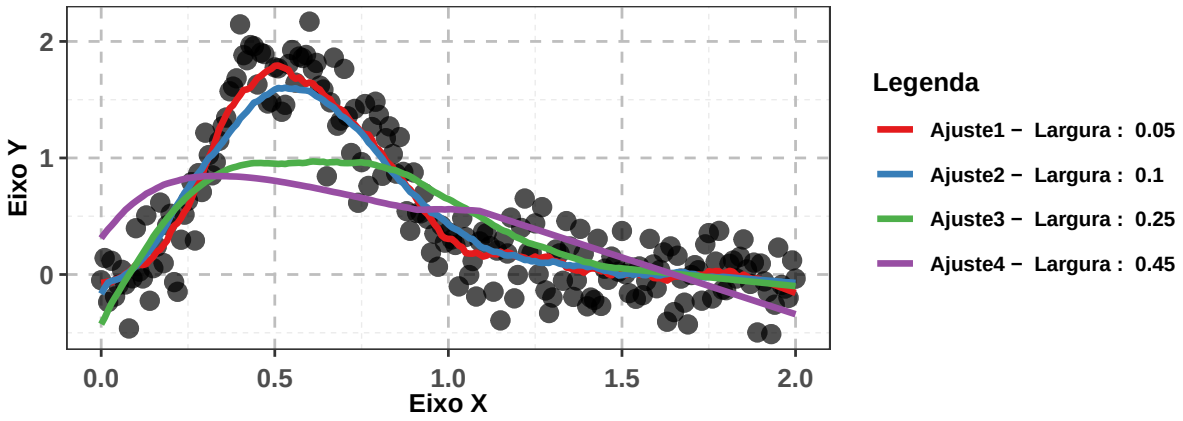

Figura 10: Linha Móvel para diferentes intervalos

adequada a tendência produzida pelos dados simulados. Os ajustes das curvas 4 e 5 (linha roxa e laranja respectivamente, Figura 11), não nos fornece uma suavização que capta de forma adequada a tendência dos dados. Quando observamos atentamente a Figura 11, concluimos que ao adotarmos valores para o *span* que estejam entre 0.15 e 0.25 teremos para estes dados um ajuste que capte a tendência dos dados levando em consideração um forma que tenha uma relação de troca entre viés e variância equilibrada.

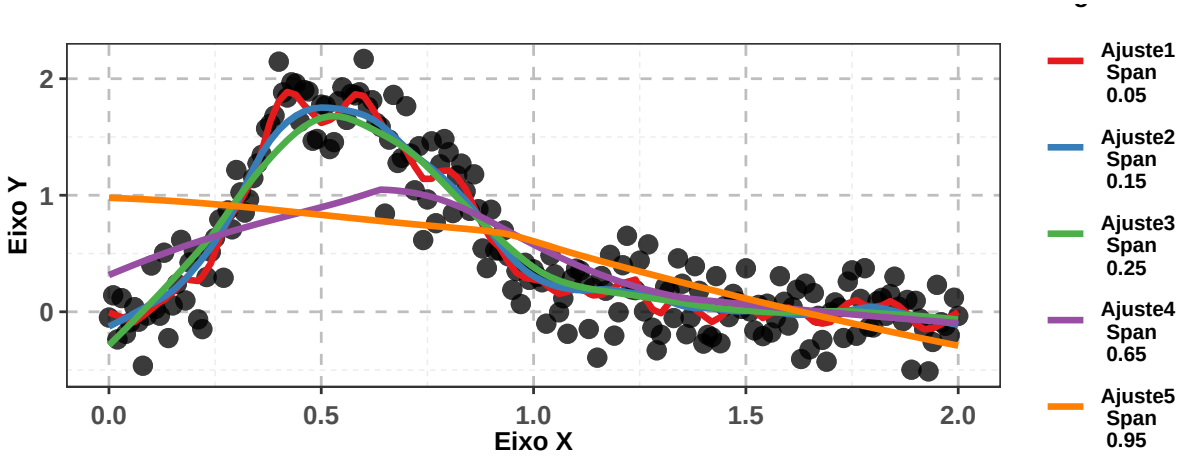

Figura 11: Ajustes Lowess para diferentes valores de span

A Figura 12 mostra alguns ajustes considerando os suavizadores com kernel (kernel gaussiano), considerando valores distintos para a largura de banda (*bandwidth*). De forma análoga a técnica *Lowess*, quando obtemos os ajustes para valores pequenos do parâmetro largura da banda (*bandwidth*), temos uma maior irregularidade em sua forma (figura 14, Ajuste1). A medida que valores maiores para a largura de banda são adotados a curva se torna altamente suavizada, de modo que o ajuste não capta a tendência dos dados (Vide figura 14). Portanto ao avaliarmos podemos dizer o melhor ajuste se encontra para valores da largura de banda de 0.05 à 0.35.

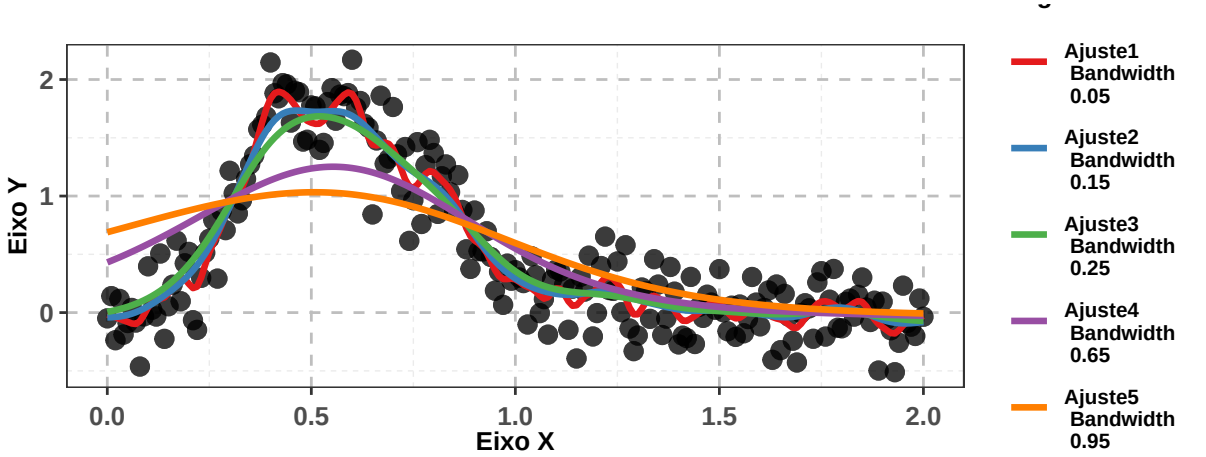

Figura 12: Ajustes com Suavizadores Kernel para diferentes valores de largura de banda

Para análise do *Spline* de Regressão iremos comparar os ajustes quando adotamos um polinômio de grau 1 e logo após utilizaremos polinômios cúbicos. Poderemos ainda verificar as consequências ao se adotar diferentes quantidades de nós para os dois casos. Utilizando o polinômio de grau 1 (ajuste linear), temos que para cada nó escolhido, este tipo de *spline*, ajusta uma reta. Como podemos observar na Figura 13 (gráfico A), ao escolhermos valores extremos para esta técnica, quando adotamos uma quantidade pequena de nós, a tendência dos dados não é muito bem explicada por poucas retas, e em contrapartida, quando observamos para muitos nós, temos uma forma irregular.

Como podemos ver os ajuste mais adequados para representar a tendência destes dados estaria próximos de 7 e 8 nós. Os ajustes anteriores está considerando uma distribuição equidistante para diferentes valores de quantis para os dados simulados. Porém para estas técnicas além da quantidade nós, a localização de cada nó, pode nos levar a uma representação diferente e mais adequada para os dados. A figura 15 (gráfico B), nos mostra um ajuste realizado escolhendo a posição utilizando dois nós.

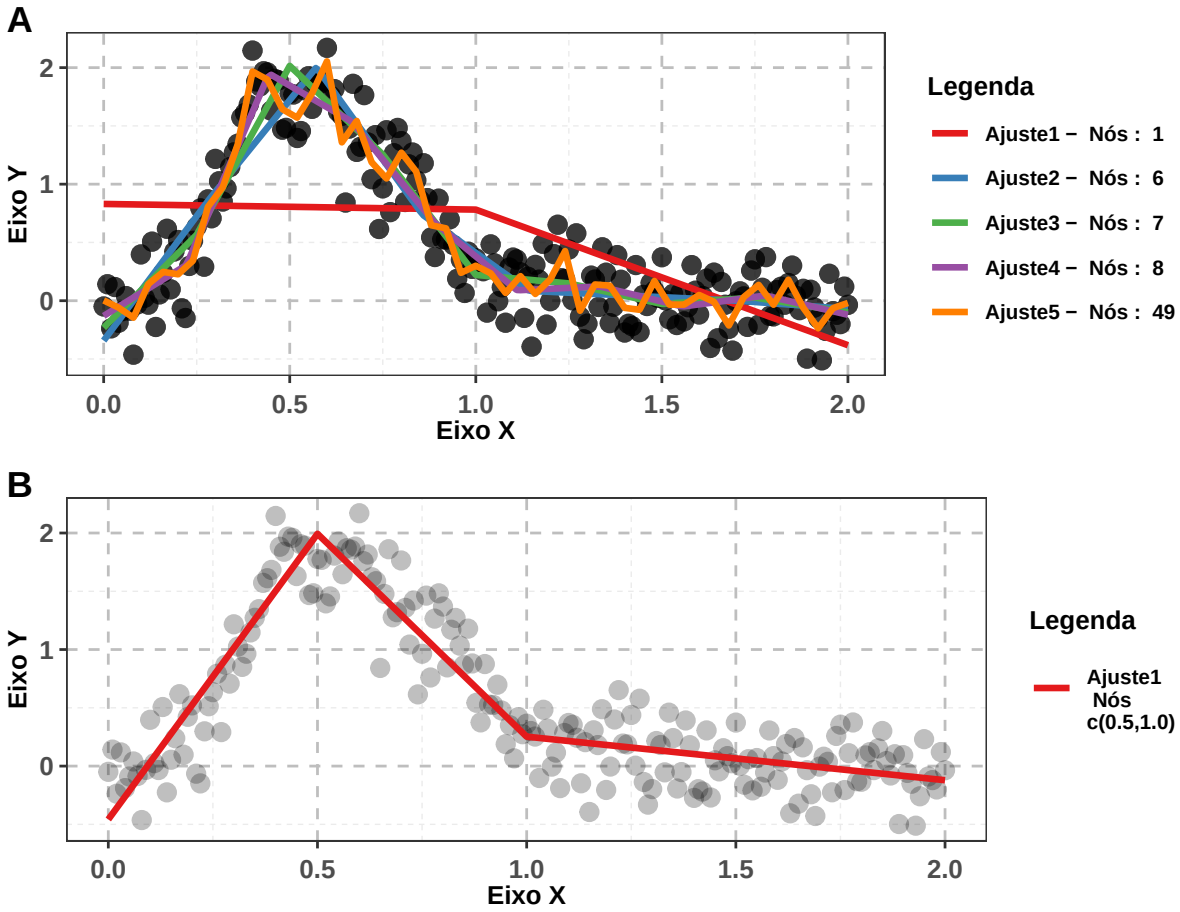

Figura 13: Splines de Regressão com polinômio de grau 1 para a escolha de nós distintos

O ajuste selecionando apenas dois nós nos quantís 0.5 e 1.0 (gráfico B) nos fornece um ajuste relativamente mais adequado para representar os dados. Conseguimos melhorar o ajuste dos dados utilizando um polinômio de grau maior.

Na Figura 14 temos alguns ajustes levando em consideração os *Splines* de Regressão Cúbico. Sendo assim, quando avaliamos as curvas geradas, observamos que o Ajuste1 (linha vermelha), não representa adequadamente os dados. O Ajuste4 (curva roxa), observamos nos extremos da curva uma leve ondulação. Logo se observamos a curva do Ajuste3 (curva verde), percebemos que esta curva se ajusta de forma suave, logo podemos concluir que para estes dados, ao utilizarmos os *Splines* de Regressão Cúbico, o melhor ajuste está quando utilizamos valores próximos de 7 nós.

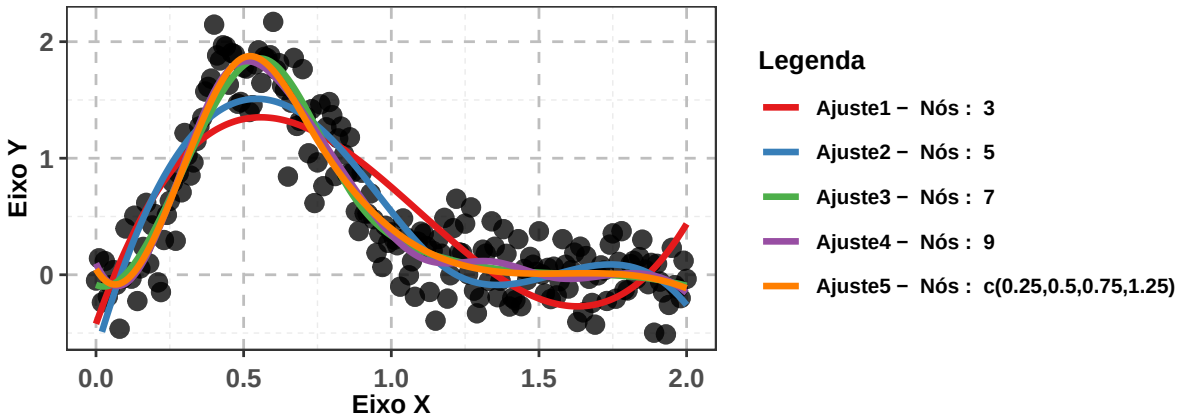

Figura 14: Ajustes com Splines de regressão para diferentes valores de span

A Figura 15 nos mostra, um comparativo entre a técnicas vista até agora.Em vermelho temos o ajuste considerando a a técnica suavizadora *Loess*, com *span* igual a 0.25. Em azul temos a curva ajustada com o suavizador Kernel (kernel gaussiano), considerando uma largura de banda de 0.22. Os dois ajustes são parecidos, e captam de forma adequada a tendência dos dados. Porém a curva em azul (*kernel smoother*), aparenta ter uma pequena oscilação no começo da curva.

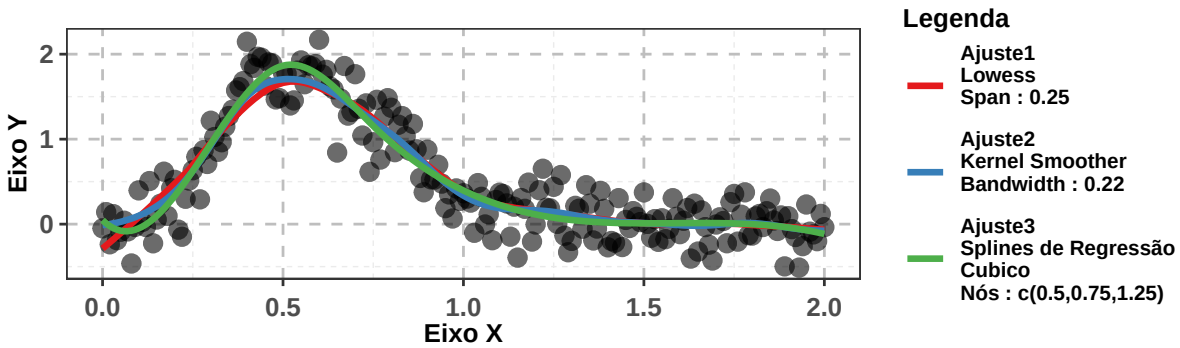

Figura 15: Ajustes comparativos para as técnicas Lowess, Kernel Smoother e Splines de Regressão Cubico

Ao observarmos o Ajuste3 (*Splines* de Regressão Cúbica), utilizando os seguintes nós localizados em $c(0, 0.5, 0.75, 1.25)$, consegue captar de forma mais adequada a tendência em torno do ponto $x = 0.5$. Logo podemos concluir que dentre as técnicas abordadas até o momento, para estes dados simulados, os *splines* de regressão cúbico aparentam ajustar a tendência dos dados de forma mais suave levando em consideração um equilibrio entre viés e variância.

Consegue recuperar os ajustes? Se sim, aplicar as métricas de qualidade da predição e concluir qual método tem melhor poder preditivo.

4.1.3 Cenário 3

Iremos agora incluir os modelos aditivos para dados simulados considerando duas covariáveis. Foram geradas 250 observações, $x_1 \sim N(5, 20)$ e $x_2 \sim N(75, 15)$ e os valores das funções reais para $f_1(x_1)$ e $f_2(x_2)$ foram gerados de funções quadráticas, sendo elas:

$f_1(x_1) = \frac{25(x_1-20)^2-(x_1-20)}{20}$ e $f_2(x_2) = \frac{-25(x_2-100)^2-(x_2-100)}{8} + 2000$

Esta sendo considerado um valor aleatório para o termo α do modelo, os erros ε são normais com média 0 e variância constante ($\varepsilon \sim N(0, 200)$). Os valores para y foram gerados da seguinte forma:

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon$$

A Figura 16 mostra o comportamento dos resíduos parciais de $f_1(x_1)$ e $f_2(x_2)$ e em vermelho suas respectivas curvas reais.

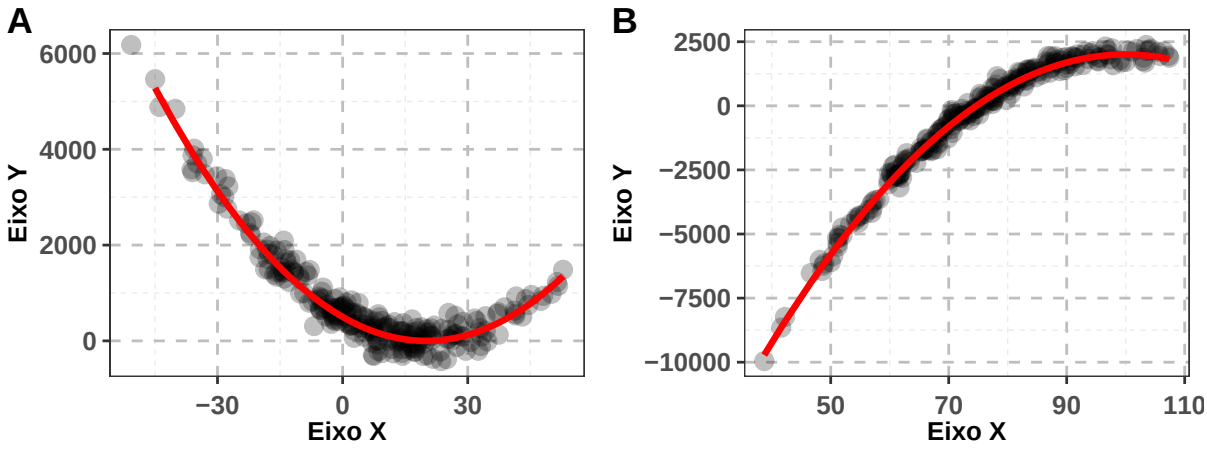

Figura 16: Gráfico de dispersão contendo as curvas reais simuladas, A - $f_1(x_1)$ e B - $f_2(x_2)$.

Iremos aplicar o algoritmo de retroajuste (*backfitting*), para conjunto de dados simulados (y, x_1, x_2) com o intuito de captar a mesma tendência quando observamos a Figura 16, comparando dois ajustes utilizando as técnicas *Loess* e *Splines* de Regressão Cúbico. Aplicando o algoritmo nos dados e obtendo as estimativas e resíduos parciais para

cada função $f_1(x_1)$ e $f_2(x_2)$. Na Figura 17 (gráfico A) temos os ajuste e resíduos parciais para $f_1(x_1)$, onde temos dois ajustes, o Ajuste1 (técnica suavizadora *Loess*) foi obtido utilizando um valor para o parâmetro $span = 0.15$. Para o Ajuste2 (técnica suavizadora *splines* de regressão cúbico) de primeiro momento foi tentado utilizar os nós distribuídos de forma equidistante, porém conseguimos um melhor ajuste utilizando um nó posicionado no quantil $x = -25$. Portanto percebemos que utilizando os parâmetros em questão para estes dados simulados, utilizando o algoritmos de reatroasjute, conseguimos de forma satisfatória recuperar a forma real para a função $f_1(x_1)$. Quando observamos os ajustes vemos as curvas dos dois ajustes são bem próximas, os pontos dos resíduos parciais se encontram dispostos ao longo da curva indicando que os ajustes sejam adequados para representar estes dados.

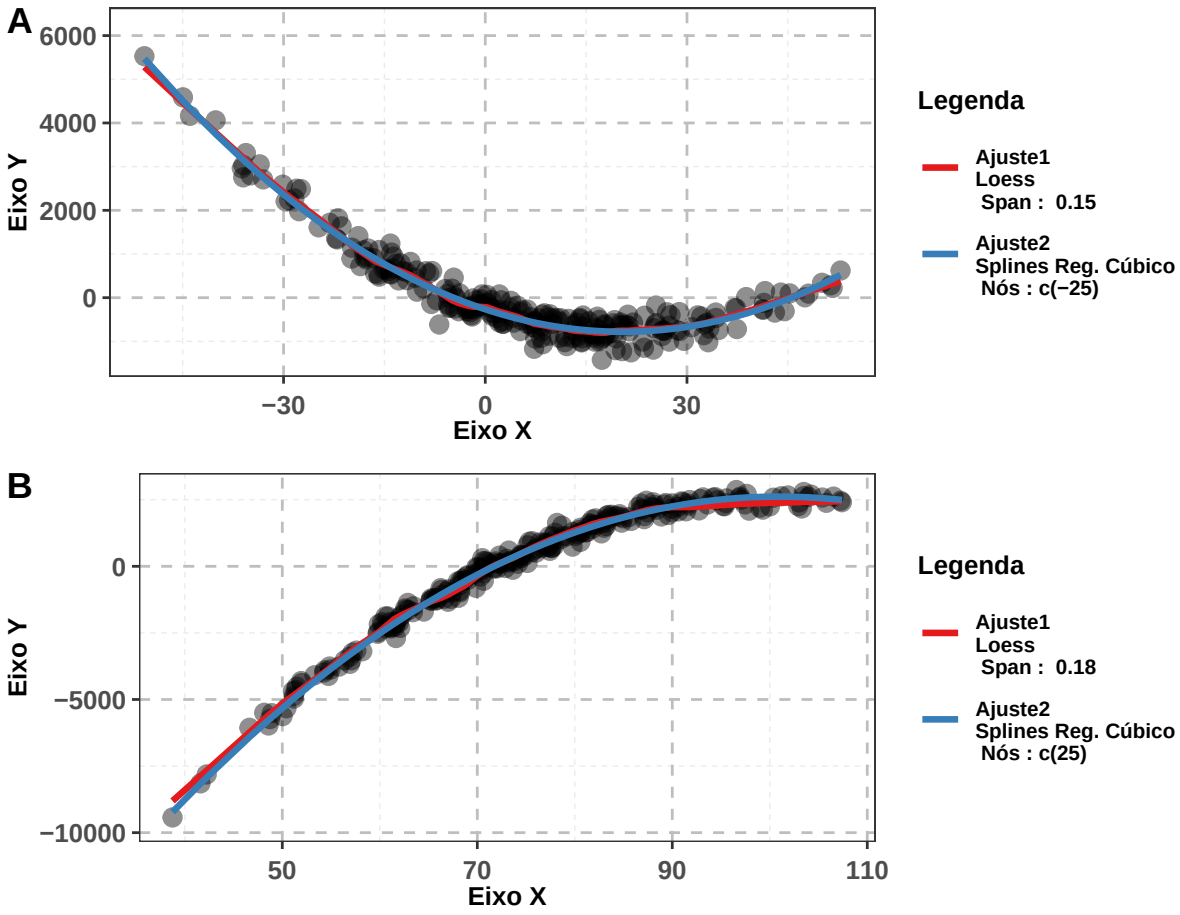

Figura 17: Ajuste e resíduos parciais para $f(x_1)$ e $f(x_2)$

Quando observamos a Figura 17 (gráfico B), para o Ajuste1 (*Loess*), utilizamos um $span = 0.18$, e para Ajuste2, utilizamos um nó posicionado no quantil $x = 25$. Quando avaliamos as curvas ajustadas para os resíduos parciais de $f_2(x_2)$ percebemos que os pontos residuais se encontram dispostos aleatoriamente em torno das curvas, logo conseguimos captar utilizando os parâmetros citados anteriormente,para estes dados, de forma adequada a tendência de $f_2(x_2)$.

Embora todas as curvas em ambos ajustes ($f_1(x_1)$ e $f_2(x_2)$), sejam semelhantes, destacamos que, quando observamos atentamente o Ajuste2 (*Splines* de Regressão Cúbico), visualmente para a simulação em questão, se adequa e capta melhor a tendência dos dados para ambas as funções. Até o momento estamos utilizando métodos visuais, utilizando gráficos de dispersão para concluirmos quais ajustes aparentam se adequar e captar melhor a tendência dos dados. Relembramos que para as técnicas vistas até o momento a escolha do melhor parâmetro *span* ou nós, pode ser uma tarefa dificil.

4.2 Aplicações

4.2.1 Aplicação 1

O banco de dados é constituído por 19 observações relacionadas ao consumo de energia em hotéis de luxo na província de Hainan, China. Além da variável consumo de energia (*enrgcons*), medida em kilowatt-horas, há outras cinco variáveis: área em quilômetros quadrados (*area*), idade em anos (*age*), número de quartos de hóspedes (*numrooms*), taxa de ocupação em porcentagem (*occrate*) e número efetivo de quartos de hóspedes (*effrooms*), esse último dado como uma função de outras duas covariáveis (nº de quartos * taxa de ocupação/100).

O objetivo é encontrar o modelo de regressão com as variáveis que melhor explicam o consumo de energia. Na Figura 18 tem-se os gráficos de dispersão de cada covariável versus a variável resposta equanto que na Tabela 2 é apresentada a matriz de correlação amostral dos dados.

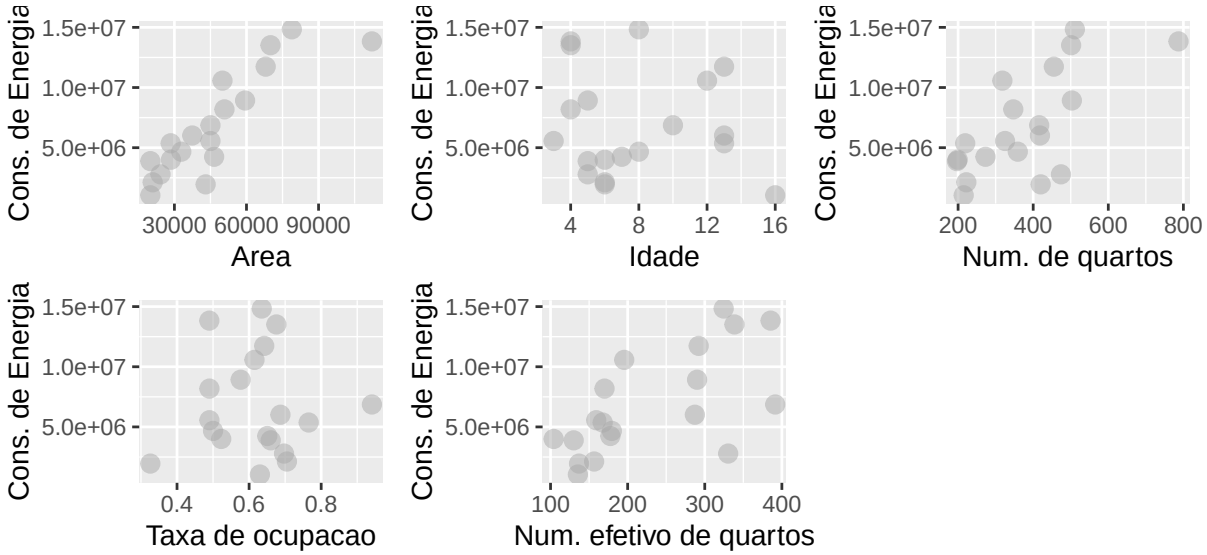

Figura 18: Gráficos de dispersão da variável resposta Consumo de energia versus cada uma das preditoras

Tabela 2: Matriz de correlação

	enrgcons	area	age	numrooms	occrate	effrooms
enrgcons	1.00	0.87	-0.09	0.68	0.031	0.65
area	0.87	1.00	-0.23	0.85	-0.21	0.65
age	-0.09	-0.23	1.00	-0.27	0.40	-0.045
numrooms	0.68	0.85	-0.27	1.00	-0.15	0.82
occrate	0.03	-0.21	0.40	-0.15	1.00	0.40
effrooms	0.65	0.65	-0.04	0.82	0.40	1.00

A partir da Tabela 2 percebe-se que o consumo de energia (*enrgcons*) tem correlação considerável com *area* (0.879) , *numrooms* (0.685) e *effrooms* (0.657). As covariáveis *area* e *numrooms* têm uma correlação relativamente alta (0.853), bem como as covariáveis *numrooms* e *effrooms* (0.826). Essa correlação entre as covariáveis pode ser vista como uma indicação de multicolinearidade. Analisando o Fator de Inflação da Variância (VIF) das variáveis explicativas é possível perceber o valor elevado para a variável *effroms* (47.45), o que faz muito sentido já que ela é função de outras covariáveis, logo essa pode ter sido a razão desse problema de multicolinearidade. Dessa forma, essa covariável foi retirada e, calculando o VIF novamente, todos os valores ficaram satisfatórios.

Logo após foi realizada a seleção de variáveis através do método *backward* e vale ressaltar que esse método leva em considereção os pvalores, logo foi preciso validar a suposição de normalidade dos dados, para isso foi analisada a normalidade dos resíduos para o modelo através dos testes de Shapiro-Wilk e Kolmogorov-Smirnov. A um nível de significância $\alpha = 5\%$, em ambos os testes não rejeitou-se a hipótese de que os resíduos provêm de uma distribuição Normal, logo há evidências da normalidade dos resíduos desse modelo.

Voltando ao método *backward*, o submodelo final ficou com as covariáveis área e occrate (taxa de ocupação), sendo que esse modelo realmente é significativo, já que através do teste F parcial entre esse e o modelo completo ficou claro que as covariáveis age e numrooms não são significativas.

Pode ser observado na Figura 19, forte relação linear entre os valores ajustados e a covariável area, Ja entre os valores ajustados e a covariável Taxa de ocupação (occrate) a relação linear não é tão acentuada. Ressalta-se que o intercepto também é significativo para o modelo.

Assim, o modelo linear final ajustado é dado por

$$\hat{y} = -6.335e^{06} + 2.070e^{02}\text{area} + 6.348e^{06}\text{occrate}$$

Os parâmetros estimados do modelo podem ser interpretados da seguinte forma:

- $\hat{\beta}_0$: nesse caso o intercepto não tem interpretação prática, já que nos dados não há

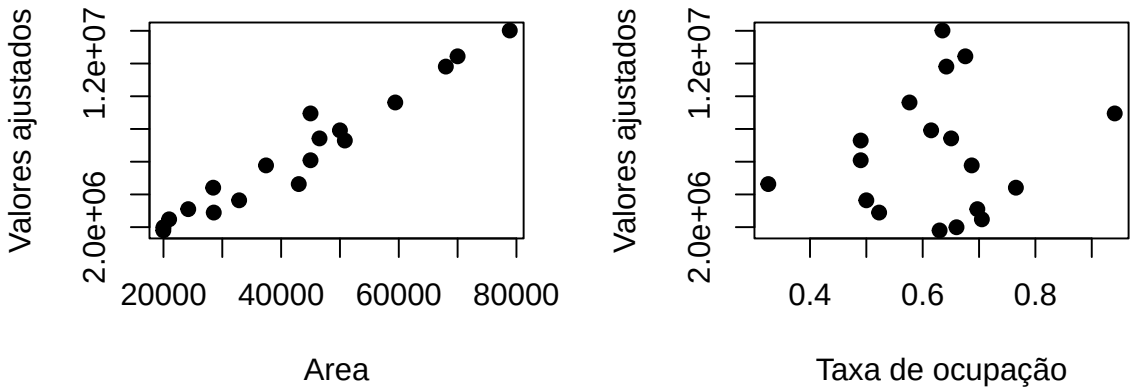

Figura 19: Gráficos de dispersão dos valores ajustados versus area e taxa de ocupação, respectivamente

uma observação na qual $x_1 = x_2 = 0$;

- $\hat{\beta}_1$: quando a taxa de ocupação (*occrate*) é constante, a cada mudança unitária no metro quadrado do hotel, é esperado que o consumo de energia aumente em $2.070e^{02}$ kilowatt-hora;

- $\hat{\beta}_2$: quando a área é constante, a cada mudança unitária na taxa de ocupação do hotel, é esperado que o consumo de energia aumente em $6.348e^{06}$ kilowatt-hora.

Nesse momento, vamos ajustar o modelo aditivo ao mesmo conjunto de dados. Foram consideradas as 5 variáveis independentes do problema inicial, sendo que as mesmas foram definidas como $X_1 : area$, $X_2 : age$, $X_3 : numrooms$, $X_4 : effrooms$ e $X_5 : occrate$. A partir disso, o modelo aditivo será da forma

$$y = \alpha + f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + f_5(X_5) + \epsilon$$

Neste caso, foi utilizado o algoritmo de retroajuste para estimar as funções do componente sistemático através de duas técnicas de suavização vistas anteriormente: Kernel e *Loess*, obtendo-se gráfico a seguir:

O valor de $\hat{\alpha}$ para o modelo foi de 6848632 e a Figura 20 é apresentado a curva ajustada pelo método *Loess* em vermelho e pelo método Kernel em verde. Pode-se perceber visualmente que o comportamento dos pontos, que são os resíduos parciais de cada função, não é explicado da mesma forma pelos dois métodos, porém ambos captam bem a tendência dos dados. Nota-se que as curvas em vermelho são mais suaves, enquanto que as curvas em verde têm uma oscilação maior. Portanto, para as técnicas apresentadas para os dados em

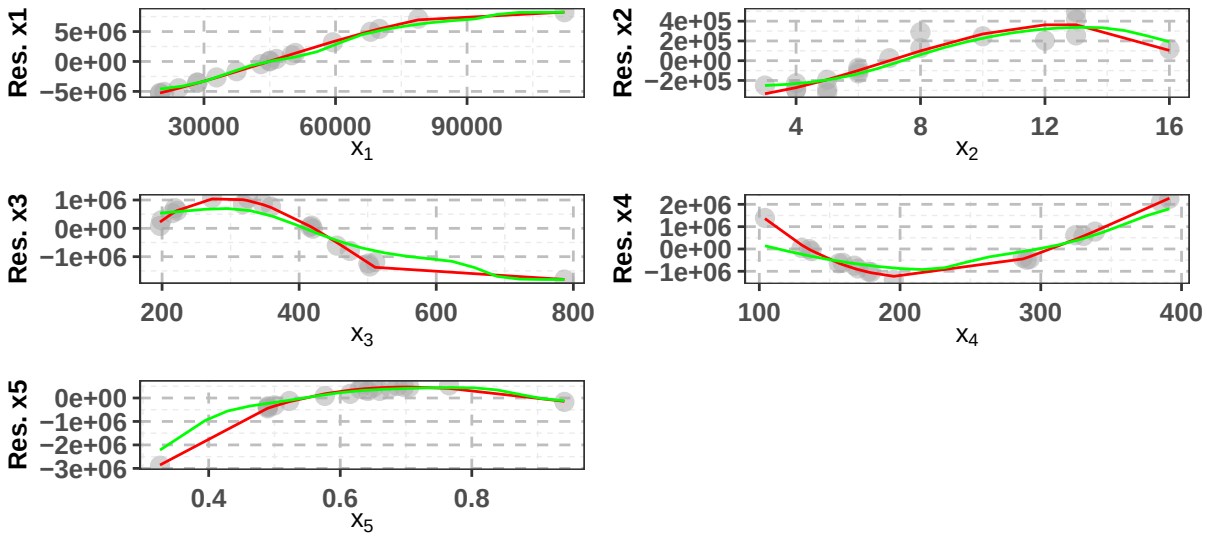

Figura 20: Curvas ajustadas

questão, o ajuste através do método *Loess* foi mais interessante, já que as curvas possuem uma variabilidade menor.

Diferente do que ocorreu no ajuste do modelo linear, podemos ver graficamente que todas as covariáveis contribuem de forma significativa para explicar a variável resposta consumo de energia (enrgcons)", já que nenhuma das funções suavizadas tem o comportamento de uma reta constante. Desta forma, tem-se um ajuste mais adequado, se comparado ao ajuste linear.

Inserir as métricas para verificar as predições

4.2.2 Aplicação 2

Os dados em questão são resultados de um experimento relacionado ao rendimento de bombas de poço contendo 54 observações composto pelas variáveis:

- **easting** : coordenadas relacionadas a posição leste do poço;
- **northing** : coordenadas relacionadas a posição norte do poço;
- **aq.resist** : resistividade do aquecedor;
- **aq.thick** : espessura;
- **anisotrophy** : coeficiente de anisotrofia;
- **bh.yield** : rendimento dos testes da bomba de poço ($litros/s$).

Iremos ajustar um modelo de regressão que explique rendimento dos testes da bomba de poço relacionando as três variáveis geologicas que são: resistividade do aquecedor;

espessura; coeficiente de anisotrofia. A Figura 21 mostra os box-plots para cada variável que será utilizada no ajuste.

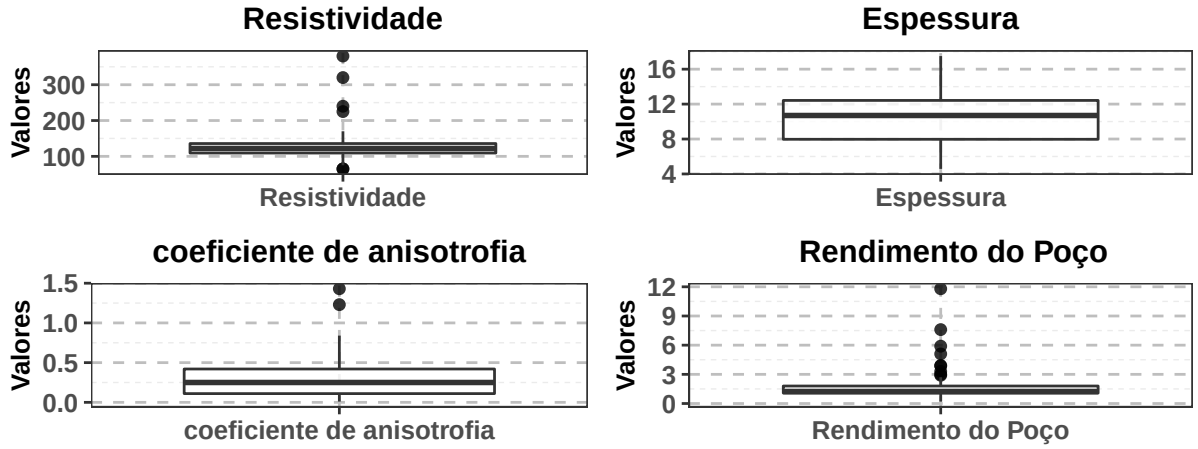

Figura 21: Box plots para as variáveis do conjunto de dados

Ao observar a variável resistividade do aquecedor podemos observar uma variância alta (2844,57), o que pode ser explicado pelos pontos discrepantes que se encontram dispersos acima de 200, enquanto 75% dos valores se encontram abaixo de 135,8. Ainda ao observarmos o rendimento dos poços encontramos pontos discrepantes que se encontram próximos e acima de 3 e que 75% dos dados se encontram abaixo 1,8. Podemos observar na variável espessura uma leve assimetria e pontos outliers na variável coeficiente de anisotrofia.

A Tabela 3 contém a matriz de correlações para as variáveis onde verifica-se que não apresentam forte correlações entre si, sendo a mais alta observada entre a variável resistividade do aquecedor e coeficiente de anisotrofia com um valor de 0,68. O valor mais alto para os fatores de inflação foi de 2,67, portanto não temos indícios de que as variáveis possuam variâncias inflacionadas devido a colinearidade, fazendo com que as estimativas para o modelo sejam ruins.

Tabela 3: Tabela contendo as correlações entre cada variável

	aq.resist	aq.thick	anisotrophy	bh.yield
aq.resist	1.00	0.50	0.69	0.92
aq.thick	0.50	1.00	0.68	0.69
anisotrophy	0.69	0.68	1.00	0.89
bh.yield	0.92	0.69	0.89	1.00

Ajustando o modelo linear, considerando todas as covariáveis, a Tabela 4 contém estimativas para o modelo ajustado completo e os valores das estatística t e F, para o teste de ANOVA, para avaliarmos se ao menos alguma das covariáveis influenciam de forma

significativa para explicar o rendimento das bombas de poço. Verificamos que o p-valor da estatistica F é zero, portanto como um nível de significância de 5%, temos indícios de que ao menos uma das covariáveis influenciam de forma significativa para explicar o redimento das bombas dos poços.

Tabela 4: Estimativas para o modelo ajustado completo

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(**Intercept**)	-2.4934	0.181	-13.7769	0	NA
I(aq.resist)	0.0209	0.0011	19.1402	0	1.9035
I(aq.thick)	0.0724	0.0181	3.9975	2e-04	1.8668
I(anisotrophy)	2.8959	0.2435	11.8907	0	2.6713
	$\mathbf{R^2}$	$\mathbf{R^2_{Adj}}$	$\mathbf{F_0}$	$\mathbf{Pr(> \|F\|)}$	NA
	0.9767	0.9753	699.4868	0	NA

Em seguida fora utilizado uma função obtida da biblioteca "olsrr", chamada "ols_step_best_subset()", que retorna os melhores ajustes de acordo com cada quantidade de covariáveis que o modelo pode obter, ou seja, a função retorna com uma única covariável o modelo obteve o melhor ajuste, em seguida seleciona qual ajuste com 2 covariáveis obtem o melhor ajuste e assim por diante. Podemos ver na tabela 5, quais covariáveis foram consideradas como melhores ajustes:

Tabela 5: 3 Melhores Ajustes

	predictors	rsquare	adjr
1	I(aq.resist)	0,84	0,84
5	I(aq.resist) I(anisotrophy)	0,97	0,97
7	I(aq.resist) I(aq.thick) I(anisotrophy)	0,98	0,98

a

Os três melhores modelos que podemos ajustar e levando em consideração o coeficiente de determinação ajustado R^2_{adj}, o melhor modelo ajustado com uma covariável obtendo assim um $R^2_{adj} = 0,83$, seria com a covariável resistividade do aquecedor (aq.resist). Ainda o melhor ajuste com duas covariáveis ajustado seria com a covariável resistividade e coeficiente de anisotrofia obtendo assim um $R^2_{adj} = 0,96$. E por fim a função sempre retornara o modelo ajustado completo. Porém, ao avaliarmos seus respectivos coeficientes de determinação podemos notar que, o segundo modelo sugerido com apenas duas covariáveis consegue explicar aproximadamente a mesma variabilidade para nossa variável resposta, quando comparamos com o modelo ajustado completo. Levando em consideração o que vimos anteriormente um modelo ajustado com duas covariáveis possuira quase o mesmo efeito explicativo quando comparado com o modelo ajustado completo. Porém

se analisarmos o p-valor para o teste de significância de β_3, nos indicia que este seja significativo para o modelo,ou seja, com um nível de singificância de 5% temos evidências de que β_3 seja diferente de zero, portanto iremos continuar considerando o modelo completo sendo o modelo ajustado dado por:

$$\hat{Y} = -2,493 + 0,021\text{aq.resist} + 0,072\text{aq.thick} + 2,8969\text{anisotrophy}$$

Onde :

- $\hat{Y}$: representa o valor esperado para o redimento da bomba dos poços em $litros/s$;
- X_1 : representa a variável resistividade do aquecedor, e dado que $X_2 = 0$ e $X_3 = 0$, nos indica que a cada aumento em uma unidade em X_1 representa o aumento em 0.021 no valor esperado do rendimento da bomba dos poços;
- X_2 : representa a variável espessura, e dado que $X_1 = 0$ e $X_3 = 0$, nos indica que a cada aumento em uma unidade em X_2, representa o aumento em 0.072 no valor esperado no rendimento da bomba dos poços;
- X_3 : representa a variável coeficiente de anisotrofia, e dado que $X_1 = 0$ e $X_2 = 0$, nos indica que a cada aumento em uma unidade em X_3, representa o aumento em 2.896 no valor esperado do rendimento da bomba dos poços;

Tentando realizar um modelo alternativo, considerando uma transformação box-cox para ajustar um modelo mais representativo dos dados. Primeiramente obtemos o coeficiente λ. Utilizando a função 'boxcox' da biblioteca "MASS", obtemos um valor de λ = 0.8, portanto teremos uma transformação em y da forma :

$$\frac{y^{0.8} - 1}{0.8}$$

Além da transformação em Y podemos testar algumas transformações nas covariáveis para afim de obter uma melhor representatividade, principalmente no intuito de melhorar e atingir o pressuposto de normalidade para o modelo. Fora realizado as seguintes transformações para as covariáveis: $X_1^* = X_1^3$, $X_2^* = X_2^2$ e $X_3^* = X_3^5$.

Portanto considerando as transformação de box-cox realizada na variável resposta, e as transformações nas covariáveis temos o ajuste sendo o seguinte modelo ajustado:

$$\hat{Y} = -1.053 + 0.01\text{aq.thick} - 1.069\text{anisotrophy}$$

Vamos agora considerar o ajuste do modelo aditivo para tentar captar a tendência dos dados referente ao rendimento de bombas de poço. Considerando as variáveis X_1 = aq.resist, X_2 = aq.thick e X_3 = anisotrophy temos o modelo aditivo,

$$y = \alpha + f_1(X_1) + f_2(X_2) + f_3(X_3) + \epsilon$$

Aplicando o método de retroajuste para ajustar a funções por meio das técnicas *Loess*, *Splines* de Regressão Linear e Cúbico, a Figura 22 apresenta as curvas residuais parciais utilizando cada método. Verificamos o gráfico A e B (figura 29), que o ajuste *Loess* capta muito bem a tendência sofrendo baixa influência de pontos mais distantes. Já para as curvas azul e verde (*Splines* de Regressão), percebemos que este sugerem um comportamento quadrático para o gráfico A e uma tendência exponecial para o gráfico B, ambos tentando captar uma maior quantidade de pontos mais distantes e dispersos quando comparamos com o ajuste *Loess*. Para o gráfico C (*anisotrophy*), percebemos que as três técnicas se ajustam de forma bem equivalente, porém a técnica *Loess* consegue captar alguns pontos mais extremos quando comparamos com as demais curvas. Percebe-se que todas as covariáveis são significativas para explicar a resposta, com as covariáveis aq.thick e anisotropphy tendo uma tendência linear.

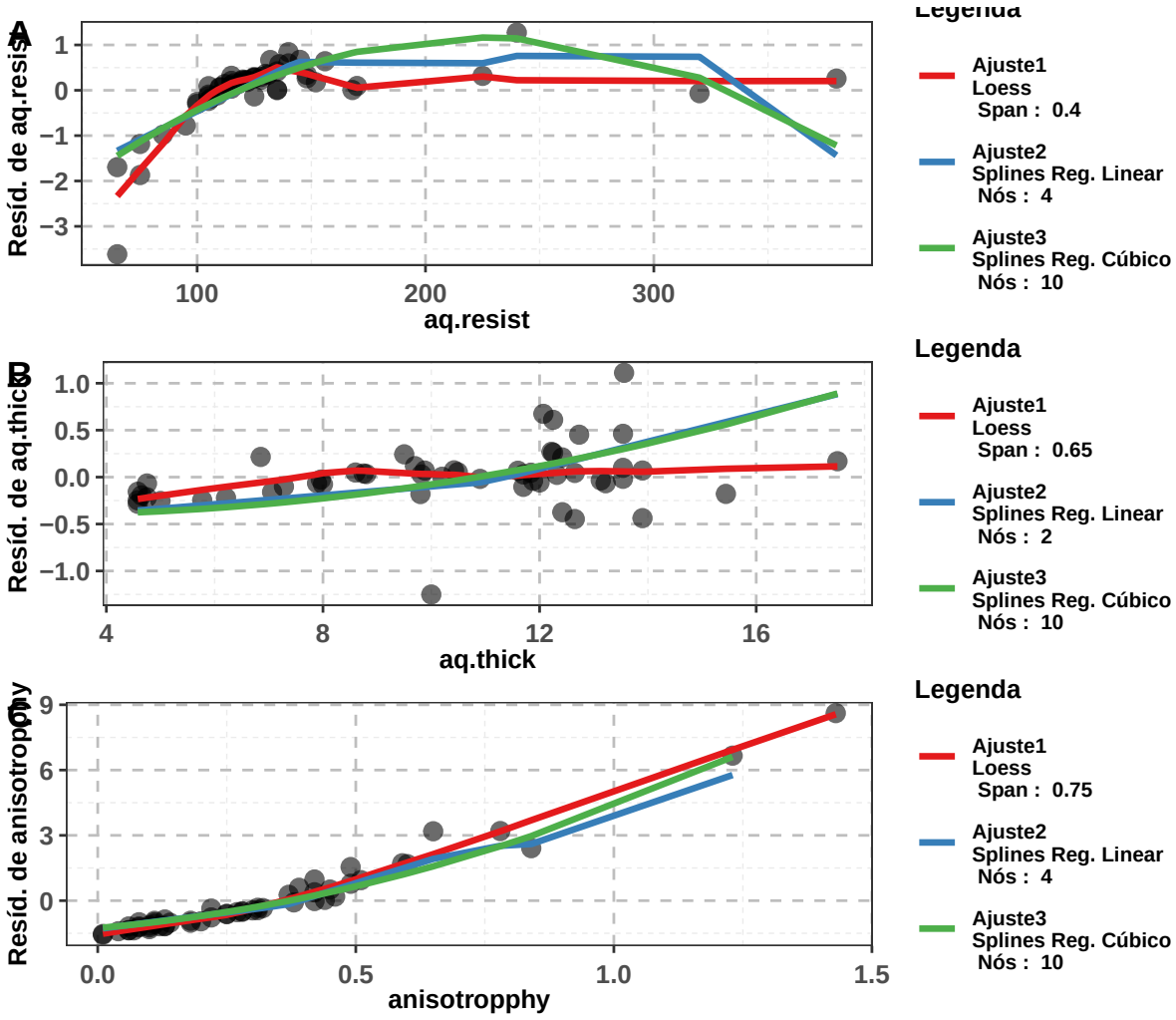

Figura 22: Resíduos parciais das variáveis aq.resist, aq.thick e anisotrophy

Inserir comparações de prediçoes

5 Conclusão

Para investigar e modelar relações entre variáveis o modelo de regressão linear pode ser utilizado, porém quando essa relação não possui forma linear, uma alternativa é o uso de ferramentas que não impõem suposições paramétricas. Nesse contexto, existem técnicas de suavização que podem ser utilizadas, inclusive na estimação das funções do componente sistemático dos modelos aditivos. Caso haja mais de uma covariável para predizer Y, o algoritmo de retroajuste pode ser uma solução.

Para validar a metodologia estudada, foram realizadas análises em dados simulados e dados reais, onde foram ajustados o modelo de regressão linear múltipla e o modelo aditivo, de acordo com o que foi estudado e com as ferramentas disponíveis. Em dados simulados em diferentes cenários, foram observados os resultados em relação ao comportamento de técnicas de suavização, bem como a aplicação do algoritmo de retroajuste em cenários simulados com modelos aditivos.

Vale ressaltar que a escolhas dos parâmetros de suavização foram feitas arbitrariamente, porém existem técnicas específicas para fazer essa escolha de modo otimizado, sendo a mais popular a validação cruzada e a validação cruzada generalizada. Além disso, destaca-se que as análises realizadas em relação à qualidade dos ajustes foram estritamente gráficas. Existem medidas numéricas adequadas para fazer essa análise, que serão abordadas em trabalhos futuros.

6 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.

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.

Izbicki, Rafael; Santos, Tiago Mendonça. **Aprendizado de máquina: uma abordagem estatística**. ISBN 978-65-00-02410-4.

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