UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Suavização Não-Paramétrica e Análise de Variância Funcional

Autor: Paulo Corrêa da Silveira Neto
Orientador: Professor Dr. Flávio Augusto Ziegelmann

Porto Alegre, 16 de janeiro de 2013.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Suavização Não-Paramétrica e Análise de Variância Funcional

Autor: Paulo Corrêa da Silveira Neto

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Flávio Augusto Ziegelmann
Professor Me. Fernando Augusto Boeira Sabino da Silva

Porto Alegre, 16 de Janeiro de 2013.

Resumo

Análise de Dados Funcionais consiste em tratar e modelar problemas estatísticos onde as unidades amostrais são funções. Tal abordagem permite que novos problemas sejam tratados, propondo novas soluções e também trazendo novas dificuldades que precisam ser levadas em consideração.

Como os dados são coletados originalmente de forma discreta, são necessários métodos de suavização, tais como decomposição de funções em bases ortogonais, Sistema de bases de Fourier, B-Splines e Regressão Local por método Kernel, a fim de obter unidades amostrais funcionais.

Neste trabalho apresentaremos estatísticas descritivas, intervalos de confiança, testes de hipóteses e modelos estatísticos definidos para o contexto de dados funcionais. Para melhor entendimento, exemplos utilizando bancos de dados reais são apresentados em todos estes casos, e os resultados devidamente interpretados.

Por fim, uma aplicação mais detalhada do modelo de Análise de Variância Funcional é abordada. Os dados coletados através de um instrumento de pesquisa específico são o tempo de reação de crianças com diferentes transtornos psiquiátricos. Foram detectadas e encontradas tanto diferença entre as médias quanto entre as variâncias dos grupos em questão.

Sumário

1. Introdução

Com a chegada de novas tecnologias, os bancos de dados hoje em dia são cada vez maiores e vem tornando-se mais fácil e comum a coleta de dados mais complexos, como, por exemplo, curvas. Assim, podem-se analisar e/ou modelar curvas e suas características, diferenças entre grupos de curvas, suas derivadas, explicações sobre a variabilidade dessas curvas, entre outras tantas. Este é o foco da Análise de Dados Funcionais, ou seja, utilizar as funções como unidades amostrais, modelá-las e analisá-las.

Uma das principais motivações é analisar amostras de curvas num contexto teórico/metodológico adequado, já que estes dados são realmente curvas, apesar de muitas vezes serem tratados vetorialmente

Como os dados, as estatísticas, os testes, etc, são objetos funcionais, é preciso exibir os dados de maneira que suas características sejam destacadas, na maioria das vezes através de gráficos, o que facilita a interpretação e a compreensão das análises.

No Capítulo 2, conceitos de suavização serão abordados. Como os dados são coletados de maneira discreta, precisamos transformar estes em curvas para realizar as análises. Falaremos sobre dois casos específicos de bases ortogonais: Séries de Fourier e os B-Splines, discutindo-se como estimar os parâmetros associados a tais bases, via Mínimos Quadrados e Mínimos Quadrados penalizados através de *Roughness Penalty*. Além disso, abordaremos Regressão Polinomial Local, outro método para transformação dos dados brutos em curvas.

No Capítulo 3, apresentaremos uma breve introdução sobre estatísticas descritivas definidas no caso em que as unidades amostrais são funções. Também é apresentado um método para calcular intervalos de confiança para tais estatísticas através de *bootstrap*. Um exemplo prático é apresentado.

No Capítulo 4 são expostos três tipos de modelos funcionais lineares, e para cada um deles é apresentado um exemplo. Os coeficientes funcionais estimados, assim como avaliações dos ajustes dos modelos, testes estatísticos funcionais e interpretação das análises são apresentados. Nos Anexos contidos no final deste trabalho, constam os códigos que geraram cada análise, estes comentados para possibilitar ao leitor tanto a reprodução dos exemplos quanto adaptação destes a novos bancos de dados.

No Capítulo 5 é apresentada uma aplicação do modelo de Análise de Variância Funcional, em um problema envolvendo o tempo de reação de cinco grupos de crianças com transtornos psiquiátricos. Nesta análise, estudamos funcionalmente os efeitos de cada transtorno tanto na média quanto na variância das curvas das crianças. Além dos testes

estatísticos, são apresentados os intervalos de confiança para os efeitos funcionais estimados tanto na média quando na variabilidade.

2. Suavização

Como os dados não são coletados diretamente em forma de curvas, e sim discretamente, faz-se necessária a passagem dos dados da forma discreta para a forma de funções, para termos valores das unidades amostrais para quaisquer valores num intervalo definido. Esta etapa já é parte do processo de modelagem e da análise, tendo em vista que, no momento que definimos as funções em cima de um sistema de bases, nossas estimativas também estarão definidas assim. Segundo Ramsay e Silverman (2005), é de muita importância um cuidadoso e bem elaborado processo de suavização, visto que, se realizado sem cuidado, pode não representar bem as curvas e suas características.

Existe uma gama de possíveis técnicas de suavização. Neste trabalho foram revisados resumidamente três tipos: Sistema de Bases de Fourier, Splines e Regressão Local por método Kernel.

2.1. Sistema de Bases

A ideia é representar uma função escolhida através de uma expansão do tipo

$$x(t) = \sum_{k=1}^{\infty} c_k \phi_k(t) \tag{2.1}$$

onde ϕ_k são as funções base. Ou seja, assumiremos que a função $x(t)$ pode ser decomposta através de uma combinação linear das funções base $\phi_k(t)$. Sistemas de base geralmente possuem bases ortogonais, como nos dois exemplos que citaremos posteriormente. Tal propriedade é desejada para que não haja possibilidade de, com diferentes coeficientes c_k, encontrarmos a mesma aproximação para $x(t)$, ou seja, as soluções são únicas.

Pensando em uma estimativa para $x(t)$, como a função não é conhecida, há apenas a possibilidade de avaliá-la em alguns pontos $(t, x(t))$, $t \in 1,\ldots,n$. Portanto, dadoque as expansões têm infinitos termos, não conseguiríamos calcular todos os coeficientes de todas as funções base. Frente a esta dificuldade, escolheremos um número K finito de funções base para truncar a aproximação. K não necessariamente precisa ser tão grande quanto possível, pois, por exemplo, se as observações $x(t)$ possuírem algum ruído de medida, não gostaríamos que este fosse incorporado às curvas estimadas. Uma solução para o problema mencionado seria diminuir o número de funções base, fazendo com que a curva fique mais

suave. De fato, quanto menos funções na base, menos elas conseguirão adaptar-se a pequenas variações e também aproximar-se-ão menos da curva observada nos pontos avaliados.

Podemos pensar também em possibilidades de Análise de Multiresolução, assim é possível utilizar poucas bases para dispensar pequenas variações e dar atenção apenas às variações maiores. E após, em um segundo estágio, utilizar um grande número de bases para levar em consideração as frequências mais altas e explorá-las separadamente.

Devemos, se possível, escolher uma base de funções que contemple as características das curvas a serem suavizadas, isto se tais características existirem e forem conhecidas. Tal escolha facilita a suavização e possibilita melhores resultados utilizando um número menor de funções base. Por consequência, menos complexa será a estimação e menos graus de liberdade serão necessários, assim como menos recursos computacionais serão exigidos.

A seguir, trataremos de dois dos sistemas de base bastante utilizados, Fourier e *B-Splines*, os quais possuem rotinas prontas para *R*, *S+* e *MATLAB*, entre outros.

2.2. Sistema de Bases de Fourier

Ideal para dados com periodicidades, o Sistema de Bases de Fourier decompõe as funções em combinações de senos e cossenos da seguinte forma:

$$x(t) = c_0 + c_1 sen(2\pi\omega t) + c_2 \cos(2\pi\omega t) + c_3 sen(4\pi\omega t) + c_4 \cos(4\pi\omega t) + \ldots \qquad (2.2)$$

onde ω é uma constante associada ao domínio de $x(t)$ de forma que o período da função seja $2\pi/\omega$. Ou seja, as primeiras parcelas de seno e cosseno oscilarão uma vez durante o domínio de $x(t)$, as parcelas associadas a c_3 e c_4 oscilarão duas vezes durante o domínio, e assim por diante. Já c_i são os coeficientes a serem estimados, que multiplicarão as bases para a composição de $x(t)$.

Como a decomposição é infinita, mas nossos pontos coletados de $x(t)$ não, truncamos a soma em K parcelas. É importante definir que, a fim de manter a ortogonalidade entre as funções da base, sempre quando uma parcela de seno entra na soma, sua parcela de cosseno com mesmo período deve entrar também. Assim como no caso da inserção de uma parcela de cosseno, sua parcela de seno associada deve ser

incluída da mesma maneira. Desta forma, K é sempre um número ímpar, uma vez que é composto por pares de parcelas mais a constante c_0.

Geralmente a quantidade de parcelas a entrarem na soma é escolhida arbitrariamente. Podemos escolher uma quantidade pequena de parcelas se desejamos analisar apenas os períodos grandes das curvas, e aumentar o número se o interesse são os ciclos mais curtos. Tal característica é atraente, pois podemos decidir qual o foco de interesse da análise ou focar separadamente em altas ou baixas frequências.

Segundo Ramsay e Silverman (2005), Séries de Fourier são muito utilizadas por sua agilidade computacional e fácil adaptação a quaisquer espécies de dados, porém, nem sempre os resultados são tão bons quanto se espera, comentam*: "A Fourier series is like margarine: It's cheap and you can spread it on practically anything, but don't expect that the result will be exciting eating"*. Segue, na Figura 2.1, as temperaturas coletadas durante os 365 dias do ano em uma estação climática. No exemplo, a curva foi suavizadas através de um Sistema de Bases de Fourier com 65 funções base.

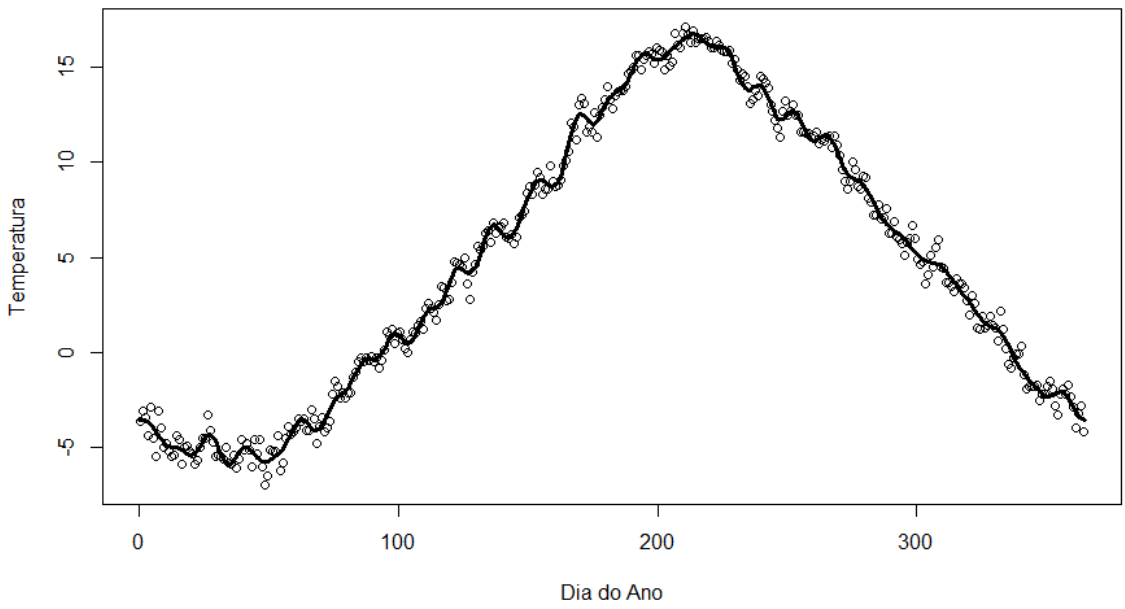


Figura 2.1 – Curva suavizada através de Séries de Fourier

2.3. *B-Splines*

B-Splines (Basis Splines) é um sistema de bases formado por polinômios ortogonais proposto por De Boor (1978), o qual tem sido muito utilizado. Segundo De Boor (2001), além de produzir ótimos resultados em termos de suavização, *B-Splines* requerem pouco

esforço computacional. Em De Boor (2001) também podem ser encontradas revisões sobre outros tipos de *Splines*, além de todas as definições sobre o que será tratado na Seção 2.3.[1]

Mais versáteis que as séries de Fourier, devem se sair melhor quando empregados para suavizar curvas onde não necessariamente há uma periodicidade específica, já que uma parcela de seno ou cosseno, uma vez adicionada, perturba a curva em toda sua extensão. Por ser uma base polinomial, suas derivadas também são suaves, contínuas e bem comportadas, o que torna essa técnica muito útil quando é de interesse do estudo analisar as derivadas das curvas e suas características.

2.3.1. Obtenção dos polinômios da base

Dado um vetor de $n+1$ nós arbitrariamente escolhidos $u = [u_0, u_2, \dots u_n]$, os polinômios de um grau p, também especificado, que formam a base são construídos pelo seguinte algoritmo:

$$N_{i,0}(u) = \begin{cases} 1 & se\ u_i \le u \le u_{i+1} \\ 0 & caso\ contrário \end{cases}, \qquad i = 0,1, \dots n,$$

$$N_{i,p}(u) = \frac{u-u_i}{u_{i+p}-u_i} N_{i,p-1}(u) + \frac{u_{i+p+1}-u}{u_{i+p+1}-u_{i+1}} N_{i+1,p-1}(u), \quad u_i < u < u_{i+p+1}\ . \quad (2.3)$$

Uma vez que $N_{i,1}$ é computado por meio de $N_{i,0}$ e $N_{i+1,0}$ e dado que $N_{i,0}$ e $N_{i+1,0}$ são não-zero em $[u_i, u_{i+1}]$ e $[u_{i+1}, u_{i+2}]$, $N_{i,1}$ é não zero em $[u_i, u_{i+2}]$. Da mesma maneira, uma vez que $N_{i,2}$ depende de $N_{i,1}$ e $N_{i+,1}$ e visto que essas duas bases são não-zero em $[u_i, u_{i+2}]$ e $[u_{i+1}, u_{i+3}]$ respectivamente, então $N_{i,2}$ é não-zero em $[u_i, u_{i+3}]$. Seguindo esta linha de raciocínio, podemos mostrar que qualquer polinômio $N_{i,p}$ é não-zero no intervalo $[u_i, u_{i+p+1}]$. O diagrama na Figura 2.2 ilustra a lógica do algoritmo. Ressaltado em vermelho, podemos observar em quais intervalos o polinômio $N_{1,2}$ é não zero.

[1] Uma revisão rápida, fácil, com muitas figuras e bastante intuitiva pode ser acessada em http://www.cs.mtu.edu/~shene/COURSES/cs3621/NOTES/spline/B-spline/ (acessado em 15/12/2012 21:00)

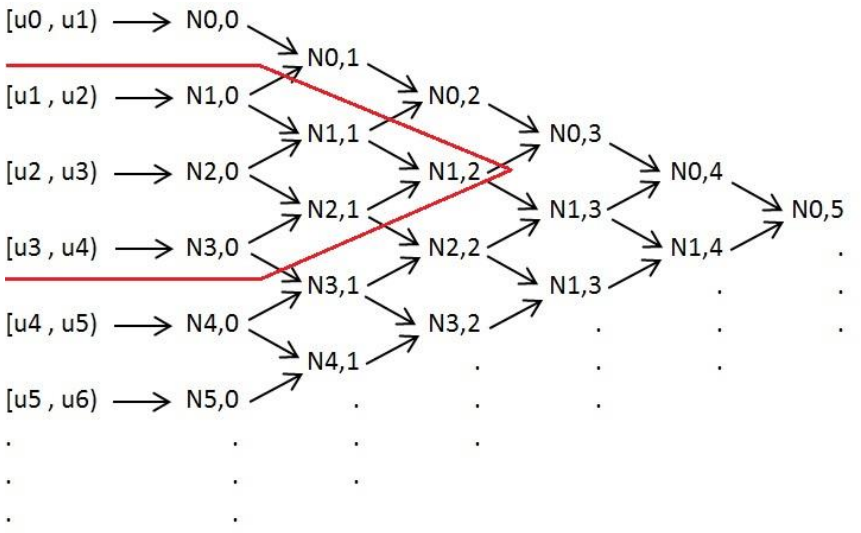


Figura 2.2 – Diagrama ilustrando como identificar o intervalo onde cada polinômio é não zero.

Seguindo a linha de construção anterior, só que de maneira inversa e com a ajuda do diagrama, podemos notar que no máximo $p+1$ polinômios $N_{.,p}$ são positivos em cada intervalo $[u_i, u_{i+1}]$. Como exemplo, a Figura 2.3 ilustra quais os quatro polinômios de grau 3 que são positivos no intervalo$[u_3, u_4]$. Porém, como mostra o diagrama, nem todo intervalo possui $p+1$ polinômios de grau p positivos, no exemplo, no intervalo $[u_2, u_3]$ existem apenas três polinômios positivos.

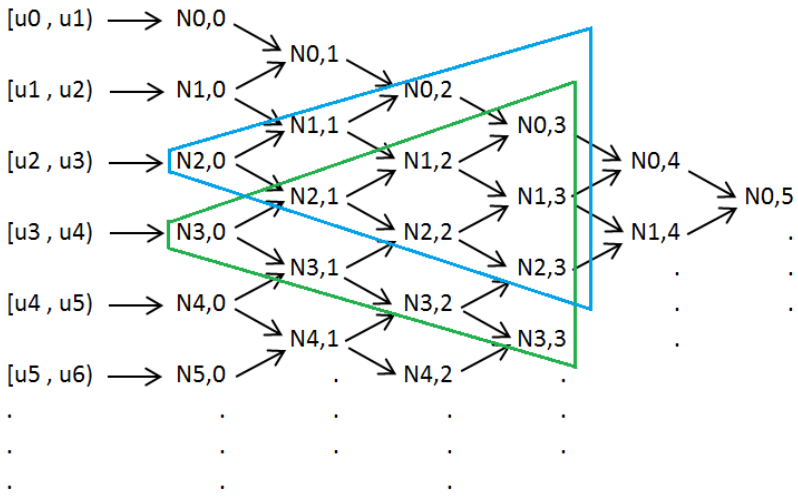


Figura 2.3 – Diagrama ilustrando a quantidade de polinômios positivos nos intervalos entre os nós

2.3.2. Escolha do grau do polinômio

Podemos definir uma base de polinômios de qualquer grau desejado. Na literatura, geralmente são utilizados polinômios de grau 3, por não haver uma perda relevanteante em relação aos de grau 4 e por serem mais fáceis de computar que os de grau maior. A Figura 2.4 mostra duas bases, a primeira definida com $p = 1$ e a segunda com $p = 2$. Note que as funções base são polinômios de grau p entre os nós.

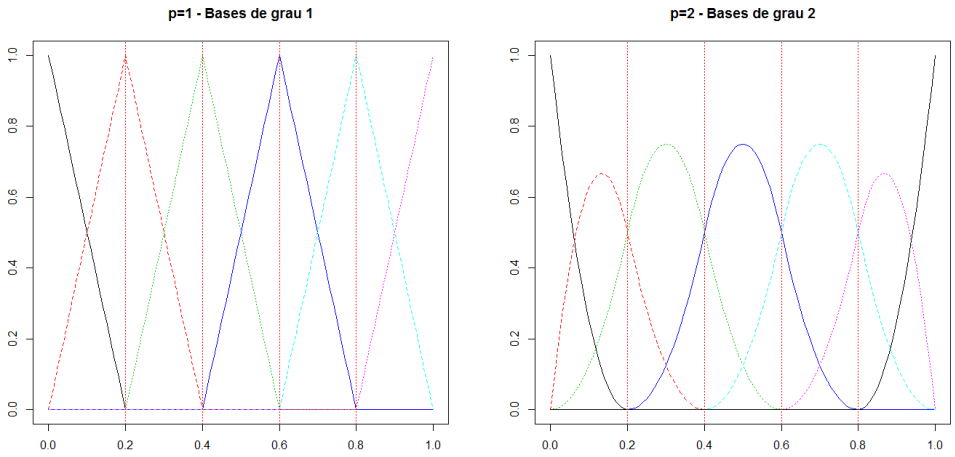


Figura 2.4 – Bases polinomiais de grau 1 e 2 respectivamente.

2.3.3. Sobre a escolha dos nós

A escolha dos nós é um procedimento muito relevante quando estamos suavizando curvas utilizando uma base formada por B-Splines. A quantidade e o posicionamento dos nós define a suavidade das funções tanto localmente como por todo o intervalo onde estão sendo suavizadas.

2.3.3.1. Número de nós

A quantidade de nós está ligada à quantidade de polinômios na base especificada, assim como ao grau dos mesmos. Uma base de polinômios de grau p com $n + 1$ nós tem exatamente $n + 1 - p - 1$ polinômios na base. Quanto mais funções na base, mais a base se ajustará aos dados, ou seja, mais perto dos pontos a curva final passará, e menos suave ela será. A Figura 2.5 ilustra a influência da quantidade de nós na quantidade de funções na base, as linhas verticais em vermelho indicam a localização de cada um dos nós.

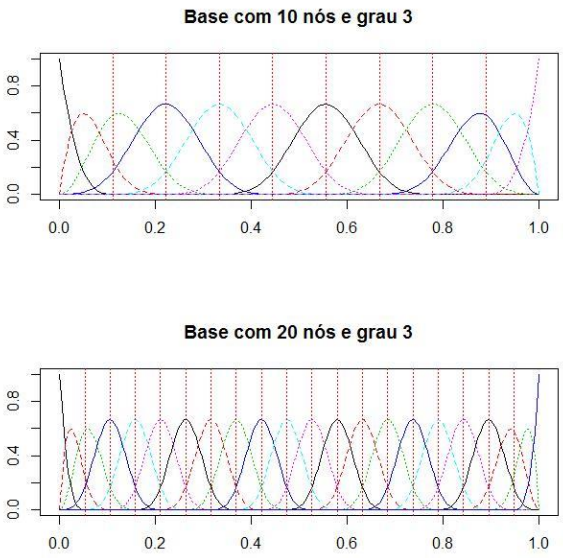


Figura 2.5 – Em cima, definida uma base com 10 nós e grau 3, abaixo, 20 nós foram utilizados.

2.3.3.2. Nós múltiplos

Como demonstramos anteriormente, um polinômio $N_{i,p}$ é não zero em $[u_i, u_{i+p+1}]$. Um artifício para forçar mais polinômios a terminarem em certo nó é repetir o nó específico no vetor $\boldsymbol{u}$. Para cada adição do nó no vetor, mais um polinômio termina naquele instante. Esta técnica é útil para quando a curva em questão tem uma forte inflexão em alguma região, fazendo assim com que ali terminem as bases e daquele local partam as próximas.

Segue na Figura 2.6 um exemplo de base onde foi sendo aumentada a multiplicidade de um nó específico. Na primeira imagem, apenas um polinômio se encerra no nó de valor 0,5, na segunda imagem, no topo à direita, dois nós terminam ali. Na terceira imagem, abaixo da primeira, aumentamos mais uma vez a multiplicidade do nó e mais uma função base se encerra ali. Na última fica bem caracterizado quando a função é forçada a decrescer rapidamente para zero quando chega perto do nó múltiplo.

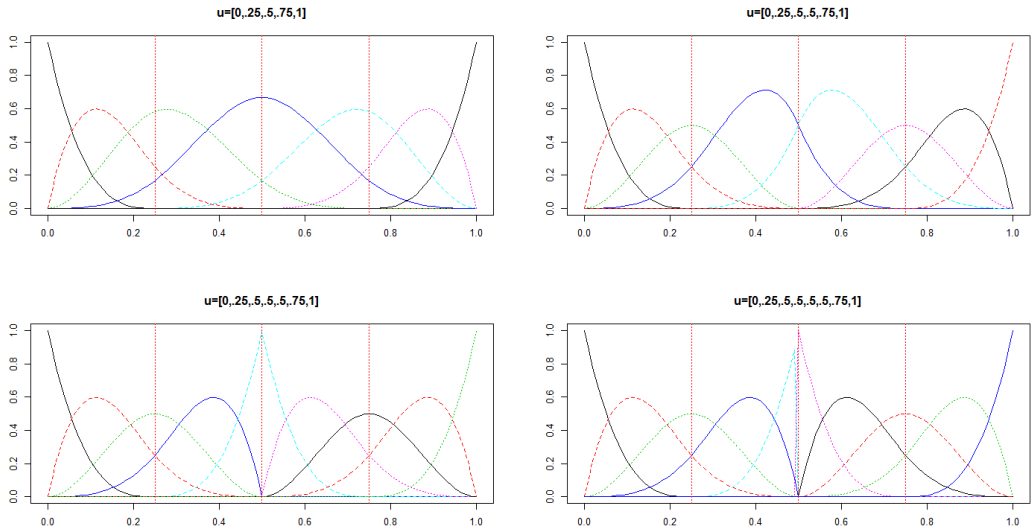


Figura 2.6 – Aumentando gradualmente a multiplicidade do nó em 0,5

2.3.3.3. Espaçamento entre os nós

Se o espaço entre os nós é constante, ou seja, $u_{i+1} - u_i = c$, então serão chamados de nós uniformes. Os nós não necessariamente precisam ser uniformes, podemos concentrar mais nós em regiões onde as curvas variam mais, ou apresentam mais oscilações. A Figura 2.7 ilustra como a distribuição dos nós afeta as funções na base. Mesmo a quantidade de polinômios não zero entre cada par nós sendo mantida, nos intervalos onde os nós são mais próximos, os polinômios tem menores durações, enquanto são mais longos nos intervalos maiores.

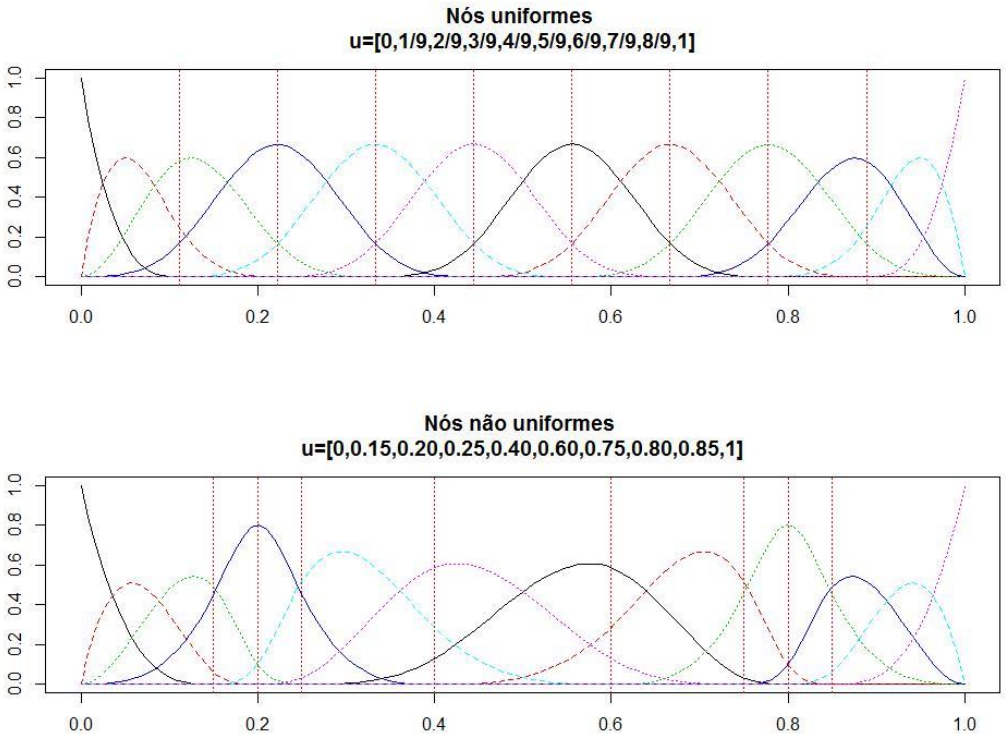


Figura 2.7 - Exemplo de bases com nós uniformes e não uniformes.

2.4. Ajustando as bases

Definido um sistema de bases, precisamos ajustar ele aos nossos dados coletados, a fim de obter as curvas estimadas para a análise. Retomando a Equação (2.1) devemos encontrar os componentes do vetor de coeficientes $\boldsymbol{c}$, os quais multiplicarão as funções da base especificada anteriormente para finalmente compor a nova função suavizada estimada. Exploraremos dois métodos de suavização, o usual método de mínimos quadrados e uma de suas extensões, mínimos quadrados com penalidade por não suavidade.

2.4.1. Mínimos quadrados

Minimizar o quadrado da distância até cada um dos pontos é o procedimento mais simples para minimizar a distância entre a nova curva suavizada e os pontos, como fazemos no modelo de regressão linear. A descrição do estimador, na forma matricial, também é a mesma. Dado um conjunto de pontos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, seja $\boldsymbol{y} =$

$[y_1, y_2, \dots, y_n]$ e $\boldsymbol{\Phi}$ uma matriz nxK contendo em nas linhas os valores das K funções base avaliadas nos pontos $x_1, x_2, \dots, x_n$ e $\boldsymbol{c} = [\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_k}]$ um vetor de tamanho K com as constantes a serem estimadas temos:

$$SSE(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})'(\boldsymbol{y} - \boldsymbol{\Phi c}), \tag{2.4}$$

derivando em relação a $\boldsymbol{c}$ nos leva a seguinte equação:

$$2\boldsymbol{\Phi\Phi}'\boldsymbol{c} - 2\boldsymbol{\Phi}'\boldsymbol{c} = 0. \tag{2.5}$$

Resolvendo para $\boldsymbol{c}$ obtemos o estimador clássico de mínimos quadrados ordinários:

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{y}. \tag{2.6}$$

Um dos problemas de estimar através de mínimos quadrados ordinários é que este estimador escolhe $\boldsymbol{c}$ de maneira que aproxima ao máximo a curva suavizada dos pontos, podendo trazer estimativas de funções pouco suaves se o número de funções na base é alto. Ou seja, dada uma base fixa, não restam opções de escolha para obtenção de curvas mais ou menos suaves.

2.4.2. *Roughness Penalty*

É possível controlar a suavização mantendo a base fixa ao se adicionar uma penalidade à falta de suavidade das curvas. Uma das maneiras mais comuns de proceder tal tarefa é adicionar à quantidade objetivo do processo de minimização um termo relacionado à segunda derivada da curva suavizada:

$$PEN(x) = \int [D^2 x(s)]^2 ds. \tag{2.7}$$

Curvas com uma variabilidade alta, que oscilam muito, produzem valores maiores de $PEN(x)$. Portanto, ao adicionar tal parcela na quantidade a ser minimizada, a minimização levará em conta também o quanto a curva estimada será suave ou não. Adicionando um parâmetro arbitrário λ multiplicando a penalidade, conseguimos decidir o quanto a penalidade pesará na estimação. A quantidade a ser minimizada será

$$\sum_{j=1}^{n}[y_j - \sum_{k}^{K} c_k \phi_k(t_j)]^2 + \lambda \int [D^2 x(s)]^2 ds. \quad (2.8)$$

Definir a magnitude de λ pode ser uma tarefa complexa. De fato, como a penalidade é definida por uma integral ao longo de toda a curva, o tamanho de λ depende também do tamanho do intervalo onde ela está definida. O valor de λ ainda depende da natureza da curva; se as curvas de interesse são menos suaves, menores valores para λ são requeridos. Ramsay e Silverman (2005) dedicam um capítulo a exemplos de suavização com penalidade e também à escolha de λ. A Figura 2.8 mostra um exemplo onde uma curva foi suavizada sem penalidade, e então foram adicionadas as penalidades $\lambda = 10^{-5}, 5 * 10^{-5}\ e\ 10^{-4}$, respectivamente. Nos três gráficos, em preto encontra-se a curva suavizada sem penalidade e em colorido a função estimada com a penalidade adicionada. Note que à medida que λ cresce, as curvas obtidas tornam-se mais suaves.

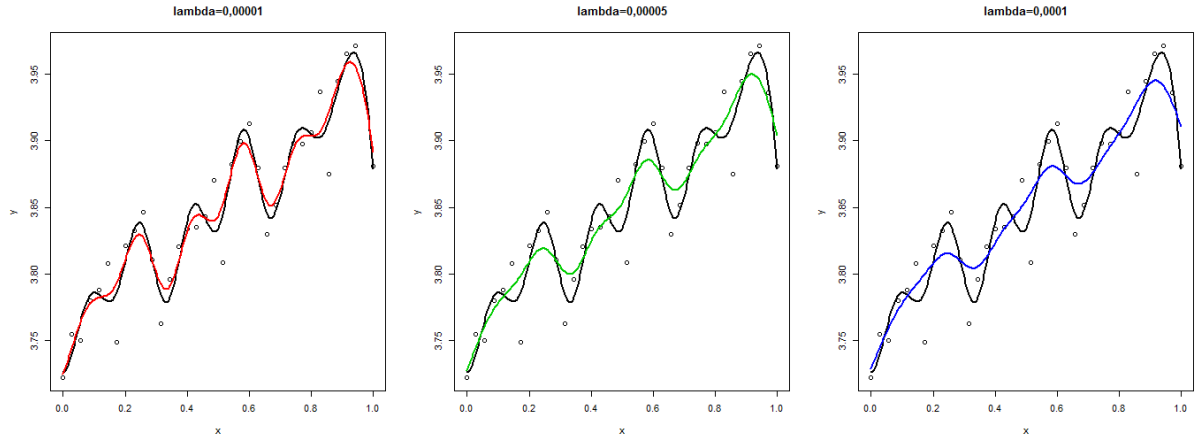


Figura 2.8 – Suavização utilizando três níveis de *Roughness Penalty*

2.5. Regressão Polinomial Local via *Kernel*

Suavizar por regressão significa investigar a associação entre uma variável explicativa X e uma variável resposta Y, como no modelo mais tradicional de regressão linear. Porém, enquanto no contexto clássico há uma suposição de que a relação entre Y e X se dá de forma linear, ao utilizarmos regressão local, permitimos que a curva estimada tenha qualquer formato, linear ou não linear, desde que com um certo grau de suavidade.

2.5.1. Estimação

Dado um conjunto de pontos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, o objetivo deste método é estimar a relação funcional entre Y e X, modelando esta como

$$Y = m(X) + \varepsilon \tag{2.10}$$

onde $E(\varepsilon|X) = 0$ *e* $Var(Y|X) = \sigma^2(X)$, no caso mais geral com possível heterocedasticidade.

Note que o termo $m(x)$, pode ser expandido em uma Série de Taylor da seguinte forma:

$$m(x) = m(x_0) + m^{(1)}(x_0)(x - x_0) + \frac{m^{(2)}(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p + \varepsilon, \tag{2.11}$$

onde x_0 é um valor próximo de x e $m^{(p)}$ a p-ésima derivada de $m(x)$. Ou seja, podemos aproximar $m(x)$ por um polinômio de ordem p. Então, tomando um ponto x no domínio de X, podemos definir o estimador polinomial local de $m(x)$ como $\hat{m}_p(x) = \hat{\beta}_0$, onde $\hat{\beta}_0$ é a solução do seguinte problema de mínimos quadrados ponderados:

$$\left[\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right] = \arg \min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^{n} \{Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j\}^2 K\left(\frac{X_i - x}{h}\right), \tag{2.12}$$

onde K(·) é uma função peso *Kernel* e h um parâmetro de suavização.

2.5.1.1. Estimadores de Nadaraya-Watson e Linear Local Kernel

Resolvendo (2.12) para $p = 0$ obtemos o estimador de Nadaraya-Watson, proposto por Nadaraya (1964) e Watson (1964) antes mesmo da proposição do problema de Regressão Polinomial Local:

$$\hat{m}_{nw}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}, \tag{2.13}$$

onde K(·) é uma função *Kernel* e h um parâmetro suavização. Calculando o estimador acima para uma grade de pontos obtemos uma curva. Para controlar a suavização existe o parâmetro h. Quanto maior o seu valor, mais suave será a suavização, já um valor de h baixo resulta em uma estimativ mais ruidosa.

Já quando tomamos $p = 1$, obtemos o estimador linear local, que tem a seguinte forma:

$$\hat{m}_l(x) = n^{-1} \sum_{i=1}^{n} \frac{[\hat{s}_2(x) - \hat{s}_1(x)(X_i - x)]K\left(\frac{X_i - x}{h}\right)Y_i}{\hat{s}_2(x)\hat{s}_0(x) - \hat{s}_1(x)}, \quad (2.14)$$

onde

$$\hat{s}_r(x) = n^{-1} \sum_{i=1}^{n} (X_i - x)^r K(\frac{X_i - x}{x}). \quad (2.15)$$

O estimador linear local se mostra melhor que o estimador de Nadaraya Watson nos limites do domínio das funções estimadas, além de possuir um viés menor. Uma revisão pode ser encontrada em Fan (1992) e Hardle et al (2004).

2.5.2. Funções *Kernel*

Geralmente uma função Kernel é uma função densidade de probabilidade simétrica ao redor do zero, ou seja, satisfazendo as seguintes suposições:

- $\int K(u)du = 1$
- $\int uK(u)du = 0$
- $K(u) = K(-u)$

Na Tabela 1, em Anexos, há alguns exemplos de funções Kernel. Note que salvo a função uniforme, a lógica das funções é designar pesos maiores para os menores valores de $\left|\frac{x-Xi}{h}\right|$, e pesos menores para os valores mais altos desta quantidade. Também é válido ressaltar que, apenas a função Kernel Normal atribui pesos a todos os vizinhos em torno de X_i, enquanto as outras dão peso zero se $\left|\frac{x-Xi}{h}\right| > 1$.

2.5.3. Parâmetro de suavização

A escolha de h, chamado também de *bandwidth*, ou tamanho de janela, influencia decisivamente na suavidade das curvas: quanto maior o tamanho de h, menor será a quantidade $\frac{x-Xi}{h}$, e, portanto, maiores os pesos dos valores que estão longe do ponto a ser estimado. A especificação deste parâmetro pode ser feita tanto de maneira subjetiva, quanto determinada automaticamente pelos dados. Existem inúmeras sugestões de como definir o tamanho de h, desde regras de bolso, como podemos encontrar em Silverman (1998), ligadas à variação dos pontos ou até ao tamanho do intervalo de suavização, passando por critérios computacionais que minimizam o erro quadrático médio integrado, ou até utilizando métodos de validação cruzada.

Na Figura 2.9 e na análise posterior, foi utilizado o critério de seleção sugerido por Ruppert, Sheather e Wand (1995). No gráfico foram sorteados vinte pontos de função especificada e a estes adicionados erros aleatórios provindos de distribuição normal. A curva vermelha é a estimada utilizando a função Kernel Normal. Olhando para as outras duas estimativas, confirmamos o que já nos mostrava o estimador de Nadaraya-Watson. Ao utilizar um h muito pequeno, na curva em verde, os pontos mais distantes acabam influenciando pouco na estimativa e ela acaba sendo mais ruidosa, já que os pontos com resíduos maiores puxam com facilidade a curva para perto. Já a curva azul, com parâmetro grande, não cedeu aos pontos mais altos e acabou oscilando menos que a curva original.

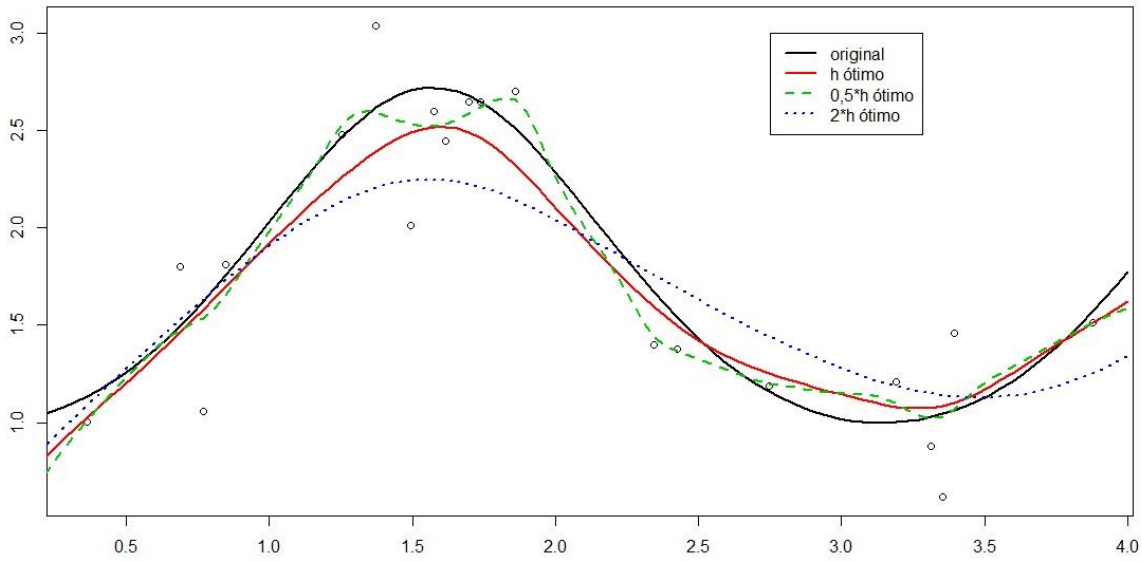


Figura 2.9 – Curva original e três estimativas utilizando h ótimo, maior e menor

3. Estatísticas Descritivas

Com as curvas definidas, podemos determinar as estatísticas descritivas clássicas para os dados funcionais, a Figura 3.1 mostra as curvas da altura de 39 meninas medidas 31 vezes do 1 aos 18 anos, o banco de dados está disponível no pacote *fda* do software *R*:

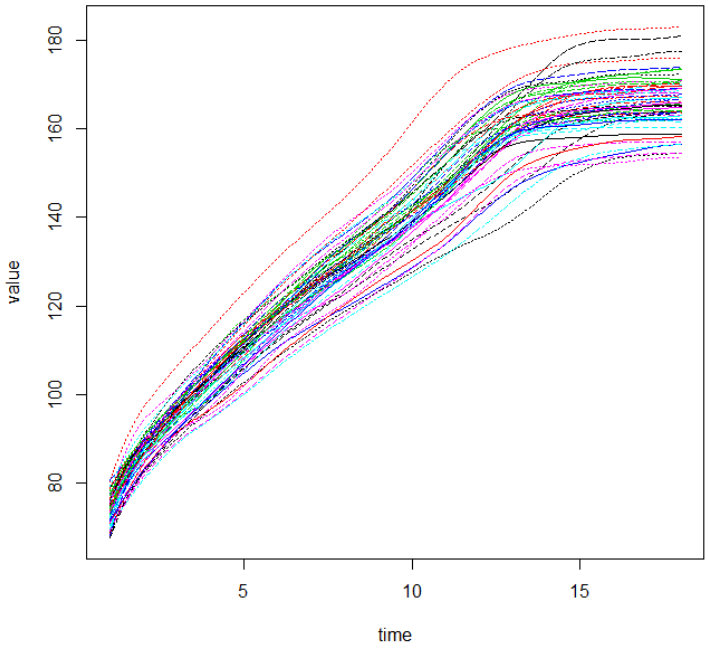


Figura 3.1 - Curvas das alturas das meninas estudadas

Dado um conjunto de curvas $x_1(t), x_2(t), \ldots, x_n(t)$, é possível definir e calcular a média e a variância dessas curvas. As fórmulas não são diferentes das que já conhecemos, porém é importante ressaltar que as observações agora são objetos funcionais:

$$\bar{x}(t) = N^{-1}\sum_{i=1}^{N} x_i(t), \qquad \mathrm{var}_X(t) = (N-1)^{-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2 \ . \tag{3.1}$$

Na Figura 3.2 estão ilustradas a curva média e a curva da variância das alturas das meninas mencionadas anteriormente. No primeiro gráfico, podemos notar que após os 15 anos, a altura das meninas tende a crescer menos. Já no segundo, é notável a alta variabilidade entre os 10 e os 15 anos, período onde as meninas crescem mais rapidamente, porém não ao mesmo tempo. Uma interpretação é de que algumas meninas demoram mais para crescer do que as outras, o que pode ter causado a alta variabilidade neste período.

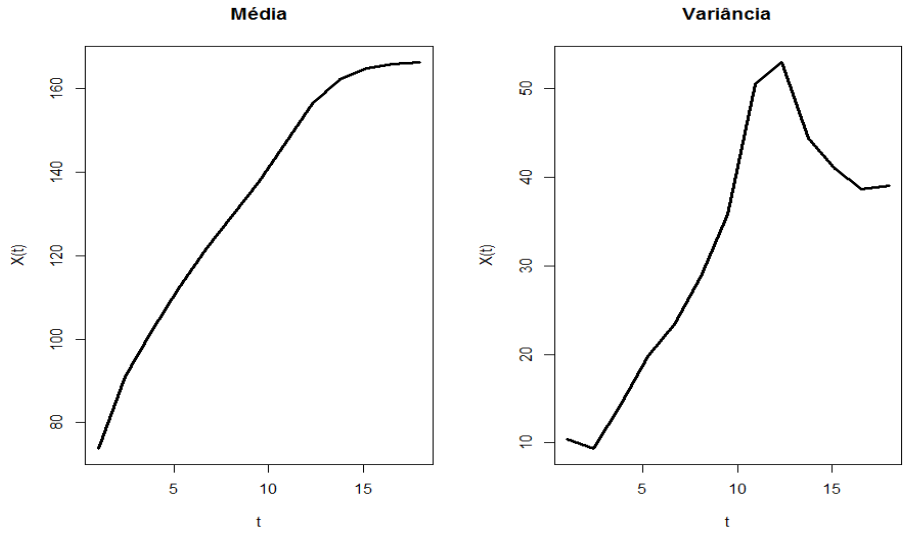


Figura 1.2- Média e Variância das curvas citadas anteriormente

3.1. Intervalos de Confiança através de *bootstrap*

Pode-se também definir intervalos de confiança para as médias e variâncias das curvas descritas anteriormente. Como não supomos distribuições de probabilidade para as curvas em seus níveis, e também não para as curvas como um todo, tal intervalo é construído através de *bootstrap*. Por exemplo, para a média, mil replicações são reamostradas com reposição das curvas com o mesmo tamanho de amostra original, formando assim mil novas curvas médias. Essas médias são ordenadas através de uma métrica definida, e assim calculamos os quantis para obtenção dos intervalos. Como as curvas são ordenadas por uma métrica que ordena as curvas como um todo, e não ponto a ponto, estes intervalos podem ser interpretados ao longo da curva sem perda do nível de significância. Uma revisão sobre intervalos de confiança através de bootstrap pode ser encontrada em Wang e Gasser (1998). Na Figura 3.3 retomamos as curvas das meninas agora mostrando o intervalo com 95% confiança para a curva média obtida via *bootstrap*.

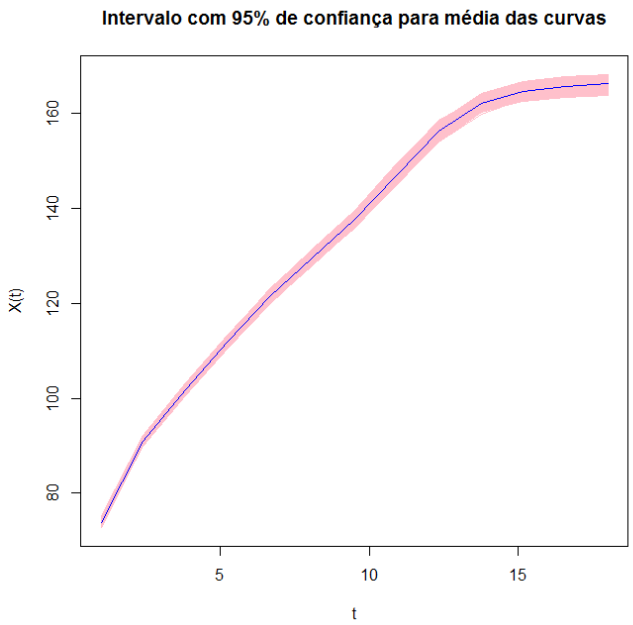


Figura 3.3 - Intervalo de confiança para a altura média das meninas através de bootstrap

4. Modelos funcionais lineares

Com o interesse de explicar o comportamento de grupos de curvas, suas causas de variação, variáveis que causam efeitos e até realizar predições, surge então a necessidade de estipularmos modelos para formular, quantificar e até testar estas relações. Como as unidades funcionais são uma classe especial de unidades amostrais, resultam três modelos distintos para modelagem envolvendo tais dados provindos de curvas. Em um primeiro caso, temos uma relação de um escalar versus um funcional, muito parecido com o modelo de regressão linear simples, onde uma ou mais das variáveis regressoras são funcionais, e estas explicarão uma variável unidimensional.

Na segunda situação utilizaremos uma variável resposta funcional contra escalares ou até fatores. Além do exemplo a ser apresentado, tal caso se enquadra na aplicação no final deste trabalho, sendo essa uma representação mais completa do processo de modelagem e interpretação dos resultados. Por último, finalmente exemplificaremos um modelo com variáveis dependentes e independentes funcionais, chamado de *concurrent model.*

É importante ressaltar que, por mais que muitas vezes a intuição e as interpretações dos resultados destes modelos sejam de fácil compreensão e aplicabilidade, a matemática por trás do estabelecimento dos espaços amostrais, distribuições de probabilidades e processos de estimação é bastante complexa e não será exploradas neste trabalho. Artigos

que contemplam toda construção dos modelos, suposições, provas, estimadores e até testes de hipóteses teóricos serão indicados em cada um dos três sistemas.

Os dados utilizados para a exemplificação dos modelos são curvas de precipitação média diária e de temperatura média diária. Estas foram medidas diariamente durante entre 1960 e 1994 em 35 estações climáticas diferentes do Canadá, distribuídas entre quatro regiões do país. Realizando a média das temperaturas e precipitações em cada dia do ano medidas ao longo dos anos, obtemos uma curva média de temperatura para cada estação. Em anexo, os códigos utilizados para obtenção dos resultados através do software *R*, utilizando o pacote *fda*.

4.1. Resposta Escalar e Covariável Funcional

A ideia de uma integral surge naturalmente quando precisamos resumir as informações de uma função em um único número. É o caso deste modelo, onde uma variável funcional é utilizada para explicar uma variável escalar. Dado um conjunto de observações da variável resposta $y_1, y_2, \dots, y_n$ e um conjunto de curvas observadas $x_1(t), x_2(t), \dots, x_n(t)$, o modelo é o seguinte:

$$y_i = \beta_0 + \int \beta_1(t) x_i(t) dt + \varepsilon_i. \tag{4.1}$$

Como $x_i(t)$ é um objeto funcional, $\beta(t)$ também será. Note que não há operações funcionais entre $\beta(t)$ e $x_i(t)$, e sim o produto de seus valores avaliados em cada ponto do domínio das duas. O que $\beta(t)$ faz é ponderar o quanto cada ponto de $x_i(t)$ contribuirá para a integral, e, logo, para a estimativa de y_i. Muitas vezes, um dos esclarecimentos que $\beta(t)$ nos traz é quais pontos, intervalos ou regiões de $x_i(t)$ influenciam mais na variável resposta. O modelo, suas suposições, processo de estimação e testes de hipóteses são apresentados por Cardot, Ferraty e Sarda (1999), entre outros.

4.1.1. Exemplo de aplicação

Para exemplificar como uma variável funcional pode explicar uma variável unidimensional quantitativa, utilizaremos como variável resposta o logaritmo da soma das precipitações diárias ao longo do ano nas 35 estações citadas anteriormente, ou seja, o logaritmo da precipitação anual. Nossa variável independente é a curva de temperatura

média diária, ou seja, veremos quanto a temperatura ao longo dos dias do ano consegue explicar a precipitação anual.

Estimando o modelo anterior, obtemos coeficientes estimados $\hat{\beta}_0 = 0.009881473$, e $\hat{\beta}_1$ como na Figura 4.1. No gráfico da esquerda, tracejado em vermelho, nas linhas verticais, estão divididas as estações do ano, sendo a primeira o inverno. Podemos notar um pico negativo em $\hat{\beta}(t)$ durante o verão e valores menores durante o outono. Observa-se também que há um período em cada estação. Intuitivamente, podemos considerar que o estimador está ponderando as temperaturas diárias dentro das estações do ano. Cruzando os valores estimados contra os observados, na direita, notamos o bom ajuste do modelo, que produziu um R^2=0,87. É preciso tomar cuidado com a quantidade de funções na base proposta para o estimador, Ramsay e Silverman (2005) indicam que altos valores de K podem causar um *overfitting*, no processo de estimação, ou seja, os dados estariam explicando parte do ruído do modelo.

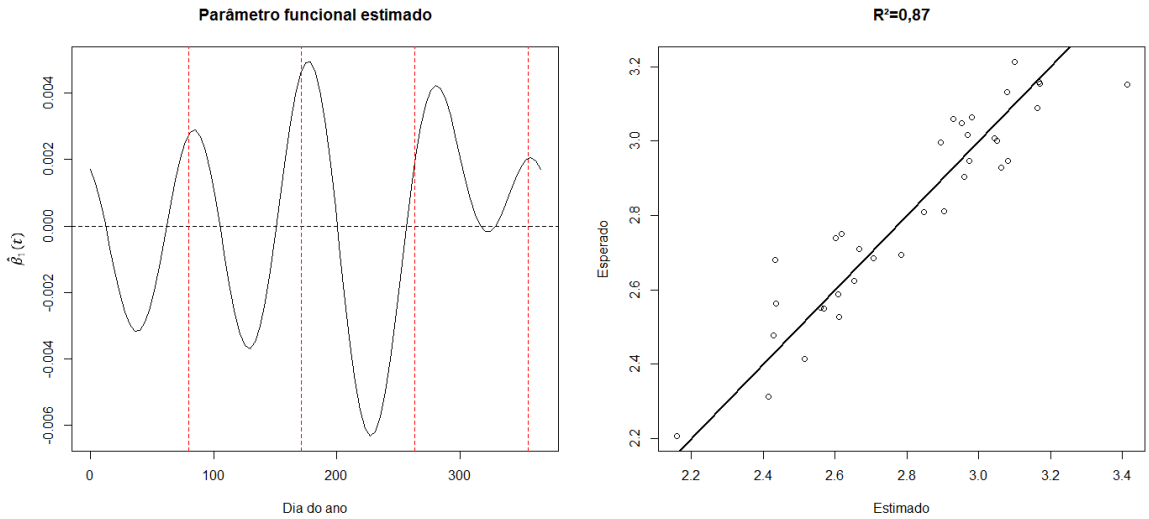


Figura 4.1 – À esquerda, $\hat{\beta}_1$ estimado, na esquerda o gráfico Estimado x Esperado.

4.2. Resposta Funcional e Covariáveis Categóricas – fANOVA

Dentro de um segundo caso, podemos ter grupos de curvas e o desejo de saber o efeito dos grupos nas curvas, ou até se há significância estatística destes efeitos. Dado um conjunto de curvas $y_1(t), y_2(t), \ldots, y_n(t)$, divididas em i grupos de curvas, pode-se definir o seguinte modelo para regressão funcional:´

$$y_{ij}(t) = \mu(t) + \alpha_i(t) + \varepsilon_{ij}(t). \quad (4.2)$$

Nesta situação, cada curva $Y_{ij}(t)$ pode ser explicada por uma curva média geral $\mu(t)$, uma curva $\alpha_i(t)$ associada ao seu fator e uma curva de erro $\varepsilon_{ij}(t)$. É possível estimar o efeito esperado de cada fator nas curvas. Como se percebe olhando em (4.2), tanto a média geral, quanto os efeitos associados aos fatores e também o erro aleatório são objetos funcionais. A parte teórica envolvendo definições matemáticaa, suposições e estimação pode ser encontrada em Zoglat (2008). Este é o modelo mais simples de Análise de Variância Funcional. Há também modelos funcionais para modelos mistos, os quais podem ser encontrados em Guo (2002). Podemos citar Spitzner, Marron e Essick (2003), como um dos trabalhos envolvendo aplicação do modelo misto funcional.

4.2.1. Estimação

Assim como o modelo de análise de variância multivariada, o modelo de análise de variância funcional é claramente uma extensão, ou até generalização do modelo clássico de ANOVA. Dada uma matriz de planejamento $\boldsymbol{X}$ e um vetor de funções $\boldsymbol{Y}(\boldsymbol{t}) = [Y_1(t), Y_2(t), \dots, Y_n(t)]$, o estimador para $\boldsymbol{\alpha}(\boldsymbol{t}) = [\alpha_1(t), \alpha_2(t), \dots, \alpha_i(t)]$ é o estimador de mínimos quadrados ordinários

$$\hat{\alpha} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \quad (4.3)$$

Este estimador, segundo Zoglat (2008), assim como no caso clássico de análise de variância, é o melhor estimador linear não-viesado (BLUE). É válido ressaltar também, que como neste caso apenas o próprio ponto t é considerado para a estimativa, é como se tal modelo realizasse uma análise de variância para cada ponto t das funções. Porém, como a teoria define o modelo como um todo, as curvas associadas aos fatores também podem ser interpretadas tanto nos pontos quanto em regiões ou até em sua totalidade. Podemos também definir estatísticas como somas de quadrados, R^2 e outras estatísticas para testes funcionalmente, como será especificado e exemplificado no exemplo que segue.

4.2.2. Exemplo de aplicação

Neste caso, utilizaremos como variável resposta as curvas de temperatura medidas ao longo do ano em 35 diferentes estações climáticas do Canadá. As funções estão divididas nas quatro regiões do país: Atlântico, Continental, Pacífico e Ártico. Na Figura 4.2 são exibidas as curvas separadas por região.

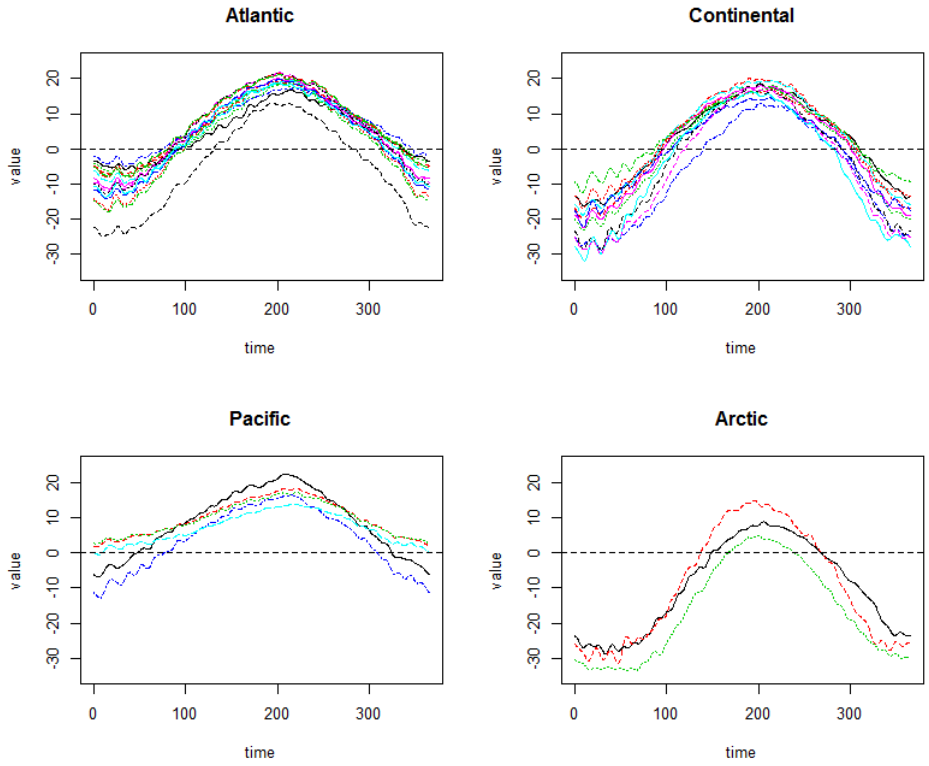


Figura 4.2 - Curvas de cada uma das estações climáticas do Canadá, divididas por região

Realizada a estimação, obtém-se a média geral e as estimativas para os efeitos, ilustrados na Figura 4.3. O primeiro gráfico mostra a média geral, seguido pelos efeitos referentes às regiões Atlântico, Continental, Pacífico e Ártico. No último gráfico, abaixo à direita, estão os quatro efeitos estimados somados à média geral.

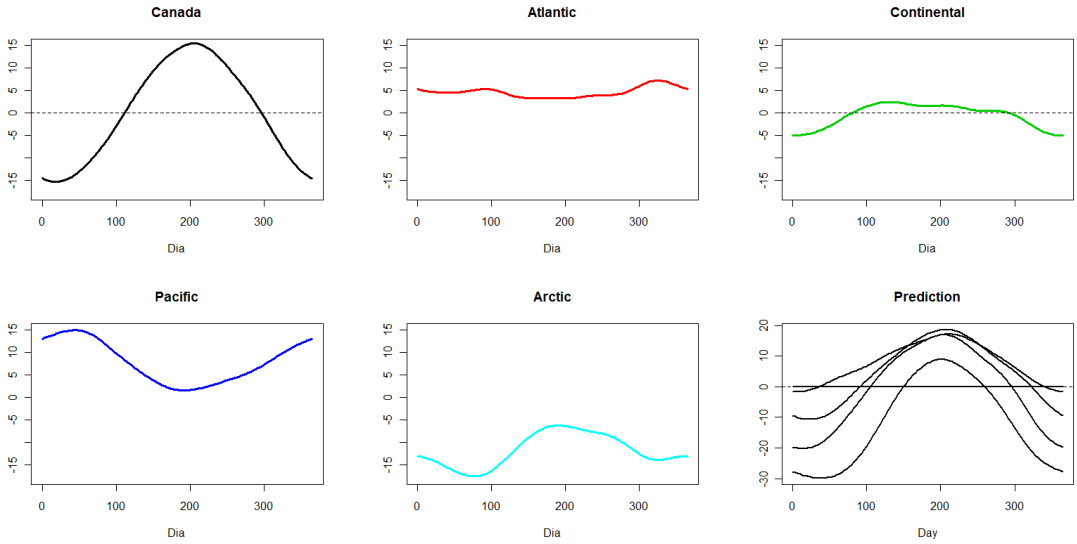


Figura 4.3 - No primeiro gráfico, a média geral, seguida pelos 4 efeitos estimados para cada região, no último as curvas de temperatura estimadas para cada região

Nota-se que, como esperado, a região do Ártico possui a temperatura anual mais baixa. Também se vê que no Pacífico as temperaturas são mais altas que a média geral. Os efeitos dos outros dois grupos oscilam menos. Na região Atlântico, as temperaturas são em média sempre 5º maiores que a média geral e na Continental são levemente mais quentes no inverno e levemente mais geladas no verão.

Podemos medir o ajuste do modelo calculando a estatística R^2 para cada instante de t, o resultado é a curva $R^2(t)$ na Figura 4.4:

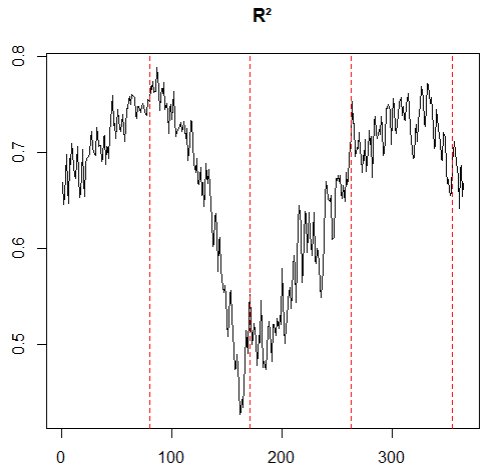


Figura 4.4 - $R^2(t)$ para o ajuste das modelos

Percebe-se que o valor do R^2 é maior durante a primavera e o outono, tem uma baixa pequena no inverno e uma grande queda no verão, onde, provavelmente, a

temperatura é mais variável dentro dos grupos, o que faria deflacionar a explicação que estes concedem sobre os dados.

4.2.3. Testes utilizando *bootstrap*

Assim como nos Intervalos de Confiança, podemos realizar testes funcionais utilizando *bootstrap*. Dada uma estatística escolhida, formamos novos grupos utilizando reamostragem e calcula-se ponto a ponto a estatística para cada um dos novos sorteios. Após muitas replicações, obtém-se uma distribuição dessa estatística, e, comparando-a com o valor resultante dos grupos originais da análise, decide-se se há ou não significância naquele ponto. Como exemplo, realizou-se um teste-F funcional para descobrir se há ou não diferença significativa entre a temperatura média das regiões do exemplo anterior. A estatística utilizada será:

$$F(t) = \frac{Var[\hat{y}(t)]}{\frac{1}{n}\sum (y_i(t) - \hat{y}_i(t))^2}. \tag{4.4}$$

É importante ressaltar que, como não foram supostas restrições quanto à distribuição de probabilidade ao longo das curvas, e também o fato de o *bootstrap* neste caso ter sido realizadao ponto a ponto, os resultados deste teste não podem ser interpretados ao longo de regiões ou por toda a curva. Ao ser executado o procedimento especificado anteriormente, obteve-se os resultados exibidos na Figura 4.5:

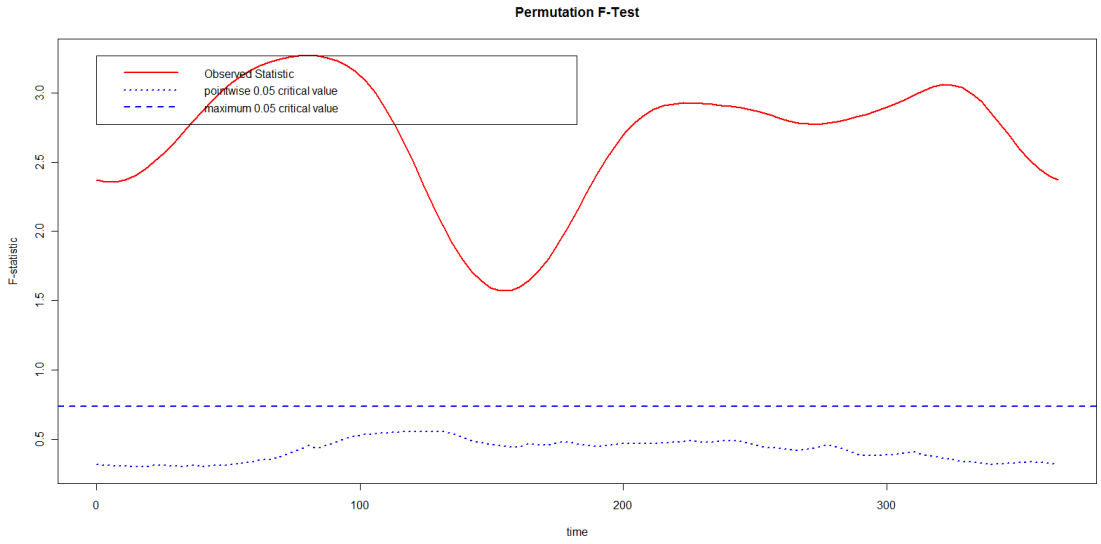


Figura 4.5 - Teste-F Funcional para a diferença entre as regiões

Na linha azul pontilhada, estão os valores críticos para 95% de confiança encontrados através do *bootstrap*, a linha tracejada indica o maior valor critico encontrado em todos os níveis. No gráfico, a estatística observada está acima da linha tracejada em toda a extensão do ano, mostrando assim, evidências estatísticas de que há efeito da região na temperatura nas estações climáticas no Canadá.

4.3. Resposta Funcional e Covariável Funcional

Neste caso mais complexo e mais geral, utilizaremos uma variável funcional para explicar outra variável funcional. Este modelo é chamado na literatura de *the concurrent model*, pelo fato de todo o domínio da variável explicativa influenciar em cada ponto da variável resposta. Dado dois conjuntos de curvas $y_1(t), y_2(t), \dots, y_n(t)$ e $x_1(t), x_2(t), \dots, x_n(t)$, tal modelo mais geral é da forma

$$Y_i(t) = \beta_0(t) + \int \beta_1(s,t) X_i(s) ds + \varepsilon_i(t). \quad (4.5)$$

Neste caso, β_1 é bivariado, ou seja, uma superfície nos domínios de Y e X. Para cada ponto em t há uma curva explicando a relação entre a função $Y(t)$ naquele ponto e $X(s)$ em todo seu domínio. De fato, o modelo apresentado na Seção 4.1 é um caso especial deste, onde o domínio de t é apenas um ponto e, portanto, β_1 apenas uma curva. Mais detalhes e especificações teóricas como processo de estimação e suposições do modelo podem ser encontradas em Chiou, Müller e Wang (2004).

4.3.1. Exemplo de aplicação

Desta vez utilizaremos as curvas de temperatura média diária ao longo do ano como variável explicativa para as curvas de precipitação ao longo do ano, ou seja, tanto a variável resposta quanto a variável indepente são funcionais. A Figura 4.6 exibe a superfície $\hat{\beta}_1$ estimada utilizando 7 funções bases de Fourier tanto para a base no eixo t quanto no eixo s.

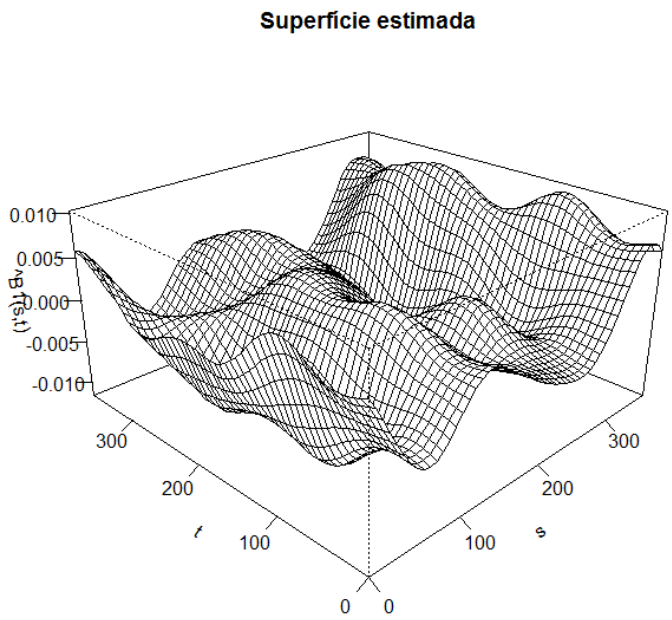


Figura 4.6 – Superfície $\widehat{\beta}_1$ estimada

É interessante notar que, ao contrário do que é intuitivo, a superfície sobre a reta $t = s$ não é saliente, ou seja, a variável explicativa, no caso a temperatura diária média, no próprio instante de tempo não explica mais sobre a precipitação. Uma possível explicação/interpretação poderia estar no fato de a que a precipitação poderia depender da temperatura do dia ou da semana anterior, ou até de uma combinação dessas.

5. Aplicação e Resultados

O problema constitui em encontrar diferença entre grupos de crianças em um instrumento de avaliação psicológica. Cada uma das crianças foi submetida a uma avaliação psiquiátrica que as separou em sete grupos:

0- Crianças Sem Transtornos
1- Crianças com Fobias
2- Crianças com Depressão
3- Crianças com Transtorno Déficit de Atenção – Hiperatividade (TDAH)
4- Crianças com Transtorno de Oposição e Desafio (TOD)

Os dados foram coletados através de um instrumento de pesquisa que mede o tempo de reação das crianças. Para cada criança eram exibidas setas para direita ou para

esquerda, e assim que aparecesse, o indivíduo deveria apertar o botão com a seta para a direção certa o mais rápido possível. Foram exibidas uma a uma 100 setas para cada criança, as setas apareciam a cada 1,5 segundos, tempo que as crianças tinham para responder ao estímulo, caso a reação demorasse mais que o tempo apontado, era computado como missing. No total, 785 crianças foram submetidas ao instrumento especificado. O objetivo é descobrir se há efeito dos grupos na média e também na variância do tempo de reação das crianças estudadas. Também é de interesse verificar se há crescimento ou decrescimento das médias e variâncias dos tempos de reações das crianças ao longo do tempo, uma das hipóteses do pesquisador é de que as crianças com TDAH ao passar do tendem a exibir variâncias maiores que as outras crianças.

5.1. Amostra

Dois outros grupos faziam parte do banco de dados, o primeiro com crianças com histórico de TDAH na família e o segundo com crianças diagnosticadas com TDAH e TOD. Por sugestão do pesquisador, tais grupos foram retirados da análise devido às suas possíveis altas correlações com o grupo controle e com o grupo TDAH, respectivamente. Como alguns dos métodos e rotinas utilizados não permitiam missings no banco de dados, estes foram substituídos por uma média ponderada dos dois trials anteriores e os dois *trials* posteriores. Duas crianças com muitos valores faltantes consecutivos não puderam ser completadas com o método descrito anteriormente e, portanto, foram também retiradas da análise. A amostra final possui 665 crianças de cinco grupos diferentes.

5.2. Suavização

Dado o comportamento dos dados, que é diferente para cada unidade amostral, e não exibe característica conhecida a prori, utilizamos o método de Regressão Polinomial Local via Kernel para a suavização, empregando o parâmetro h ótimo sugerido por Ruppert, Sheather e Wand (1995). Na Figura 5.1, os pontos são os 100 tempos de reação coletados de uma criança ao longo do tempo, a linha em preto é a função estimada:

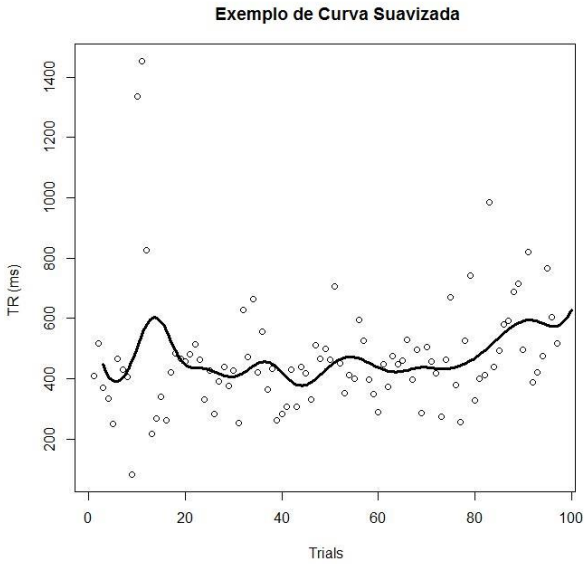


Figura 5.1 - Exemplo de suavização, os pontos são os tempos de reação registrados e a curva estimada

5.3. Análise das médias

Primeiramente calculou-se a média de cada grupo, para exploratória e intuitivamente percebermos se deve haver ou não diferenças entre os grupos. Na Figura 5.2, nota-se que a média das crianças com TOD (em rosa) é mais baixa que as outras, e que as crianças dos outros transtornos têm médias ligeiramente maiores que as do grupo controle (em preto).

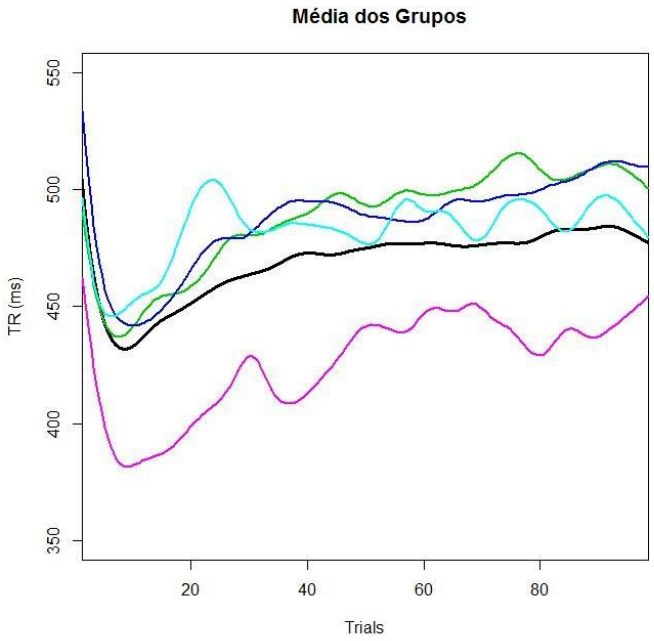


Figura 5.2- Gráfico com as médias dos grupos

Realizando através de bootstrap o teste-F, ilustradi na Figura 5.3, constatou-se que há diferença significativa em alguns períodos de tempo entre as médias dos grupos e que então devemos calcular o efeito de cada grupo para descobrirmos em quais grupos o efeito é significantemente diferente de zero.

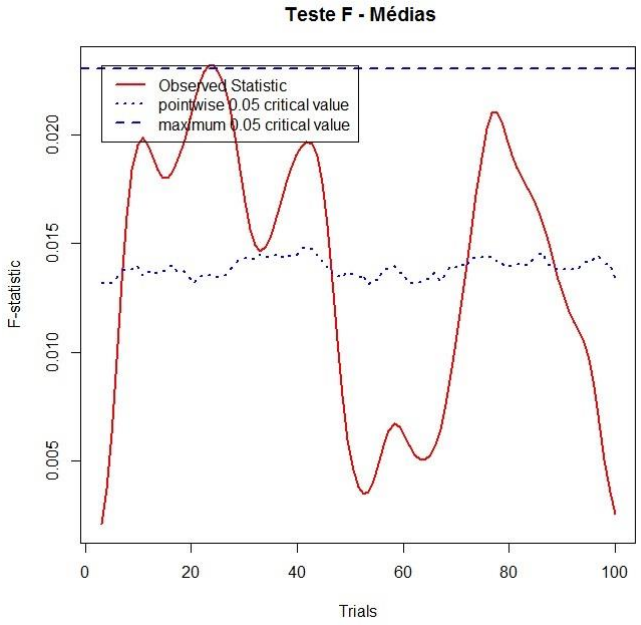


Figura 5.3 - Teste-F para o efeito na média

Para a estimativa dos efeitos utilizamos o modelo funcional citado anteriormente

$$y_{ij}(t) = \mu(t) + \alpha_i(t) + \varepsilon_{ij}(t). \quad (5.1)$$

portanto, tem-se uma curva de efeito estimada associada a cada grupo do estudo. Na Figura 5.4 encontram-se os intervalos com 95% de confiança associados a cada um dos grupos em questão. Mesmo com algumas pequenas áreas fugindo do zero nos grupos de crianças com Fobias e TDAH, é evidente que o efeito negativo no grupo das crianças com TOD é significantemente diferente de zero em grande parte do tempo de realização da tarefa.

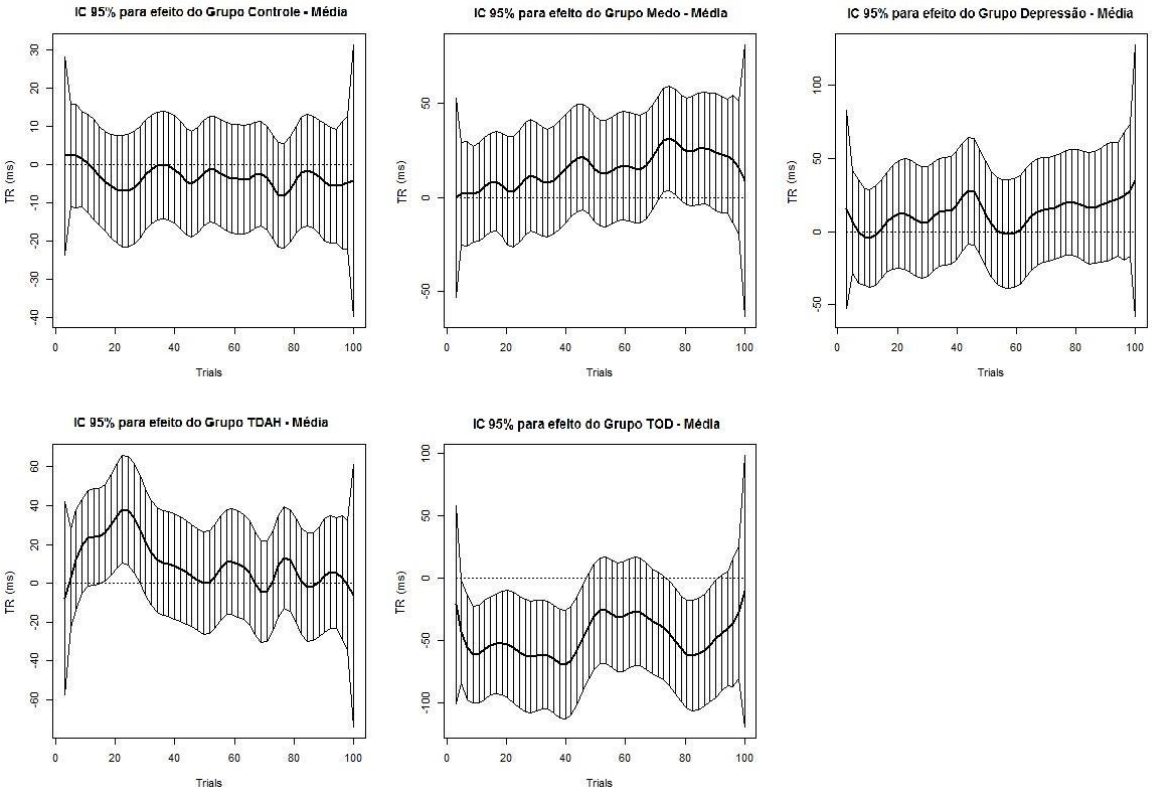


Figura 5.4- Intervalos de Confiança para os efeitos dos grupos na média geral

5.4. Análise das Variâncias

Para verificar se há diferença entre a variância dos grupos, foram removidas das curvas as médias dos seus respectivos grupos. Ou seja, das crianças sem transtorno foi subtraída a média do grupo das sem transtorno (Grupo 0). O mesmo procedimento foi realizado em cada grupo de curvas. Assim a média dos grupos foi reduzida a zero, e elevando ao quadrado os valores assim obtidos, adquirimos as curvas da variabilidade de cada criança. Ao realizar a média dos quadrados, obtemos a variância dos grupos. Desta maneira analisaremos se há efeito dos transtornos na variabilidade das crianças. Apanhadas

as curvas das variâncias individuais das crianças, as mesmas técnicas da análise das médias foram adotadas.

Na Figura 5.5 são exibidas as variâncias dos grupos, logo nota-se que as crianças com TDAH (em azul claro) possuem uma variabilidade quase sempre mais alta que as dos outros grupos, que se entrelaçam sugerindo que não devem ser diferentes umas das outras.

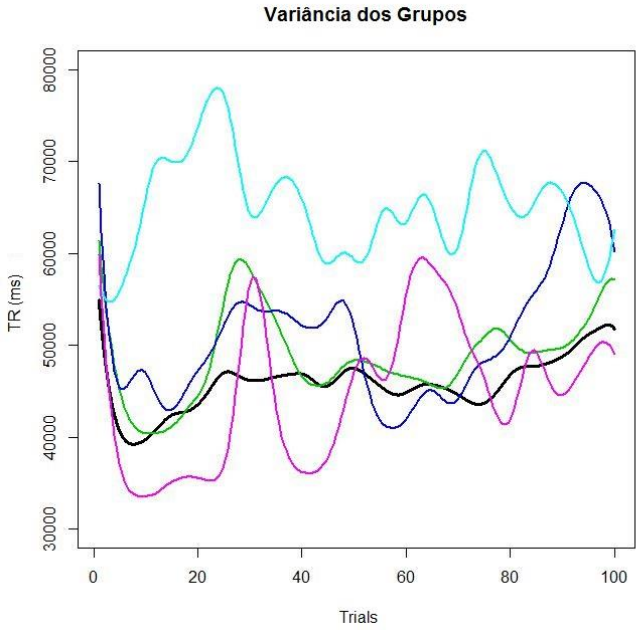


Figura 5.5- Variâncias dos grupos

O teste-F, na Figura 5.6, confirma a hipótese mencionada anteriormente. Salvas algumas regiões, as estatísticas são quase sempre significativas e em algumas partes até maiores que o ponto de corte mais conservador, na linha tracejada. Calculamos então os efeitos dos grupos nas médias das variâncias das unidades amostrais.

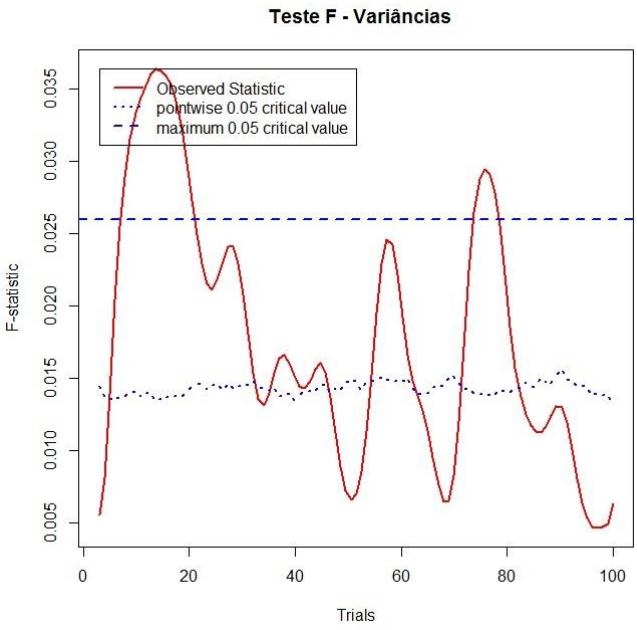


Figura 5.6 - Teste-F para efeito dos grupos na média das variâncias

Os intervalos para os efeitos confirmam novamente nossas intuições. Somente o efeito do grupo com TDAH, embaixo na esquerda, afeta positivamente a média das variâncias das crianças, ou seja, os dados evidenciam que as crianças com TDAH tenham uma variabilidade mais alta no instrumento aplicado.

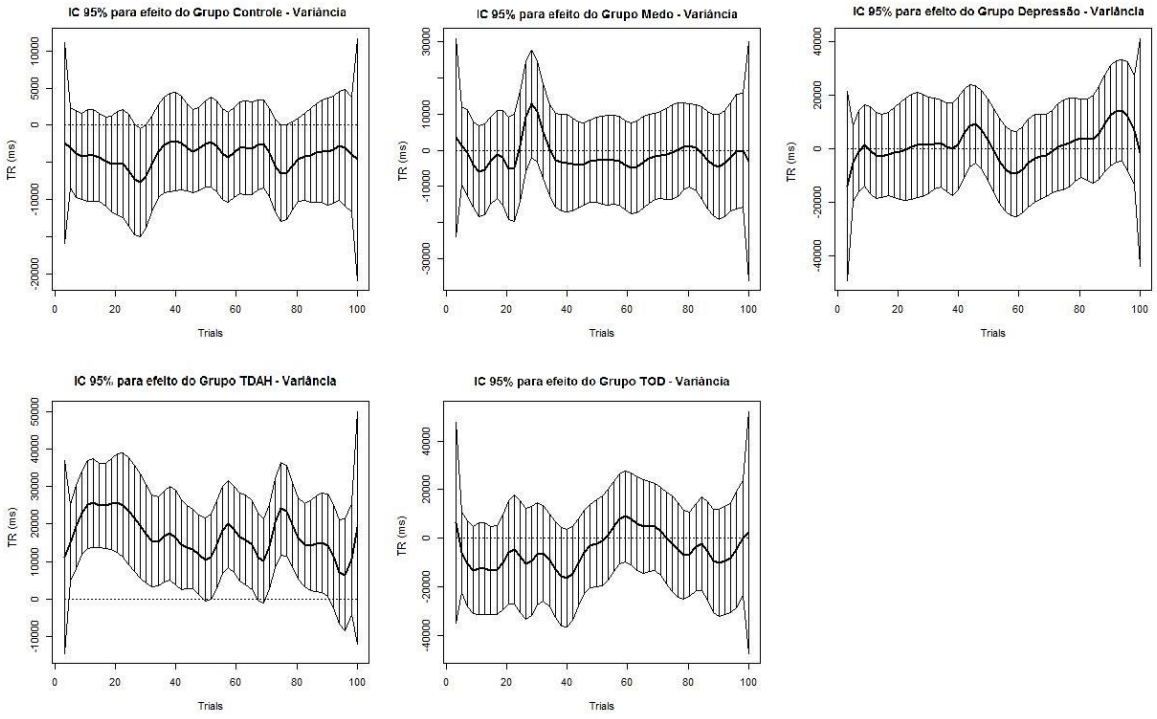


Figura 5.72 - Intervalos de Confiança para os efeitos dos grupos na variância média

6. Considerações Finais

Como exibimos e exemplificamos durante este trabalho, existem inúmeros problemas, aplicações e pontos a serem levados em consideração que surgem ao levantarmos o conceito de análise de dados funcionais. Tratando curvas como unidades amostrais, agregamos o problema da dimensionalidade infinita, sendo necessárias novas propostas para distribuição de tais elementos, assim como suas propriedades, suposições e testes de hipóteses. Porém, a matemática que envolve estas situações (conceitos de análise funcional, autofunções, espaços de Hilbert) é bastante complexa.

Sistema de Bases de Fourier, B-Splines e Regressão Polinomial Local via *Kernel* são métodos de suavização largamente utilizados. Outros problemas relacionados à suavização como seleção da quantidade de bases e seleção do parâmetro relacionado à penalidade também podem ser abordados.

Em relação aos modelos funcionais lineares, mostramos que têm alto poder explicativo, mas como citado em 4.1.1, devemos tomar cuidado com a quantidade de parâmetros na base que definiremos para os parâmetros funcionais a fim de evitar alguma espécie de *overfitting*.

Atualmente é pequeno o número de rotinas já programadas para a comparação entre parâmetros estimados do modelo de Análise de Variância Funcional. Uma sugestão de método envolvendo *bootstrap* e contrastes ortogonais para estimar se há diferença significativa entre efeitos estimados pode também ser encontrada em Ramsay e Silverman (2005, p. 233-235).

Exemplos de aplicações, análises ricas em detalhes e interpretações muito interessantes podem ser encontrados em Ramsay e Silverman (2002), é uma ótima primeira leitura para habituação com a ideia de dados funcionais e para obtenção de uma boa noção das capacidades que essa nova abordagem tem. Um manual de como aplicar Análise de Dados Funcionais utilizando o software *R* com a ajuda do pacote *fda* está contido em Ramsay, Hooker e Graves (2009). Avanços, aplicações mais recentes e novas metodologias são abordados em Ferraty (2011), Dabo-Niang e Ferraty (2008), Bathia, Yao e Ziegelmann (2010), entre outros.

Referências Bibliográficas

BATHIA, N.; YAO, Q.; ZIEGELMANN, F.; **Identifying the finite dimensionality of curve time series.** Annals of Statistics, 38:3352-3386, 2010

CARDOTA, H.; FERRATY, F. and SARDAB, P. **Functional Linear Model.** Statistics & Probability Letters 45: 11-22, 1999.

CHIOU, J.M., MULLER, H.G. WANG, J.L. **Functional Response Models** - Statistica Sinica, 2004.

DE BOOR, C. **A Practical Guide to Splines**, Springer-Verlag, New York, 1978.

DE BOOR, C. **A Practival Guide to Splines** (Revised Edition). Springer, 2001.

FAN, J. **Design-adaptative nonparametric regression**. Journal of the American Statistical Association, 87:998-1004, 1992.

FERRATY, F. **Recent Advances in Functional Data Analysis and Related Topics**. Springer, 2011.

GUO, W. **Functional Mixed Effects Model**, Biometrics, 58, 121-128, 2002.

HARDLE W.; MULLER M.; SPERLICH S. and WERTAZ A. **Nonparametric and Semiparametric Models**, Springer, 2004.

NADARAYA, E. A. **On estimating regression.** Theory of Probability and Its Applications, 10:186-190, 1964.

DABO-NIANG, S.; FERRATY, F. **Funcional and Operational Statistics,** Springer, 2008

RAMSAY, J.O. HOOKER, G. GRAVES, S. **Functional Data Analysis With R and MATLAB**, Springer, 2009.

RUPPERT, D.; SHEATHER, S. J. and WAND, M. P. **An Effective Bandwidth Selector for Local Least Squares Regression**. Journal of the American Statistical Association, 90: 1257-1270, 1995.

SILVERMAN, B.W. and RAMSAY, J.O. **Functional Data Analysis**, Springer, 2005.

SILVERMAN, B.W. and RAMSAY, J.O.
, Springer, 2002

SILVERMAN, B.W. **Density Estimation for Statistics and Data Analysis**. Chapman & Hall/CRC, 1998.

SPITZNER D. J., MARRON J. S. and ESSICK, G. K. **Mixed-Model Functional ANOVA for Studying Human Tactile Perception**. Journal of the American Statistical Association, Vol. 98, No. 462, 2003, pp. 263-272.

WANG, K. and GASSER, T. **Asymptotic and Bootstrap Confidence Bounds for the Structural Average of Curves**. The Annals of Statistics, 26:972-991, 1998.

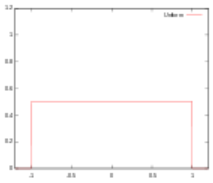
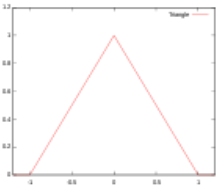
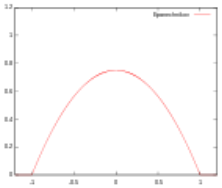
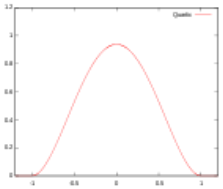
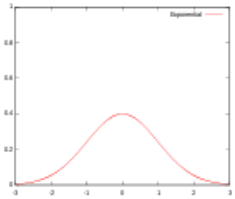
WATSON, G.S. **Smooth regression analysis**. Sankhya Series A, 26:101-116, 1964.

ZOGLAT, A. **Functional Analysis of Variance, Applied Mathematical Sciences**, Vol. 2, no. 23, 1115 - 1129, 2008.

Anexos

Tabela de Funções Kernel:

Funções Kernel, *K*(*u*)

Uniforme	**Triangular**
$K(u) = \frac{1}{2}\, I\{\lvert u\rvert \leq 1\}$	$K(u) = (1 - \lvert u\rvert)\, I\{\lvert U\rvert \leq 1\}$
	
Epanechnikov	**Quartic(biweight)**
$K(u) = \frac{3}{4}(1 - u^2)\, I\{\lvert u\rvert \leq 1\}$	$\boldsymbol{K}(u) = \frac{15}{16}(1 - u^2)^2\, I\{\lvert u\rvert \leq 1\}$
	
Normal	
$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	
	

Programa que gera o exemplo 3.1.1

```
library(fda)
#rotina para modelo escalar x funcional
#exemplo modelando tipo yi= B0 + int(B1(t)X(t)) + ei

#variável Resposta

annualprec = log10(apply(daily$precav,2,sum))
y=annualprec ##dados escalares

#suavizando
tempbasis=create.fourier.basis(c(0,365),65) #base funcional
yfunc=daily$tempav                          #matriz valores funcoes
xfunc=day.5                                 #valores da variável x que foram coletadas as funções
tempSmooth=smooth.basis(day.5,daily$tempav,tempbasis)      #suavizando
tempfd=tempSmooth$fd                        #pegando apenas o objeto funcional do resultado
da suavização

#organizando as variáveis regressoras
templist = vector ("list",2)                #montando uma lista com as variáveis regressoras
(constante + funcional) [dá pra botar mais covariáveis escalares e funcionais!!]
templist[[1]] = rep(1,35)                   #constante, 35 é o n
templist[[2]] = tempfd                      #funcional,

#organizando uma lista pras bases dos coeficientes a serem estimados
conbasis = create.constant.basis(c(0,365))          #base pra constante
betabasis= create.fourier.basis(c(0,365),9)         #base para o parâmetro funcional
betalist = vector("list",2)                 #declara a lista
betalist[[1]]=conbasis
betalist[[2]]=betabasis

#estimando o modelo
m1=fRegress(annualprec,templist,betalist)           #estima, essa função fRegress estima tudo
que for pra funcional, é tipo a lm dos dados funcionais
betaest=m1$betaestlist                      #adiciona os parâmetros estimados em uma lista
coef(betaest[[1]])                          #parâmetro B0 estimado
par(mfrow=c(1,2))
plot(betaest[[2]]$fd, ylab='B^1(t)',xlab='Dia do ano',main='Parâmetro funcional estimado')
#Parâmetro beta estimado
abline(v=c(80,171,263,355),lty=2,col=2)             #separando as estações do ano, para melhor
visualização

#ajuste do modelo
est=m1$yhatfdobj
plot(y,est,ylab='Esperado',xlab='Estimado',main='R²=0,87')                          #real x
estimado
abline(a=0,b=1,lwd=2)
res=y-est
sse1=sum(res^2)
sse0=sum((y-mean(y))^2)
rsq=(sse0-sse1)/sse0
```

Programa que gera o exemplo 3.2.2

```
library(fda)

#análise de variância funcional
x=day.5                                  # pontos onde foram coletados os valores das funções
y=daily$tempav                           # matriz com os valores coletados

#suavizando
tempbasis =create.fourier.basis(c(0,365),65)          # propoe a base
tempSmooth=smooth.basis(x,y,tempbasis)                # suaviza
tempfd =tempSmooth$fd                                 # salva o objeto funcional

#organizando os fatores
regions = unique(CanadianWeather$region)              # lista com os grupos
p = length(regions) + 1                               # quantidade de parametros
regionList = vector("list", p)                        # declarando 6 listas
regionList[[1]] = c(rep(1,35),0)                      # primeira referente ao parametro da média, último valor é 0 devido à condição de identificabilidade

for (j in 2:p) {                                      # coloca nas outras 5 listas 1 se está no fator, 0 se não, último valor é 1 devido à condição de identificabilidade
xj = CanadianWeather$region == regions[j-1]
regionList[[j]] = c(xj,1)                             ### importante, nesta etapa também podem ser adicionadas listas com variáveis explicativas quantitativas!!!
}

coef = tempfd$coef                                    # salva os coeficientes das bases das 35 curvas
coef36 = cbind(coef,matrix(0,65,1))                   # coloca mais um vetor de 0, adiciona um fator neutro, já que a matriz dos fatores tem uma linha a mais, para X'X ser quadrada!
temp36fd = fd(coef36,tempbasis,tempfd$fdnames)        # salva um novo fd com o 36o fator neutro

plot.fd(temp36fd,main='Curvas de Temperatura', xlab='Dia do Ano', ylab= '°C') #plota as temperaturas anuais


#organizando uma lista pras bases dos coeficientes a serem estimados

betabasis = create.fourier.basis(c(0, 365), 65, 365)     # propoe uma base pros estimadores, a mesma das funções
betafdPar = fdPar(betabasis)                          # coloca os parametros da base em uma lista
betaList = vector("list",p)
for (j in 1:p) betaList[[j]] = betafdPar


#estimando o modelo

m1 = fRegress(temp36fd, regionList,betaList)    # estima a média e os parâmetros relacionaos aos fatores
betaest = m1$betaestlist                 # adiciona as estimativas dos parâmetros a uma lista
est = m1$yhatfd                          # adiciona as curvas estimadas de cada regiâo a uma lista
```

```
regions = c("Canada", regions)                ###plotando
par(mfrow=c(2,3),cex=1)
for (j in 1:p)
plot(betaest[[j]]$fd, lwd=3,col=j,ylim=c(-18,15),xlab="Dia",ylab="", main=regions[j])
plot(est, lwd=2, col=1, lty=1,xlab="Day", ylab="",main="Prediction")

#avaliando ajuste do modelo, o R² calculado aqui será pointwise

estmat=eval.fd(day.5,est$fd)        # valores das estimativas das funções
estmat=estmat[,1:35]               # retirando a unidade 0 adicionada pra estimação
resmat=y-estmat                   # resíduos

ym=rep(0,365)
for (i in 1:365)
ym[i]=mean(y[i,])
resmat0=y-ym%*%matrix(1,1,35)

sse0=apply((resmat0)^2,1,sum)
sse1=apply(resmat^2,1,sum)
rsq=(sse0-sse1)/(sse0)
plot(rsq,type='l',main='R²')

F.res = Fperm.fd(temp36fd, regionList, betaList)         # teste F
```

Programa que gera o exemplo 3.3.1

```
library(fda)

#organizando os dados
#variável funcional dependente

t=day.5    #valores onde foram avaliadas as funções
y=log(daily$precav) #valores coletados
base=create.fourier.basis(c(0,365),33)   #propondo a base
yfd=smooth.basis(t,y,base)$fd

#variável funcional independente

x=daily$tempav                    #valores coletados de x
xfd=smooth.basis(t,x,base)$fd          #propondo a mesma base

#definindo lista com 2 bases, de B0, e de B1, que é bivariada!

u=7                                  #quantidade de bases em B_1
baseest=create.fourier.basis(c(0,365),u)         #base r2
p=matrix(1:(u*u),ncol=u,nrow=u)                #definindo uma matriz para alocação dos coeficientes
de B_1
base2=bifd(p,baseest,baseest)                  #colocando em um funcional bivariado a base de B_1
beta0Par=fdPar(base,2,0)                      #definindo a base para B_0
beta1Par=bifdPar(base2,2,0)                    #definindo a base para B_1 gerada anteriormente
betalist=list(beta0Par,beta1Par)               #lista com as duas bases

#estimando o modelo
m1=linmod(xfd,yfd,betalist)                    #estima o modelo

b0=m1$beta0estfd                            #acessando B_0
b1=m1$beta1estbifd                           #acessando B_1
est=m1$yhatfdobj                            #acessando valores estimados pras funções

q=eval.bifd(teval,teval,b1)                    # esses três comandos realizam o plot
teval=seq(0,365,length=53)
persp(teval,teval,q,theta=-
45,phi=25,r=3,expand=.5,ticktype='detailed',xlab='s',ylab='t',zlab='^B1(s,t)',main='Superfície
estimada')

estmat=eval.fd(day.5,est)          # valores das estimativas das funções
resmat=y-estmat                   # resíduos

ym=rep(0,365)                      #calculando o R²
for (i in 1:365)
ym[i]=mean(y[i,])
resmat0=y-ym%*%matrix(1,1,35)

sse0=apply((resmat0)^2,1,sum)
sse1=apply(resmat^2,1,sum)
rsq=(sse0-sse1)/(sse0)
plot(rsq,type='l',main='R²')
```