TÓPICOS EM ECONOMETRIA NÃO-PARAMÉTRICA

Victor Rodrigues de Oliveira

Programa de Pós-Graduação em Desenvolvimento Econômico

Universidade Federal do Paraná

E–mail: victor5491@gmail.com

Sumário

1. Prolegômenos

1.1 Classes de Diferenciabilidade

O conjunto de funções $\mathcal{C}^0$ é a primeira classe de diferenciabilidade que veremos. Como o nome diferenciabilidade pode sugerir, nas classes de diferenciabilidade estaremos lidando com as funções derivadas parciais. Além disso, precisaremos que essas derivadas parciais sejam contínuas.

Definição 1.1.1 (Função de Classe $\mathcal{C}^0$)**.** *Uma função $f: X \subset \mathbb{R}^2 \to \mathbb{R}$ é dita ser de classe $\mathcal{C}^0$ (ou simplesmente dita ser $\mathcal{C}^0$) se é contínua em todos os pontos do seu domínio.*

Definição 1.1.2 (Função de Classe $\mathcal{C}^1$)**.** *Uma função $f: X \subset \mathbb{R}^2 \to \mathbb{R}$, com derivadas parciais $f_x(x, y)$ e $f_y(x, y)$, é dita ser de classe $\mathcal{C}^1$ (ou simplesmente dita ser $\mathcal{C}^1$) se é contínua em todos os pontos do seu domínio e f_x e f_y são contínuas em X.*

O algoritmo abaixo dá todos os passos para verificarmos se uma função f dada é de classe $\mathcal{C}^1$ ou não. Uma observação se faz necessária antes de prosseguirmos: o processo de determinação é longo e confuso, porém não é dos mais difíceis. É recomendado seguir por partes seguindo os exatos passos abaixo.

1. f é contínua?
   (a) Se sim, calcule as derivadas parciais. As derivadas parciais são contínuas? Se sim, f é $\mathcal{C}^1$.
   (b) Se sim, calcule as derivadas parciais. As derivadas parciais são contínuas? Se não, f não é $\mathcal{C}^0$.
   (c) Se não, f não é $\mathcal{C}^0$ e f não é $\mathcal{C}^1$.

Primeiro verificamos a continuidade de f, para depois calcularmos as derivadas parciais e verificar a continuidade das mesmas.

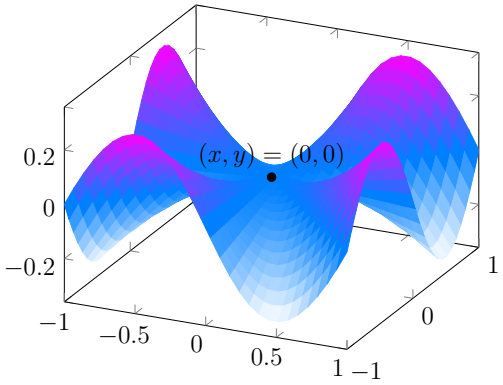
Definição 1.1.3 (Função de Classe $\mathcal{C}^k$)**.** *Uma função $f: X \subset \mathbb{R}^2 \to \mathbb{R}$, $k \in \mathbb{N}$, é dita ser de classe $\mathcal{C}^k$ (ou simplesmente dita ser $\mathcal{C}^k$) se é contínua em todos os pontos do seu domínio e possui todas as derivadas até ordem k contínuas. No caso em que a função é de classe $\mathcal{C}^k$ para qualquer $k \in \mathbb{N}$, dizemos que a função é de classe $\mathcal{C}^\infty$.*

Verifique se a função

$$\begin{cases} \dfrac{xy^3 - x^3y}{x^2 + y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0) \end{cases} \tag{1.1.1}$$

é de classe $\mathcal{C}^1$.

Figura 1.1.1 – Gráfico da Função $\dfrac{xy^3 - x^3y}{x^2 + y^2}$



Observemos que o domínio de f é $X = \mathbb{R}^2$, de forma que o primeiro passo é determinar se a função f é contínua em X.

1. f é contínua em $X = \mathbb{R}^2$?

   Observemos agora que a função f é dada por duas partes: uma fora da origem $(0, 0)$ e $f(0, 0) = 0$. Desta forma, precisamos analisar a continuidade da função em $(x, y) \neq (0, 0)$ e $(x, y) = (0, 0)$. Separemos os dois casos:

   (a) $(x, y) \neq (0, 0)$

   Esse geralmente é o caso mais simples. Isso se dá porque em $(x, y) \neq (0, 0)$ a função é dada por $\dfrac{xy^3 - x^3y}{x^2 + y^2}$ e, neste caso, temos divisão de polinômios. Como polinômios são contínuos e divisão de funções

contínuas é uma função contínua, temos f contínua desde que $x^2+y^2 \neq 0$, ou seja, $x(,y) \neq (0,0)$.

(b) $(x,y) = (0,0)$

Lembremos que, a fim de determinar a continuidade de uma função num determinado ponto, a definição de continuidade nos diz que devemos observar o comportamento da função ao redor do ponto em questão e esse comportamento também deve ser igual ao valor da função no ponto. No nosso caso em particular, como estamos analisando o ponto $(x,y) = (0,0)$, a fim de que f seja contínua no ponto devemos mostrar que

$$\lim_{(x,y)\to(0,0)} f(x,y) = \lim_{(x,y)\to(0,0)} \frac{xy^3 - x^3y}{x^2+y^2} = 0 = f(0,0) \tag{1.1.2}$$

Desta forma, devemos analisar o limite $\lim_{(x,y)\to(0,0)} \dfrac{xy^3 - x^3y}{x^2+y^2}$.

Observe inicialmente que a substituição de (x,y) por $(0,0)$ nos dá uma indeterminação. Para lidar com o limite, escreva

$$\begin{aligned} \lim_{(x,y)\to(0,0)} \frac{xy^3 - x^3y}{x^2+y^2} &= \lim_{(x,y)\to(0,0)} \left(\frac{xy^3}{x^2+y^2} - \frac{x^3y}{x^2+y^2} \right) \\ &= \lim_{(x,y)\to(0,0)} xy\frac{y^2}{x^2+y^2} - \lim_{(x,y)\to(0,0)} xy\frac{x^2}{x^2+y^2} \end{aligned} \tag{1.1.3}$$

Observe que as funções $\dfrac{y^2}{x^2+y^2}$ e $\dfrac{x^2}{x^2+y^2}$ são limitas e $\lim_{(x,y)\to(0,0)} xy = 0$. Em matemática, uma função é dita limitada se sua imagem é um conjunto limitado. Analogamente, dizemos que uma função é ilimitada quando ela não é limitada. Uma função real $f\colon D \to \mathbb{R}$ é limitada se existe uma constante $M \geq 0$ tal que $|f(x)| \leq M, \ \ \forall x \in D$.

Desta forma, temos

$$\lim_{(x,y)\to(0,0)} xy\frac{y^2}{x^2+y^2} = 0 \tag{1.1.4}$$

$$\lim_{(x,y)\to(0,0)} xy\frac{x^2}{x^2+y^2} = 0 \tag{1.1.5}$$

Assim,

$$\lim_{(x,y)\to(0,0)} \frac{xy^3 - x^3y}{x^2+y^2} = \lim_{(x,y)\to(0,0)} xy\frac{y^2}{x^2+y^2} - \lim_{(x,y)\to(0,0)} xy\frac{x^2}{x^2+y^2} = 0 \tag{1.1.6}$$

Desta forma, concluímos que

$$\lim_{(x,y)\to(0,0)} f(x,y) = \lim_{(x,y)\to(0,0)} \frac{xy^3 - x^3y}{x^2+y^2} = 0 = f(0,0) \tag{1.1.7}$$

e a função f é contínua em $(0,0)$. Como ela é contínua em $(0,0)$ e $(x,y) \neq (0,0)$, f é contínua em $X = \mathbb{R}^2$.

2. Calcule as derivadas parciais.

   Como as derivadas parciais são funções definidas em pontos do domínio de f, devemos calculá-las em todos os pontos. Por mais que pareça uma tarefa impossível, o procedimento é muito parecido com o da continuidade: temos os casos $(x,y) \neq (0,0)$ e $(x,y) = (0,0)$. Separemos novamente os dois casos:

   (a) $(x,y) \neq (0,0)$

   Para $(x,y) \neq (0,0)$, a função é dada por $f(x,y) = \dfrac{xy^3 - x^3y}{x^2+y^2}$. Neste caso, as derivadas parciais são calculadas diretamente pela regra do quociente:

$$f_x(x,y) = \frac{(y^3 - 3x^2y)(x^2+y^2) - (xy^3 - x^3y)2x}{(x^2+y^2)^2} = \frac{y^5 - x^4y - 4x^2y^3}{(x^2+y^2)^2} \tag{1.1.8}$$

$$f_y(x,y) = \frac{(3xy^2 - x^3)(x^2+y^2) - (xy^3 - x^3y)2y}{(x^2+y^2)^2} = \frac{-x^5 + x^4y + 4x^3y^2}{(x^2+y^2)^2} \tag{1.1.9}$$

   (b) $(x,y) = (0,0)$

   Observe que as derivadas parciais obtidas no item anterior não estão definidas em $(0,0)$. Desta forma, devemos utilizar a definição de derivada parcial no ponto para determinar $f_x(0,0)$ e $f_y(0,0)$:

$$f_x(0,0) = \lim_{h\to 0} \frac{f(0+h,0) - f(0,0)}{h} = \lim_{h\to 0} \frac{f(h,0) - 0}{h} = \frac{f(h,0)}{h} \tag{1.1.10}$$

$$f_y(0,0) = \lim_{h\to 0} \frac{f(0,0+h) - f(0,0)}{h} = \lim_{h\to 0} \frac{f(0,h) - 0}{h} = \frac{f(0,h)}{h} \tag{1.1.11}$$

Observe que como $h \to 0$, temos $h \neq 0$ bem próximo de 0. Desta forma, $f(h,0)$ e $f(0,h)$ devem ser calculados fazendo $(x,y) = (h,0)$ e $(x,y) = (0,h)$:

$$f(h,0) = \frac{h\cdot 0 - h^3 \cdot 0}{h^2 + 0^2} = 0 \tag{1.1.12}$$

$$f(0,h) = \frac{0\cdot h^3 - 0 \cdot h}{0^2 + h^2} = 0 \tag{1.1.13}$$

Substituindo nas expressões para as derivadas no ponto $(0,0)$, encontramos

$$f_x(0,0) = \lim_{h\to 0} \frac{0}{h} = \lim_{h\to 0} 0 = 0 \tag{1.1.14}$$

$$f_y(0,0) = \lim_{h\to 0} \frac{0}{h} = \lim_{h\to 0} 0 = 0 \tag{1.1.15}$$

Juntando os itens (a) e (b), encontramos que as funções derivadas parciais são dadas por

$$f_x(x,y) = \begin{cases} \dfrac{y^5 - x^4y - 4x^2y^3}{(x^2+y^2)^2}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0) \end{cases} \tag{1.1.16}$$

$$f_y(x,y) = \begin{cases} \dfrac{-x^5 + x^4y + 4x^3y^2}{(x^2+y^2)^2}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0) \end{cases} \tag{1.1.17}$$

O próximo passo é determinar se as derivadas parciais são contínuas.

3. As derivadas parciais são contínuas?

   No item anterior obtivemos duas funções derivadas parciais. Precisamos verificar, de maneira análoga ao item 1, se cada uma das duas derivadas é contínua.

   (a) Verifiquemos se f_x é contínua.

   Em $(x, y) \neq (0, 0)$ temos que f_x é contínua por ser divisão de polinômios.

   No ponto $(0, 0)$, precisamos mostrar que

$$\lim_{(x,y)\to(0,0)} f_x(x, y) = 0 = f_x(0, 0) \tag{1.1.18}$$

   A fim de calcular o limite, escreva

$$\begin{aligned}
\lim_{(x,y)\to(0,0)} f_x(x, y) &= \lim_{(x,y)\to(0,0)} \frac{y^5 - x^4y - 4x^2y^3}{(x^2 + y^2)^2} \\
&= \lim_{(x,y)\to(0,0)} \frac{y^5}{(x^2 + y^2)^2} - \lim_{(x,y)\to(0,0)} \frac{x^4y}{(x^2 + y^2)^2} \\
&- 4 \lim_{(x,y)\to(0,0)} \frac{x^2y^3}{(x^2 + y^2)^2} \\
&= \lim_{(x,y)\to(0,0)} y\frac{y^4}{(x^2 + y^2)^2} - \lim_{(x,y)\to(0,0)} y\frac{x^4}{(x^2 + y^2)^2} \\
&- 4 \lim_{(x,y)\to(0,0)} y\frac{x^2y^2}{(x^2 + y^2)^2}
\end{aligned} \tag{1.1.19}$$

   Observe que $\lim_{(x,y)\to(0,0)} y = 0$. Suspeitamos que a função $\dfrac{y^4}{(x^2 + y^2)^2}$ seja limitada. Lembre que $\dfrac{y^2}{x^2 + y^2}$ é limitada, isto é $0 \leq \dfrac{y^2}{x^2 + y^2} \leq 1$. Elevando as três parcelas ao quadrado, temos:

$$0 \leq \frac{y^4}{(x^2 + y^2)^2} \leq 1 \tag{1.1.20}$$

   Desta forma, a função acima é limitada. Assim,

$$\lim_{(x,y)\to(0,0)} y\frac{y^4}{(x^2+y^2)^2} = 0 \tag{1.1.21}$$

De maneira análoga, $\dfrac{x^4}{(x^2+y^2)^2}$ é limitada. Portanto,

$$\lim_{(x,y)\to(0,0)} y\frac{x^4}{(x^2+y^2)^2} = 0 \tag{1.1.22}$$

Agora temos que verificar se $\dfrac{x^2y^2}{(x^2+y^2)^2}$ é limitada:

$$0 \leq x^2y^2 \leq 2x^2y^2 \leq x^4 + 2x^2y^2 + y^4 = (x^2+y^2)^2 \tag{1.1.23}$$

Após dividir os extremos por $(x^2+y^2)^2$, temos:

$$0 \leq \frac{x^2y^2}{(x^2+y^2)^2} \leq 1 \tag{1.1.24}$$

e a função $\dfrac{x^2y^2}{(x^2+y^2)^2}$ é limitada. Assim,

$$4\lim_{(x,y)\to(0,0)} y\frac{x^2y^2}{(x^2+y^2)^2} = 0 \tag{1.1.25}$$

Logo,

$$\begin{aligned}\lim_{(x,y)\to(0,0)} f_x(x,y) = &\lim_{(x,y)\to(0,0)} y\frac{y^4}{(x^2+y^2)^2} - \lim_{(x,y)\to(0,0)} y\frac{x^4}{(x^2+y^2)^2} \\ &- 4\lim_{(x,y)\to(0,0)} y\frac{x^2y^2}{(x^2+y^2)^2} = 0\end{aligned} \tag{1.1.26}$$

de onde segue que f_x é contínua em $(0,0)$ e, portanto, a função f_x é contínua em $\mathbb{R}^2$.

(b) A função f_y é contínua seguindo exatamente o mesmo raciocínio e os

passos exatos são deixados a cargo do leitor. Recomenda-se, todavia, repetir o procedimento para melhor entendimento.

Portanto, como as derivadas parciais são contínuas, concluímos que f é de classe $\mathcal{C}^1$.

Uma das classes de diferenciabilidade mais especiais é a $\mathcal{C}^2$, a classe de funções contínuas, de derivadas parciais contínuas e segundas derivadas parciais contínuas. Isso se deve por causa de um resultado bastante importante que nos diz que funções de classe $\mathcal{C}^2$ 2 possuem derivadas mistas iguais.

Teorema 1.1.1 (Teorema de Clairaut-Schwarz)**.** *Seja $f\colon X \subset \mathbb{R}^2 \to \mathbb{R}$ uma função de classe $\mathcal{C}^2$. Então, para qualquer $(x_0, y_0) \in X$, temos*

$$\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) = \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0) \tag{1.1.27}$$

Em termos mais explícitos, o Teorema de Clairaut-Schwarz (também conhecido como Teorema de Clairaut ou Teorema de Schwarz) diz que se uma função $f(x, y)$ é de classe $\mathcal{C}^2$, então suas derivadas mistas f_{xy} e f_{yx} devem ser iguais. Observemos aqui que o resultado continua válido para uma função de três ou mais variáveis. Por exemplo, se $f(x, y, z)$ é de classe $\mathcal{C}^2$, então as derivadas mistas devem satisfazer:

$$f_{xy} = f_{yx} \tag{1.1.28}$$
$$f_{xz} = f_{zx} \tag{1.1.29}$$
$$f_{yz} = f_{zy} \tag{1.1.30}$$

Uma aplicação importante do Teorema de Clairaut-Schwarz é a sua negativa: se $f_{xy}(x_0, y_0) \neq f_{yx}(x_0, y_0) \Rightarrow$ f não é de classe $\mathcal{C}^2$.

Verifique se a função

$$\begin{cases} \dfrac{xy^3 - x^3y}{x^2 + y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0) \end{cases} \tag{1.1.31}$$

é de classe $\mathcal{C}^2$.

As funções derivadas parciais são dadas por

$$f_x(x,y) = \begin{cases} \dfrac{y^5 - x^4y - 4x^2y^3}{(x^2+y^2)^2}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0) \end{cases} \tag{1.1.32}$$

$$f_y(x,y) = \begin{cases} \dfrac{-x^5 + x^4y + 4x^3y^2}{(x^2+y^2)^2}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0) \end{cases} \tag{1.1.33}$$

Desta forma, para verificar que f é de classe $\mathcal{C}^2$, devemos calcular as derivadas parciais de cada derivada parcial, ou seja, devemos calcular f_{xx}, f_{xy}, f_{yx} e f_{yy} e verificar se todas as quatro são contínuas. Comecemos pelas derivadas mistas f_{xy} e f_{yx} . Pelo Teorema de Clairaut-Schwarz, a fim de que f seja de classe $\mathcal{C}^2$ devemos ter $f_{xy}(x,y) = f_{yx}(x,y)$ e, em particular, $f_{xy}(0,0) = f_{yx}(0,0)$. Comecemos pelas derivadas mistas no ponto $(0,0)$:

$$\begin{aligned} f_{xy}(0,0) &= \frac{\partial f_x}{\partial y}(0,0) = \lim_{h\to 0} \frac{f_x(0,0+h) - f_x(0,0)}{h} = \lim_{h\to 0} \frac{f_x(0,h) - 0}{h} \\ &= \lim_{h\to 0} \frac{f_x(0,h)}{h} \end{aligned} \tag{1.1.34}$$

$$\begin{aligned} f_{yx}(0,0) &= \frac{\partial f_y}{\partial x}(0,0) = \lim_{h\to 0} \frac{f_y(0+h,0) - f_y(0,0)}{h} = \lim_{h\to 0} \frac{f_y(h,0) - 0}{h} \\ &= \lim_{h\to 0} \frac{f_y(h,0)}{h} \end{aligned} \tag{1.1.35}$$

Observe que como $h \to 0$, h é um número diferente de zero suficientemente próximo dele. Assim

$$f_x(0,h) = \frac{h^5 - 0^4 \cdot y - 4 \cdot 0^2 \cdot h^3}{(0^2+h^2)^2} = h \tag{1.1.36}$$

$$f_y(h,0) = \frac{-h^5 + h \cdot 0 + 4 \cdot h^3 \cdot 0^2}{(0^2+h^2)^2} = -h \tag{1.1.37}$$

Substituindo nas expressões para $f_{xy}(0,0)$ e $f_{yx}(0,0)$, obtemos

$$f_{xy}(0,0) = \lim_{h\to 0} \frac{f_x(0,h)}{h} = \lim_{h\to 0} \frac{h}{h} = \lim_{h\to 0} 1 = 1 \tag{1.1.38}$$

$$f_{yx}(0,0) = \lim_{h\to 0} \frac{f_y(h,0)}{h} = \lim_{h\to 0} -\frac{h}{h} = \lim_{h\to 0} -1 = -1 \tag{1.1.39}$$

As derivadas parciais mistas em $(0,0)$ não são iguais. Pela negação do Teorema de Clairaut-Schwarz, f não é de classe $\mathcal{C}^2$.

1.2 Expansão de Taylor

A linha tangente se aproxima de $f(x)$ e fornece uma boa aproximação perto do ponto de tangência x_0. À medida que você se afasta de x_0, no entanto, a aproximação fica menos precisa. Da definição de derivada segue que:

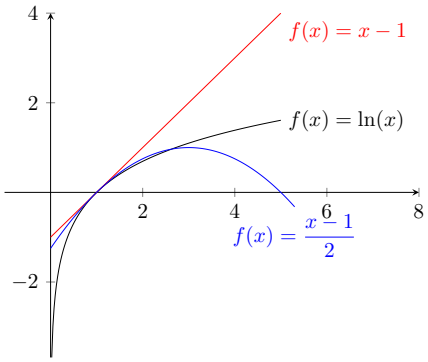
$$f(x) \approx f(x_0) + f'(x_0)(x - x_0). \tag{1.2.1}$$

Exemplo 1.2.1. *Seja $f(x) = \ln x$ no ponto $x_0 = 1$. Suponha que queremos aproximar essa função por meio da derivada de primeira ordem. Então,*

$$\begin{aligned} f(x) &\approx f(x_0) + f'(x_0)(x - x_0) \\ &\approx \ln(1) + \frac{1}{x}\Big|_{x=1} (x-1) \\ &\approx x - 1. \end{aligned} \tag{1.2.2}$$

Logo, $\ln(x) \approx x - 1$. Graficamente, temos:

Figura 1.2.1 – Gráfico da Função $f(x) = \ln x$



Às vezes precisamos de aproximações mais acuradas. A aproximação quadrática, que permite um melhor ajuste a uma função, é apresentada a seguir:

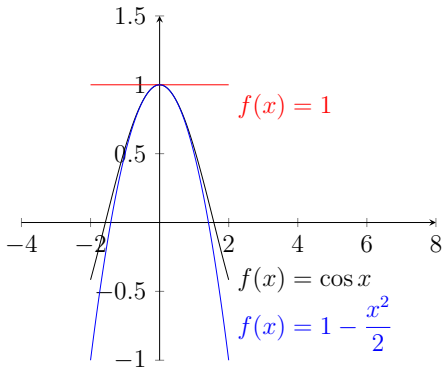
$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x)}{2}(x - x_0)^2. \tag{1.2.3}$$

Exemplo 1.2.2. *Assim, se quiséssemos aproximar a função* $f(x) = \cos x$ *em torno de* $x_0 = 0$ *por meio da expansão de Taylor, obteríamos:*

$$\begin{aligned} f(x) &\approx \cos(0) - \sin(0)(x - 0) - \frac{\cos(0)}{2}(x - 0)^2 \\ &\approx 1 - 0(x) - \frac{1}{2}x^2 \\ &\approx 1 - \frac{1}{2}x^2. \end{aligned} \tag{1.2.4}$$

Graficamente, temos:

Figura 1.2.2 – Gráfico da Função $f(x) = \cos x$



Generalizando, em matemática, uma série de Taylor é a série de funções da forma:

$$f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n \quad \text{sendo} \quad a_n = \frac{f^{(n)}(a)}{n!}, \tag{1.2.5}$$

em que $f(x)$ é uma função analítica[1] dada. Neste caso, a série acima é dita ser a série de Taylor de $f(x)$ em torno do ponto $x = a$. Associadamente, o polinômio de Taylor de ordem n em torno de $x = a$ de uma dada função n-vezes diferenciável neste ponto é dado por

$$p(x) = f(a) + f'(a)\frac{(x-a)^1}{1!} + f''(a)\frac{(x-a)^2}{2!} + \ldots + f^{(n)}(a)\frac{(x-a)^n}{n!}. \tag{1.2.6}$$

1.3 Relações de Ordem e Taxas de Convergência

Definição 1.3.1. *Duas sequências a_n e b_n são assintoticamente equivalentes se $\frac{a_n}{b_n} \to 1$. Essa relação afirma que, para n grande, os números das duas sequências são aproximadamente iguais.*

[1] Uma função analítica é uma função que é dada localmente por uma série de potências convergentes.

Vamos considerar três outras relações entre sequências, denotadas por o, $\asymp$ e O, que correspondem a a_n ser de ordem menor, igual, ou menor ou igual a b_n, respectivamente.

Definição 1.3.2. *Dizemos que* $a_n = o(b_n)$ *se* $\dfrac{a_n}{b_n} \to 0$ *quando* $n \to \infty$. *Quando* a_n *e* b_n *tendem para infinito, isso indica que* a_n *tende para o infinito mais lentamente do que* b_n*; quando ambos tendem para* 0*, ele afirma que* a_n *tende para zero mais rápido do que* b_n.

Exemplo 1.3.1. *Veja que* $\dfrac{1}{n^2} = o\left(\dfrac{1}{n}\right)$ *dado que* $\dfrac{1/n^2}{1/n} = \dfrac{n}{n^2} \to 0$.

Exemplo 1.3.2. *Suponha que*

$$a_n = \frac{1}{n} - \frac{2}{n^2} + \frac{4}{n^3} \tag{1.3.1}$$

Podemos escrever $a_n = \dfrac{1}{n} + R_n$*, em que* $R_n = -\dfrac{2}{n^2} + \dfrac{4}{n^3} = o\left(\dfrac{1}{n}\right)$. *Isso significa que* $o\left(\dfrac{1}{n}\right)$ *denota qualquer quantidade que tende a* 0 *mais rápido do que* $\dfrac{1}{n}$.

Podemos depreender as seguintes propriedades:

1. $a_n = o(1)$ se $a_n \to 0$ quando $n \to \infty$
2. $a_n = o(b_n) \implies a_n = b_n o(1) \implies \dfrac{a_n}{b_n} = o(1)$
3. se $a_n \to 0$ e $b_n \to 0$, $a_n = o(b_n) \implies \dfrac{a_n}{b_n} \to 0$ (logo, a_n vai mais rápido para zero do que b_n)
4. se $a_n = o(b_n)$ e $b_n = o(c_n)$, então, $a_n = o(c_n)$.
5. se $a_n = o(b_n)$, então $ca_n = o(b_n)$.
6. para qualquer $c_n \neq 0$ e $a_n = o(b_n)$, temos que $c_n a_n = o(c_n b_n)$.
7. se $a_n = o(b_n)$ e $c_n = o(d_n)$, então:

   (a)

$$\begin{aligned} a_n c_n &= o(b_n)o(d_n) \\ &= b_n d_n o(1) \\ &= o(b_n d_n) \end{aligned} \tag{1.3.2}$$

(b)

$$|a_n|^s = o\left(|b_n|^s\right),\ s > 0 \tag{1.3.3}$$

(c)

$$a_n + c_n = o(\max\{|b_n|, |d_n|\}) \tag{1.3.4}$$

Definição 1.3.3. *Duas sequências a_n e b_n são ditas de mesma ordem $a_n \asymp b_n$ se $\left|\frac{a_n}{b_n}\right|$ é limitada entre 0 e ∞, isto é, se existem constantes $0 < m < M < \infty$ e um inteiro n_0 tal que $m < \left|\frac{a_n}{b_n}\right| < M$ para $n > n_0$.*

Definição 1.3.4. *Dizemos que $a_n = O(b_n)$ se existe um número real M positivo e um positivo inteiro n_0 tal que $\left|\frac{a_n}{b_n}\right| < M$ para $n \geq n_0$.*

1. se $a_n = O(b_n)$ e $c_n = o(d_n)$, então:

$$a_n c_n = O(b_n)o(d_n) = b_n d_n \underbrace{O(1)o(1)}_{=o(1)} = o(b_n d_n) \tag{1.3.5}$$

2. se $a_n = O(b_n)$ e $b_n = o(c_n)$, então:

$$a_n = O(1)b_n = o(c_n)O(1) = c_n o(1)O(1) = o(c_n) \tag{1.3.6}$$

Definição 1.3.5. *Dizemos que a_n e b_n são equivalentes assintoticamente se $\frac{a_n}{b_n} \to 1$, quando $n \to \infty$, denotamos $a_n \sim b_n$.*

Definição 1.3.6. *Se uma sequência a_n satisfaz $a_n \sim r_n c_n$, então r_n é a taxa de convergência de a_n.*

Exemplo 1.3.3. *Suponha que X segue uma distribuição binomial $b(n, p)$ correspondendo a n tentativas com probabilidade de sucesso p. O estimador padrão $\frac{X}{n}$ é não-viesado e tem variância $\frac{pq}{n}$. Uma classe de estimadores interessantes é*

$$\delta(X) = \frac{a + X}{a + b + n} \tag{1.3.7}$$

Dado que

$$\mathbb{E}[\delta(X)] = \frac{a + np}{a + b + n} \tag{1.3.8}$$

o viés de δ é

$$\begin{aligned} \textit{viés de } \delta &= \frac{a + np}{a + b + n} - p \\ &= \frac{aq - bp}{a + b + n} \end{aligned} \tag{1.3.9}$$

em que $q = 1 - p$.

De forma similar,

$$\mathrm{V}[\delta(X)] = \frac{npq}{(a + b + n)^2} \tag{1.3.10}$$

A acurácia do estimador é mensurada pelo erro quadrático médio

$$\mathrm{MSE} = \left(\frac{aq - bp}{a + b + n}\right)^2 + \frac{npq}{(a + b + n)^2} \tag{1.3.11}$$

Assim,

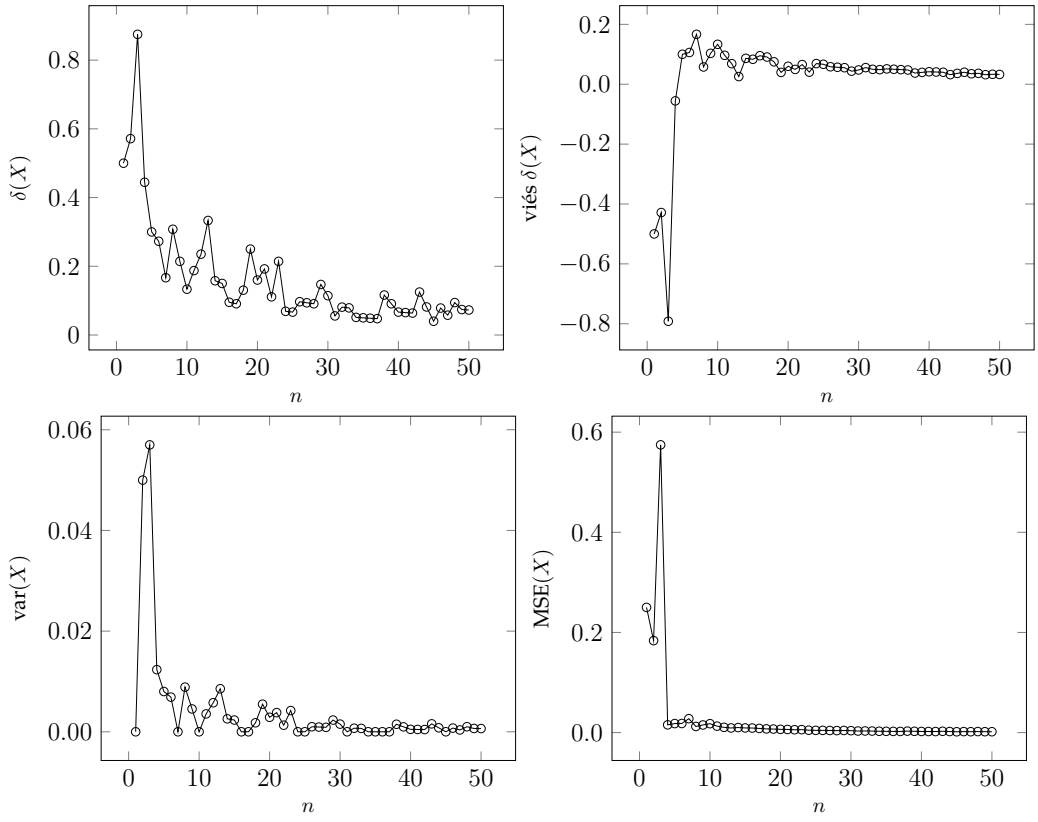
$$\frac{(\textit{viés de } \delta)^2}{\mathrm{V}[\delta(X)]} \approx \frac{(aq - bp)^2}{npq} \to 0 \textit{ quando } n \to \infty \tag{1.3.12}$$

e, desse modo,

$$(\textit{viés de } \delta)^2 = o\left[\mathrm{V}[\delta(X)]\right] \tag{1.3.13}$$

Ambos os termos tendem a zero, mas o quadrado do viés tende muito mais rápido: à taxa de $1/n^2$ em comparação com a taxa de $1/n$ para a variância. O viés, portanto, contribui relativamente pouco para o erro quadrático médio.

Figura 1.3.1 – Distribuição Binomial $b(n, p)$



Nota: Foi gerada uma amostra aleatória a partir de uma distribuição binomial $b(50, 0.5)$. Fixou-se $a = 2$ e $b = 3$.

Exemplo 1.3.4. *Sejam as sequências*

$$
\begin{aligned}
a_{1,n} &= \frac{2}{n} + \frac{50}{n^2} \\
a_{2,n} &= \frac{\sin\left({}^{n}/_{5} + 2\right)}{n^{5/4}} \\
a_{3,n} &= 3\frac{\left(1 + 5\exp\left(-\dfrac{(n - 55.5)^2}{200}\right)\right)}{n} \\
a_{4,n} &= \left(2\log_{10}(n)\left(\frac{n+3}{2n}\right)\right)^{-1} + a_{3,n} \\
a_{5,n} &= \left(4\log_2\left(\frac{n}{2}\right)\right)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
a_{6,n} &= \left(\log(n^2+n)\right)^{-1} \\
a_{7,n} &= \frac{\log(5n+3)^{-1/4}}{2} \\
a_{8,n} &= \left(4\log(\log(10n+2))\right)^{-1} \\
a_{9,n} &= \left(2\log(\log(n^2+10n+2))\right)^{-1} \\
b_n &= \frac{1}{\log(n)}
\end{aligned}
$$

Figura 1.3.2 – Diferenças e Similaridades entre o e O

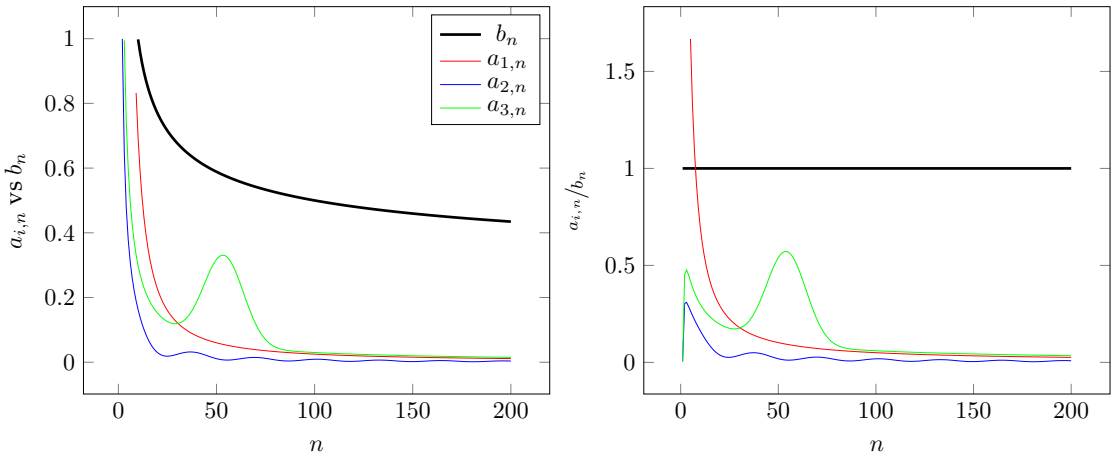


Figura 1.3.3 – Diferenças e Similaridades entre o e O

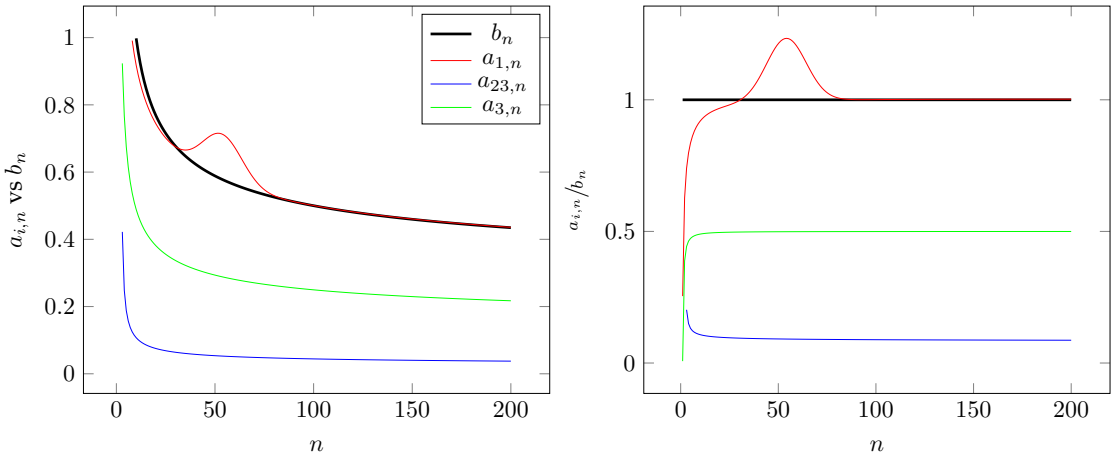
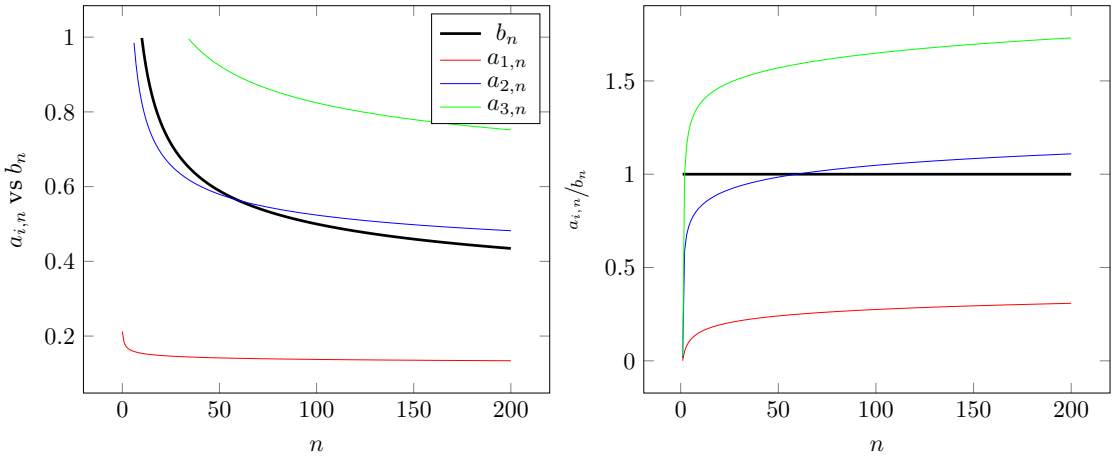


Figura 1.3.4 – Diferenças e Similaridades entre o e O



As sequências $a_{1,n}$, $a_{2,n}$ e $a_{3,n}$ são $o(b_n)$ (portanto também são $O(b_n)$). As sequências $a_{4,n}$, $a_{5,n}$ e $a_{6,n}$ são $O(b_n)$ (mas não são $o(b_n)$). Finalmente, as sequências $a_{7,n}$, $a_{8,n}$ e $a_{9,n}$ não são $O(b_n)$ (e portanto não são nem $o(b_n)$).

1.4 Convergência Estocástica

Definição 1.4.1. *Convergência em probabilidade: seja uma sequência $X_1, X_2, \ldots, X_n$ de variáveis aleatórias e X uma variável aleatória. Dizemos que $\{X_n\}_{n\geq 1}$ converge em probabilidade em X se, dado $\varepsilon > 0$,*

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0 \quad \textit{quando} \quad n \to \infty \quad \therefore \quad X_n \xrightarrow{P} X \tag{1.4.1}$$

Exemplo 1.4.1. *Suponha que $X_1, \ldots \sim U[0,1]$. Vamos definir $X_{(n)} = \max_{1\leq i\leq n} X_i$. Vamos verificar que $X_{(n)}$ converge em probabilidade para 1.*

Para ver isso, observe que

$$\begin{aligned}
\mathbb{P}\left(\left|X_{(n)} - 1\right| \geq \varepsilon\right) &= \mathbb{P}\left(X_{(n)} \leq 1 - \varepsilon\right) \\
&= \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq 1 - \varepsilon\right) \\
&= (1-\varepsilon)^n \to 0
\end{aligned} \tag{1.4.2}$$

Definição 1.4.2. *Dada uma sequência estritamente positiva $\{b_n\}_{n\geq 1}$ e uma sequência de variáveis aleatórias X_n,*

$$\begin{aligned} X_n = o_p(b_n) \colon &\iff \frac{X_n}{b_n} \xrightarrow{P} 0 \\ &\iff \lim_{n\to\infty} \mathbb{P}\left[\frac{|X_n|}{b_n} > \varepsilon\right] = 0, \quad \varepsilon > 0 \end{aligned} \tag{1.4.3}$$

Definição 1.4.3. *Dada uma sequência estritamente positiva $\{b_n\}_{n\geq 1}$ e uma sequência de variáveis aleatórias X_n,*

$$\begin{aligned} X_n = O_p(b_n) \colon &\iff \forall \varepsilon > 0, \exists C_\varepsilon > 0, n_0(\varepsilon) \in \mathbb{N} \\ &\quad \forall n \geq n_0(\varepsilon), \mathbb{P}\left[\frac{|X_n|}{b_n} > C_\varepsilon\right] < \varepsilon \\ &\iff \lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}\left[\frac{|X_n|}{b_n} > C\right] = 0 \end{aligned} \tag{1.4.4}$$

Definição 1.4.4. *Desigualdade de Chebyschev: seja uma variável aleatória com $\sigma^2 < +\infty$. Então, para qualquer $\varepsilon > 0$*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{\mathrm{V}(X)}{\varepsilon^2} \tag{1.4.5}$$

Exemplo 1.4.2. *Seja $X_1, X_2, \ldots, X_n$ variáveis aleatórias i.i.d. e $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Assim, $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, $\mathbb{E}\left(\overline{X}_n\right) = \mu$ e $\mathrm{V}\left(\overline{X}_n\right) = \dfrac{\sigma^2}{n}$. Então,*

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ quando } n \to \infty \;\; \therefore \;\; \overline{X}_n \xrightarrow{P} X \tag{1.4.6}$$

Pela desigualdade de Chebyschev,

$$\mathbb{P}\left(\left|\frac{X - \mathbb{E}(X)}{\sqrt{\mathrm{V}(X)}}\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \tag{1.4.7}$$

Se $\eta = \dfrac{1}{\varepsilon^2}$ *e* $C = \varepsilon = \dfrac{1}{\sqrt{\eta}}$,

$$\mathbb{P}\left(\left|\frac{X - \mathbb{E}(X)}{\mathrm{V}(X)}\right| \geq C\right) \leq \eta \tag{1.4.8}$$

Nesse caso, $\dfrac{X - \mathbb{E}(X)}{\mathrm{V}(X)} = O_p(1) \implies X = \mathbb{E}(X) + \sqrt{\mathrm{V}(X)}O_p(1)$.
Assim, com $\sigma^2 < +\infty$, *temos:*

$$X_n - \mathbb{E}(X_n) = O_P\left(\sqrt{\mathrm{V}(X_n)}\right) \tag{1.4.9}$$

Definição 1.4.5. *Uma sequência de variáveis aleatórias* $X_1, X_2, \ldots, X_n$ *converge quase certamente para uma variável aleatória* X, *denotado por* $X_n \overset{a.s.}{\to} X$, *se*

$$\mathbb{P}\left(\left\{s \in S : \lim_{n\to\infty} X_n(s) = X(s)\right\}\right) = 1. \tag{1.4.10}$$

1.5 Quantidade Pivotal

Uma quantidade pivotal que possui duas características:

1. é uma função da amostra aleatória e do parâmetro θ, em que θ é a única quantidade desconhecida, isto é, devemos obter $Q = Q(X_1, \ldots, X_n, \theta)$
2. tem uma função distribuição de probabilidade que não depende do parâmetro θ

Por exemplo, seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição normal de média θ e variância 9, isto é, X_i é *i.i.d.* com $\mathcal{N}(\theta, 9)$. Verifique se são quantidades pivotais:

$$\overline{X} - \theta, \quad \frac{\overline{X} - \theta}{3/\sqrt{n}}, \quad \frac{\overline{X}}{\theta} \tag{1.5.1}$$

Temos que $X \sim \mathcal{N}(\theta, 9)$, então $\overline{X} \sim \mathcal{N}\left(\theta, 9/n\right)$.

1. $Q = \overline{X} - \theta$ depende da amostra aleatória e de θ e além disso $\mathbb{E}(Q) = \mathbb{E}(\overline{X}) - \theta = 0$. Também $\mathrm{V}(Q) = \mathrm{V}(\overline{X}) = \dfrac{9}{n}$ e assim $Q = \overline{X} - \theta \sim \mathcal{N}\left(0, {}^{9}/_{n}\right)$. E, portanto, por definição, $Q = \overline{X} - \theta$ é uma quantidade pivotal.

2. $Q = \dfrac{\overline{X} - \theta}{{}^{3}/_{\sqrt{n}}}$ depende da amostra aleatória de θ e além disso $\mathbb{E}(Q) = \mathbb{E}\left(\dfrac{\overline{X} - \theta}{{}^{3}/_{\sqrt{n}}}\right) = 0$. Também $\mathrm{V}(Q) = \mathrm{V}\left(\dfrac{\overline{X} - \theta}{{}^{3}/_{\sqrt{n}}}\right) = \dfrac{n}{9}\mathrm{V}(\overline{X}) = 1$ e assim $Q = \dfrac{\overline{X} - \theta}{{}^{3}/_{\sqrt{n}}} \sim \mathcal{N}(0, 1)$. E, portanto, por definição, $Q = \dfrac{\overline{X} - \theta}{{}^{3}/_{\sqrt{n}}}$ é uma quantidade pivotal.

3. $Q = \dfrac{\overline{X}}{\theta}$ depende da amostra aleatória de θ e além disso $\mathbb{E}(Q) = \dfrac{1}{\theta}\mathbb{E}(\overline{X}) = 1$. Também $\mathrm{V}(Q) = \dfrac{1}{\theta^2}\mathrm{V}(\overline{X}) = \dfrac{9}{\theta^2 n}$ e assim $Q = \dfrac{\overline{X}}{\theta} \sim \mathcal{N}\left(1, {}^{9}/_{\theta^2 n}\right)$. E, portanto, por definição, $Q = \dfrac{\overline{X}}{\theta}$ não é uma quantidade pivotal, pois sua distribuição depende de θ.

1.6 Diferencial de Hadamard

Definição 1.6.1. *Sejam $f: U \subset X \to Y$, X e Y espaços vetoriais topológicos, U aberto de X, $\mathcal{G}$ a classe das funções $g: I \to U$, $I \in \mathbb{R}$ vizinhança de 0, tais que $g(0) = x_0$ e $\exists g'(0) = \lim_{t\to 0} \dfrac{g(t) - g(0)}{t} \in X$. A função f é Hadamard diferenciável em $x_0 \in U$ se e somente se $\exists {}^{H}f'(x_0) \in L(X, Y)$ satisfazendo $\forall g \in \mathcal{G}, \exists (f \circ g)'(0) = \lim_{t\to 0} \dfrac{(f \circ g)(t) - (f \circ g)(0)}{t} \in Y$ e $(f \circ g)'(0) = {}^{H}f'(x_0)g'(0)$.*

Para mostrarmos que uma função f é Hadamard diferenciável, podemos utilizar o seguinte resultado, dado por Daniel Henry, que é equivalente à definição anterior no caso de espaços de Banach.

Teorema 1.6.1. *Sejam X e Y espaços de Banach reais, $x_0 \in U \subset X$, com U aberto de X e $f: U \to Y$. A função f é Hadamard diferenciável em x_0 com derivada ${}^{H}f'(x_0) = T$ se e somente existe $T \in L(X, Y)$ tal que $\forall v \in X$,*

$$Tv = \lim_{\substack{(t,l)\to(0,0) \\ (t,l)\in\mathbb{R}\times X}} \frac{f(x_0 + tv + tl) - f(x_0)}{t} \tag{1.6.1}$$

A diferencial de Hadamard pode ser uma excelente ferramenta para a resolução de problemas, que verificamos não serem deriváveis no sentido de Fréchet. Por exemplo, queremos linearizar fluxos de sistemas dinâmicos, em torno de um ponto fixo, com o intuito de obter uma conjugação local com o sistema original. Uma possibilidade seria a derivada de Gâteaux que, por ser a "mais fraca", quase não apresenta propriedades que possam ser usadas para analisar problemas relevantes. Nesse contexto, a derivada de Hadamard se mostra uma ótima ferramenta, já que possui muitas propriedades.

1.7 Intervalo de Confiança

Grande parte da inferência não paramétrica é dedicada a encontrar um estimador $\hat{\theta}_n$ de alguma quantidade de interesse θ. Aqui, por exemplo, θ poderia ser uma média, uma densidade ou uma função de regressão. Mas também queremos fornecer conjuntos de confiança para essas quantidades. Existem diferentes tipos de conjuntos de confiança, como explicamos agora.

Seja $\mathscr{F}$ uma classe de funções de distribuição F e seja θ alguma quantidade de interesse. Assim, θ pode ser o próprio F, ou F' ou a média de F e assim por diante. Seja C_n um conjunto de possíveis valores de θ que dependem dos dados $X_1, \ldots, X_n$ (observe que C_n é aleatório). Para enfatizar que as probabilidade dependem de F às vezes escreveremos $\mathbb{P}_F$.

Definição 1.7.1. *C_n é um conjunto de confiança de amostra finita de $(1-\alpha)\%$ se*

$$\inf_{F\in\mathscr{F}} \mathbb{P}_F[\theta \in C_n] \geq 1-\alpha, \quad \forall n \tag{1.7.1}$$

Definição 1.7.2. *C_n é um conjunto de confiança uniforme assintótico de $(1-\alpha)\%$ se*

$$\liminf_{n\to\infty} \inf_{F\in\mathscr{F}} \mathbb{P}_F[\theta \in C_n] \geq 1-\alpha \tag{1.7.2}$$

Definição 1.7.3. *C_n é um conjunto de confiança pontual assintótico de $(1-\alpha)\%$ se*

$$\text{para todo } F \in \mathscr{F}, \liminf_{n\to\infty} \mathbb{P}_F[\theta \in C_n] \geq 1-\alpha \tag{1.7.3}$$

Se $\|\cdot\|$ denota alguma norma e $\hat{f}_n$ é uma estimativa de f, então uma bola de confiança para f é um conjunto de confiança da forma

$$C_n = \left\{ F \in \mathscr{F} \colon \|f - \hat{f}_n\| \leq s_n \right\} \tag{1.7.4}$$

em que s_n pode depender dos dados. Suponha que f está definida em um conjunto $\mathcal{X}$. Um par de funções (ℓ, u) é uma banda de confiança ou envelope de confiança (é uma família de intervalos aleatórios) de nível de $(1-\alpha)\%$ se

$$\inf_{f \in \mathscr{F}} \mathbb{P}[\ell(x) \leq f(x) \leq u(x) \text{ para todo } x \in \mathcal{X}] \geq 1 - \alpha \tag{1.7.5}$$

Bolas e bandas de confiança podem ser amostras finitas, assintóticas pontuais e assintóticas uniformes como acima. Ao estimar uma quantidade real avaliada em vez de uma função, C_n é apenas um intervalo e chamamos C_n de intervalo de confiança.

Idealmente, gostaríamos de encontrar conjuntos de confiança de amostra finita. Quando isso não é possível, tentamos construir conjuntos de confiança assintóticos uniformes. O último recurso é um intervalo de confiança assintótico pontual. Se C_n é um conjunto de confiança assintótica uniforme, então o seguinte é verdadeiro: para qualquer $\delta > 0$ o existe um $n(\delta)$ tal que a cobertura de C_n é pelo menos $1 - \alpha - \delta$ para todo $n > n(\delta)$. Com um conjunto de confiança assintótica pontual, pode não existir um $n(\delta)$ finito. Nesse caso, o tamanho da amostra em que o conjunto de confiança tem cobertura próxima a $1 - \alpha$ dependerá de f (que não conhecemos).

1.8 Condição de Lindeberg

Na teoria da probabilidade, a condição de Lindeberg é uma condição suficiente (e sob certas condições também uma condição necessária) para que o TCL seja válido para uma sequência de variáveis aleatórias independentes. Ao contrário do TCL clássico, que exige que as variáveis aleatórias em questão tenham variância finita e sejam independentes e distribuídas de forma idêntica, o TCL de Lindeberg requer apenas que tenham variância finita, satisfaçam a condição de Lindeberg e sejam independentes.

Seja $(\Omega, \mathcal{F}, \mathbb{P})$ um espaço de probabilidade e $X_k \colon \Omega \to \mathbb{R}, k \in \mathbb{N}$, variáveis

aleatórias independentes definidas nesse espaço de probabilidade. Assuma que o valor esperado $\mathbb{E}[X_k] = \mu_k$ e as variâncias $\mathrm{V}[X_k] = \sigma_k^2$ existem e são finitas. Também seja $s_n^2 := \sum_{k=1}^{n} \sigma_k^2$.

Se esta sequência de variáveis aleatórias independentes X_k satisfazem a condição de Lindeberg:

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \mathbb{E}\left[(X_k - \mu_k)^2 \cdot \mathbb{I}_{\{|X_k-\mu_k|>\varepsilon s_n\}}\right] = 0 \tag{1.8.1}$$

para todo $\varepsilon > 0$, em que $\mathbb{I}$ é uma função indicadora, então o TCL é válido, isto é, as variáveis aleatórias

$$Z_n := \frac{\sum_{k=1}^{n}(X_k - \mu_k)}{s_n} \tag{1.8.2}$$

convergem em distribuição para uma variável normal padrão quando $n \to \infty$.

1.9 Teorema de Fubini

Teorema 1.9.1 (Teorema Fraco de Fubini)**.** *Se $f(x, y)$ é uma função contínua ao longo da região retangular $\mathcal{D} = \{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, então:*

$$\iint_{\mathcal{D}} f(x,y)dA = \int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dydx. \tag{1.9.1}$$

Teorema 1.9.2 (Teorema Forte de Fubini)**.** *Seja $f(x, y)$ uma função contínua ao longo da região $\mathcal{D}$.*

1. *Se $a \leq x \leq b$ e $g_1(x) \leq y \leq g_2(x)$ com $g_1(x)$ e $g_2(x)$ contínuas em $[a, b]$, então*

$$\iint_{\mathcal{D}} f(x,y)dA = \int_a^b \int_{g_1(x)}^{g_2(x)} f(x,y)dydx. \tag{1.9.2}$$

2. *Se* $c \leq y \leq d$ *e* $h_1(x) \leq x \leq h_2(x)$ *com* $h_1(y)$ *e* $h_2(y)$ *contínuas em* $[c, d]$*, então*

$$\iint_{\mathcal{D}} f(x,y)dA = \int_c^d \int_{h_1(x)}^{h_2(x)} f(x,y)dxdy. \tag{1.9.3}$$

1.10 Convolução

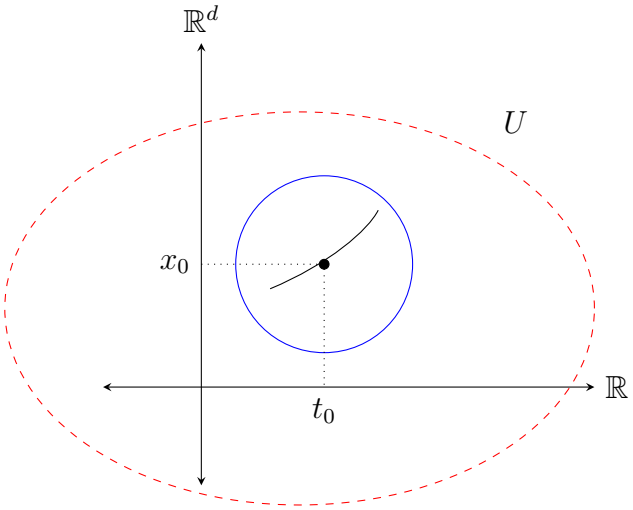
Se f e g são duas funções não-negativas nos reais $\mathbb{R}$, então sua convolução é definida como

$$(f \star g)(x) := \int_{-\infty}^{\infty} f(y)g(x-y)dy \tag{1.10.1}$$

1.11 Função Lipschitz

Definição 1.11.1 (Função Localmente Lipschitziana)**.** *Uma função é localmente Lipschitziana se para qualquer* $(t_0, x_0) \in U$*,* U *aberto de* $\mathbb{R}$*, existe uma constante* $C > 0$ *e uma constante* $\varepsilon > 0$ *tais que* $\|F(t, x_1) - F(t, x_2) \leq C\|x_1 - x_2\|$*,* $\forall x_1, x_2 \in B(x_0, \varepsilon)$ *e* $\forall t \in B(t_0, \varepsilon)$*.*

Figura 1.11.1 – Função Localmente Lipschitziana



2. Estimação Não-Paramétrica de Densidade

O objetivo da inferência estatística é usar dados para inferir uma quantidade desconhecida. No jogo de inferência, geralmente há uma compensação entre eficiência e generalidade, e essa compensação é controlada pela força das suposições feitas no processo de geração de dados.

A inferência paramétrica favorece a eficiência. Dado um modelo (uma forte suposição sobre o processo de geração de dados), a inferência paramétrica fornece um conjunto de métodos (estimativa de ponto, intervalos de confiança, teste de hipótese, etc.) adaptados para esse modelo. Todos esses métodos são os procedimentos inferenciais mais eficientes se o modelo corresponder à realidade, ou seja, se o processo de geração de dados realmente atender às premissas. Caso contrário, os métodos podem ser inconsistentes.

A inferência não paramétrica favorece a generalidade. Dado um conjunto de suposições mínimas e fracas (por exemplo, certa suavidade de uma densidade ou existência de momentos de uma variável aleatória), ele fornece métodos inferenciais que são consistentes para situações amplas, em troca de perda de eficiência para tamanhos de amostra pequenos ou moderados. Em termos gerais, uma técnica estatística se qualifica como "não paramétrica" se não se baseia em suposições paramétricas, que normalmente têm uma natureza de dimensão finita.

Portanto, para qualquer processo específico de geração de dados, existe um método paramétrico que domina sua contraparte não paramétrica em eficiência. Mas o conhecimento do processo de geração de dados raramente é o caso na prática. Esse é o apelo de um método não paramétrico: ele terá um desempenho adequado não importa qual seja o processo de geração de dados. Por esse motivo, os métodos não paramétricos são úteis:

1. quando não temos ideia do que poderia ser um bom modelo paramétrico.

2. para criar testes de adequação empregados para validar modelos paramétricos.

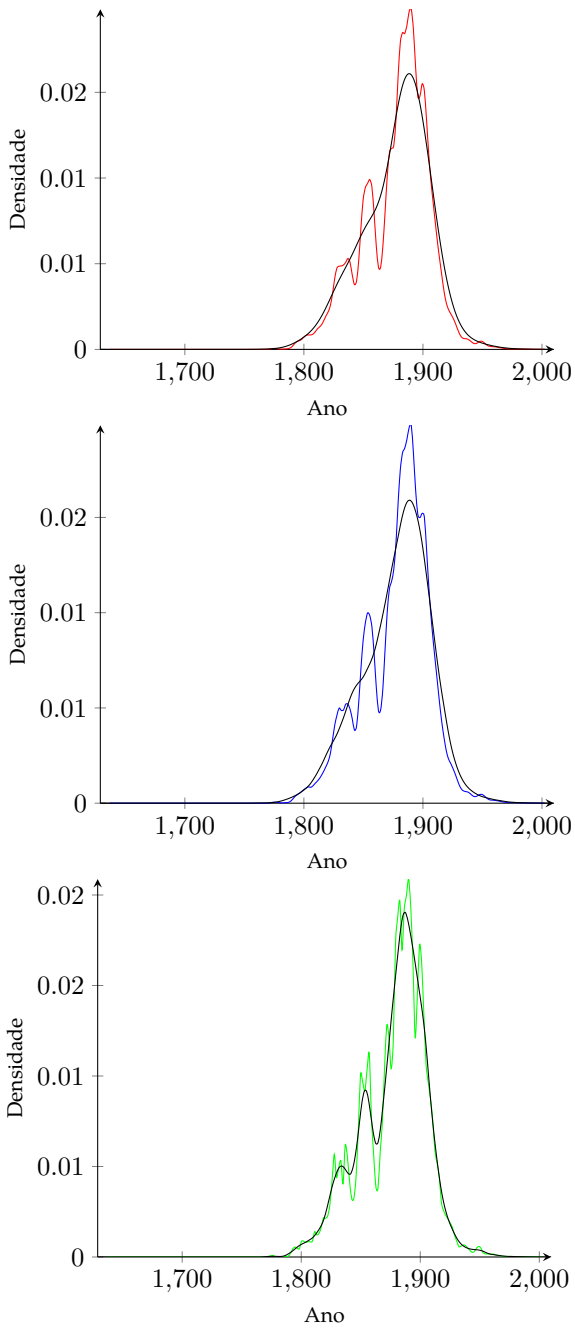
Uma função de densidade de probabilidade é um conceito-chave por meio do qual a variabilidade pode ser expressa com precisão. Na modelagem estatística, seu papel é frequentemente capturar a variação suficientemente bem, dentro de um modelo onde o principal interesse reside em termos estruturais, como coeficientes de regressão. No entanto, existem algumas situações em que a própria forma da função de densidade é o foco da atenção. O exemplo abaixo ilustra isso.

A estimativa da densidade *kernel* é um procedimento que historicamente precede a regressão do *kernel*. Também leva naturalmente a uma família simples de procedimentos para classificação não paramétrica.

Uma variável aleatória X é completamente caracterizada por sua função distribuição acumulada. Portanto, uma estimativa da FDA produz estimativas para diferentes características de X como produtos colaterais, conectando, nessas características, a FDA empírica F_n em vez de F.

Apesar de sua utilidade e muitas vantagens, as FDAs são difíceis de visualizar e interpretar, uma consequência de sua definição baseada em cumulativos. As densidades, por outro lado, são fáceis de visualizar e interpretar, tornando-as ferramentas ideais para a exploração de dados de variáveis aleatórias contínuas. Eles fornecem informações gráficas imediatas sobre as regiões de maior densidade, modas e forma do suporte de X. Além disso, as densidades também caracterizam completamente variáveis aleatórias contínuas. Porém, mesmo que uma função densidade de probabilidade seja derivada de uma função densidade acumulada pela relação $f = F'$, a estimativa da densidade não segue imediatamente a partir da FDA empírica F_n, uma vez que essa função não é diferenciável. Daí a necessidade de procedimentos específicos para estimar f que veremos neste capítulo.

Um estimador paramétrico é definido pelo modelo $\hat{f}(x|\theta)$, em que $\theta \in \Theta$. Provou-se surpreendentemente difícil formular uma definição de trabalho para o que constitui um estimador de densidade não paramétrico. Uma definição heurística pode ser proposta com base na condição necessária de que o estimador "trabalhe" para uma "grande" classe de densidades verdadeiras. Uma noção útil é que um estimador não paramétrico deve ter muitos parâmetros, na verdade, talvez um número infinito, ou um número que diverge em função do tamanho da amostra. Tapia e Thompson (1978) tendem a favorecer a noção de que o estimador não paramétrico deve ter dimensão infinita. Silverman (1986) simplesmente indica que



Fonte.— Cameron Blevins, Paper Trails: The US Post and the Making of the American West. New York: Oxford University Press, 2021. Conjunto de dados históricos e espaciais de 166.140 correios que operaram nos Estados Unidos entre 1639–2000.
Nota: Em preto está a densidade estimada com $h = 10$. O primeiro plot acima foi obtido com o *kernel* gaussiano; o segundo com o *kernel epanechnikov*; e o último com o *kernel biweight*.

uma abordagem não paramétrica faz "suposições menos rígidas sobre a distribuição dos dados observados". Mas quantos parâmetros um histograma possui? E quanto ao estimador de série ortogonal com um número infinito de termos que é igual (na distribuição) à função de densidade empírica para qualquer tamanho de amostra?

Uma definição surpreendentemente elegante está implícita no trabalho de Terrell (Terrell e Scott, 1992), que mostra que todos os estimadores, paramétricos ou não paramétricos, são estimadores de kernel generalizados, pelo menos assintoticamente. Terrell introduz a ideia da influência de um ponto de dados x_i sobre a densidade em x. Se $\hat{f}(x)$ é um estimador não paramétrico, a influência de um ponto deve desaparecer assintoticamente se $|x - x_i| > \varepsilon$ para qualquer $\varepsilon > 0$, enquanto a influência de pontos distantes não desaparece para um estimador paramétrico. Grosso modo, estimadores não paramétricos são assintoticamente locais, enquanto estimadores paramétricos não são. No entanto, estimadores não paramétricos não devem ser muito locais para serem consistentes.

2.1 Estimação de Densidade Univariada

O clássico histograma de frequência é formado pela construção de um conjunto completo de intervalos não sobrepostos, chamados *bins*, e contando o número de pontos em cada *bin*. Para que as contagens de *bins* sejam comparáveis, todas os *bins* devem ter a mesma largura. Nesse caso, o histograma é completamente determinado por dois parâmetros, a largura do *bin*, h, e a origem do *bin*, t_0, que é qualquer ponto final do intervalo do *bin* escolhido convenientemente. Frequentemente, a origem do *bin* é escolhida para ser $t_0 = 0$.

Embora a ideia de agrupar dados na forma de um histograma seja pelo menos tão antiga quanto o trabalho de Graunt em 1662, nenhuma orientação sistemática para projetar histogramas foi fornecida até a breve nota de Herbert Sturges em 1926. Seu trabalho fez uso de um dispositivo que foi defendido de forma mais geral por Tukey (1977). Tomando a densidade normal como um ponto de referência ao pensar sobre os dados, Sturges simplesmente observou que a distribuição binomial[1],

[1] Em teoria das probabilidades e estatística, a distribuição binomial é a distribuição de probabilidade discreta do número de sucessos numa sequência de n tentativas tais que: (i) cada tentativa tem exclusivamente como resultado duas possibilidades, sucesso ou fracasso; (ii) cada tentativa é independente das demais;(iii) a probabilidade de sucesso p a cada tentativa permanece constante independente das demais, e; (iv) a variável de interesse, ou pretendida, é o número de sucessos k nas n tentativas.

$B(n, p = 0,5)$, poderia ser usada como um modelo de um histograma construído de forma otimizada com dados normais adequadamente dimensionados.

Nesse sentido, seja um histograma de frequência com k *bins*, cada um com largura 1 e centralizado nos pontos $i = 0, 1, \ldots, k-1$. Escolha a contagem de *bin* do i-ésimo *bin* para ser o coeficiente binomial[2] $\binom{k-1}{i}$. À medida que k aumenta, este histograma de frequência ideal assume a forma de uma densidade normal com média $\left(\frac{k-1}{2}\right)$ e variância $\left(\frac{k-1}{4}\right)$. O tamanho total da amostra é

$$n = \sum_{i=0}^{k-1} \binom{k-1}{i} = 2^{k-1} \tag{2.1.1}$$

pela expansão binomial. Segue, da regra de Sturges, que

$$k = 1 + \log_2 n \tag{2.1.2}$$

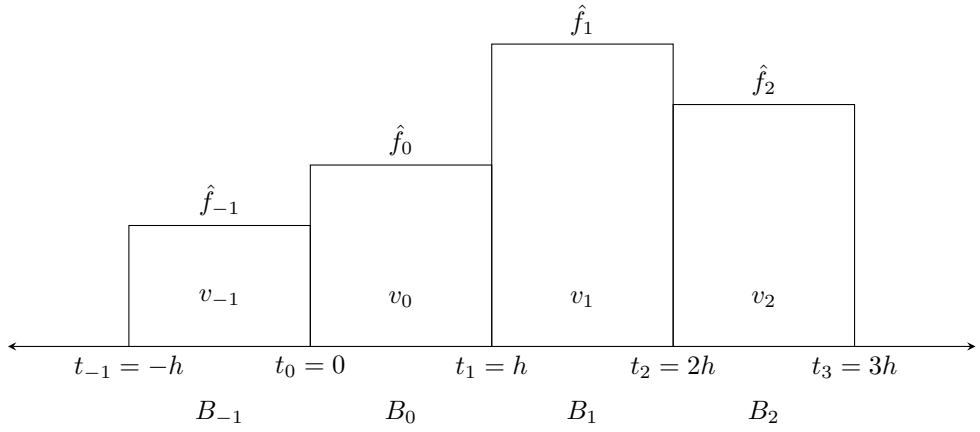
A diferença entre um histograma de frequência e um histograma de densidade é que o último é normalizado para ter integral igual a 1. O histograma é completamente determinado pela amostra $\{X_1, \ldots, X_n\}$ de $f(x)$ e pela escolha de uma malha $\{t_k, -\infty < k < \infty\}$. Assuma que $B_k = [t_k, t_{k+1})$ denote o k-ésimo *bin*. Suponha que $t_{k+1} - t_k = h$ para todo k; então, diz-se que o histograma tem largura fixa h. Um histograma de frequência é construído usando blocos de altura 1 e largura h empilhados nas caixas apropriadas. A integral de tal figura é claramente igual a nh. Assim, um histograma de densidade usa blocos de construção de altura $\frac{1}{nh}$, então cada bloco tem área igual a $\frac{1}{n}$. Seja v_k a contagem do número de pontos de amostra caindo no *bin* B_k. Então, o histograma é definido como

$$\hat{f}(x) = \frac{v_k}{nh} = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{I}_{[t_k, t_{k+1})}(x_i) \quad \text{para } x \in B_k \tag{2.1.3}$$

em que $v_k \sim B(n, p_k)$.

[2] O coeficiente binomial, também chamado de número binomial, de um número n, na classe k, consiste no número de combinações de n termos, k a k.

Figura 2.1.1 – Notação para Construção de um Histograma Igualmente Espaçado



Seja X uma variável aleatória com função de distribuição contínua $F(x)$ e densidade $f(x) = \dfrac{d}{dx}$. O objetivo é estimar $f(x)$ a partir de uma amostra aleatória $\{X_1, \ldots, X_n\}$. A função de distribuição $F(x)$ é naturalmente estimada pela função de distribuição empírica $\widehat{F}(x) = \dfrac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$. Pode parecer natural estimar a densidade $f(x)$ como a derivada de $\widehat{F}(x)$, mas esse estimador será um conjunto de pontos de massa, não uma densidade, e como tal não é uma estimativa útil de $f(x)$.

O método mais simples para estimar uma densidade f de uma amostra *i.i.d.* é o histograma. Do ponto de vista analítico, a ideia é agregar os dados em intervalos da forma $[x_0, x_0 + h)$ e, em seguida, usar sua frequência relativa para aproximar a densidade em $x \in [x_0, x_0 + h)$. Assim, a estimativa de $f(x)$ é

$$\begin{aligned} f(x_0) &= F'(x_0) \\ &= \lim_{h \to 0^+} \frac{F(x_0 + h) - F(x_0)}{h} \\ &= \lim_{h \to 0^+} \frac{\mathbb{P}[x_0 < X < x_0 + h]}{h} \end{aligned} \tag{2.1.4}$$

Mais precisamente, dada uma origem t_0 e uma largura de banda $h > 0$, o histograma constrói uma função constante por partes nos intervalos

$$\{B_k := [t_k, t_{k+1}) : t_k = t_0 + hk, k \in \mathbb{Z}\} \tag{2.1.5}$$

contando o número de pontos de amostra dentro de cada um deles. Esses intervalos de comprimento constante também são chamados de *bins*. O fato de terem comprimento constante h é importante: podemos facilmente padronizar as contagens em qualquer *bin* por h, a fim de termos frequências relativas por comprimento nos *bins*. O histograma em um ponto x é definido como

$$\hat{f}_H(x; t_0, h) := \frac{1}{nh} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in B_k : x \in B_k\}} \tag{2.1.6}$$

De forma equivalente, se denotarmos o número de observações $X_1, \ldots, X_n$ em B_k como v_k, o histograma pode ser escrito como

$$\hat{f}_H(x; t_0, h) = \frac{v_k}{nh}, \quad \text{se } x \in B_k \text{ para um certo } k \in \mathbb{Z} \tag{2.1.7}$$

A análise de $\hat{f}_H(x; t_0, h)$ como uma variável aleatória é simples, uma vez que se reconheça que a contagem dos *bins* v_k é distribuída como uma binomial $B(n, p_k)$, com $p_k := \mathbb{P}[X \in B_k] = \int_{B_k} f(t)dt$. Se f é contínua[3], pelo teorema do valor médio, $p_k = hf(\xi_{k,h})$ para um $\xi_{k,h} \in [t_k, t_{k+1})$. Assuma que $k \in \mathbb{Z}$ tal que $x \in B_k = [t_k, t_{k+1})$. Portanto,

$$\mathbb{E}[\hat{f}_H(x; t_0, h)] = \frac{np_k}{nh} = f(\xi_{k,h}) \tag{2.1.8}$$

$$\mathrm{V}[\hat{f}_H(x; t_0, h)] = \frac{np_k(1-p_k)}{n^2h^2} = \frac{f(\xi_{k,h})\left[(1 - hf(\xi_{k,h}))\right]}{nh} \tag{2.1.9}$$

$$\text{Viés}[\hat{f}_H(x; t_0, h)] = \frac{np_k}{nh} - f(x) = \frac{p_k}{h} - f(x) \tag{2.1.10}$$

Os resultados acima fornecem informações interessantes[4]:

1. se $h \to 0$, então $\xi_{k,h} \to x$, resultando em $f(\xi_{k,h}) \to f(x)$ e, portanto, (2.1.6) torna-se um estimador assintoticamente não-viesado (quando $h \to 0$) de $f(x)$.

[3] Exige-se que seja localmente Lipschitziana.

[4] Note que k depende de h porque $t_k = t_0 + kh$. Portanto, o k para o qual $x \in [t_k, t_{k+1})$ irá mudar quando $h \to 0$.

2. mas se $h \to 0$, a variância aumenta. Para a variância decrescer, $nh \to \infty$ é requerido.

3. a variância é diretamente dependente de $f(\xi_{k,h})(1-hf(\xi_{k,h})) \to f(x)$ (quando $h \to 0$), portanto, há mais variabilidade em regiões com maior densidade

Vamos focar no exemplo abaixo na dependência de t_0.

Figura 2.1.2 – Histograma de uma Distribuição Uniforme com $t_0 = 0$ e $h = 0.2$

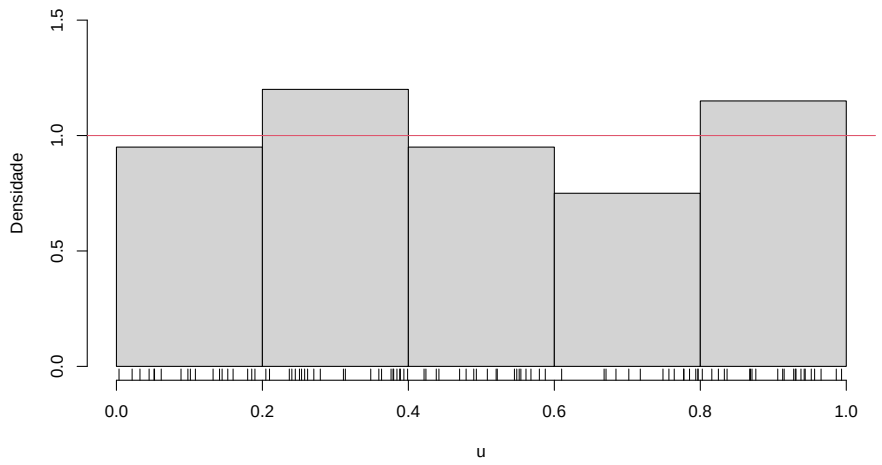


Figura 2.1.3 – Histograma de uma Distribuição Uniforme com $t_0 = -0.1$ e $h = 0.2$

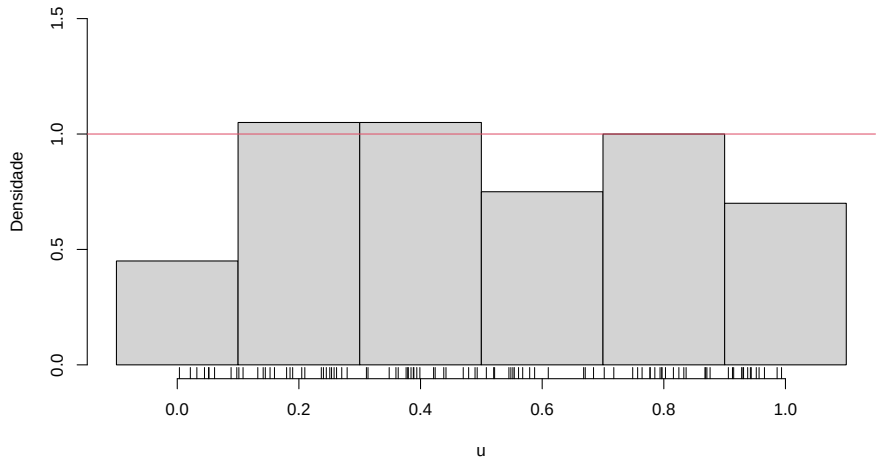


Figura 2.1.4 – Histograma de uma Distribuição Uniforme com $t_0 = 0$ e $h = 0.05$

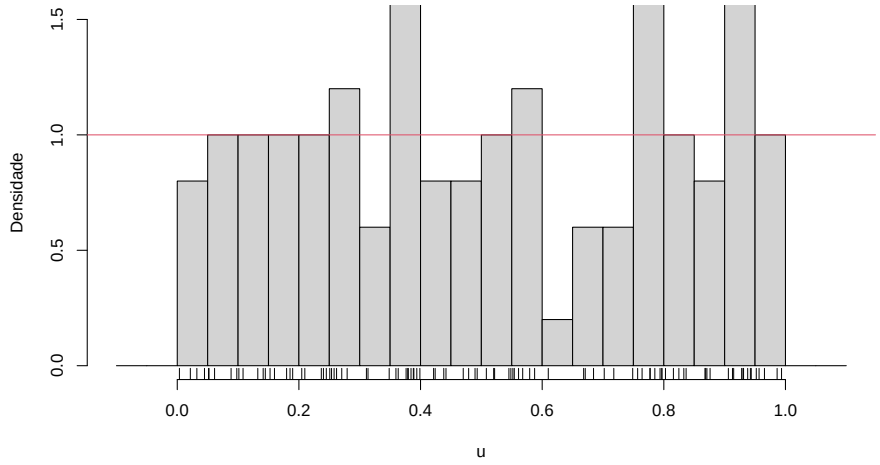
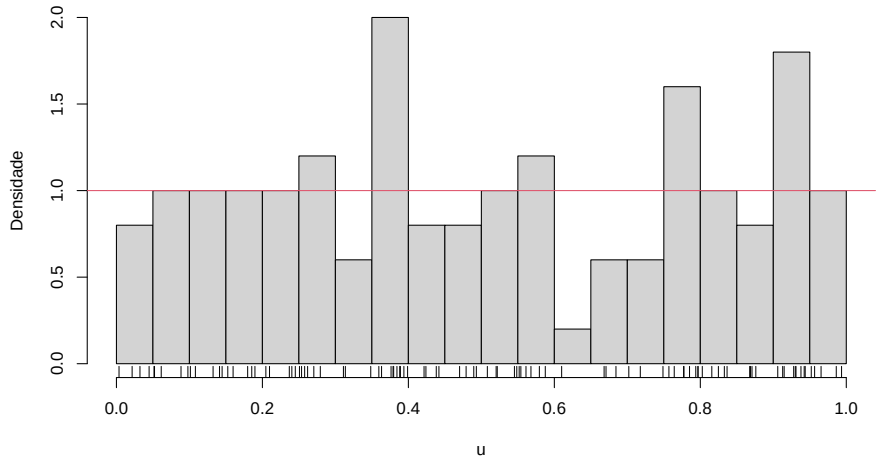


Figura 2.1.5 – Histograma de uma Distribuição Uniforme com $t_0 = -0.1$ e $h = 0.05$



Claramente, a subjetividade introduzida pela dependência de t_0 é algo da qual gostaríamos de nos livrar. Podemos fazer isso permitindo que os *bins* sejam dependentes de x (o ponto em que queremos estimar $f(x)$, em vez de fixá-los de antemão.

Uma alternativa para evitar a dependência de t_0 é o histograma móvel, também conhecido como *naive density estimator*. Para algum $h > 0$, faça,

$$\begin{aligned} f(x) &= F'_x(x) \\ &= \lim_{h\to 0^+} \frac{F(x+h) - F(x-h)}{2h} \\ &= \lim_{h\to 0^+} \frac{\mathbb{P}[x-h < X < x+h]}{2h} \\ \widehat{f}_N(x,h) &= \frac{\text{número de elementos na amostra entre } (x-h) \text{ e } (x+h)}{2nh} \\ \widehat{f}_N(x,h) &= \frac{\sum_{i=1}^{n} \mathbb{I}_{[x-h,x+h]}(x_i)}{2nh} \end{aligned} \tag{2.1.11}$$

Assim, o i-ésimo elemento da amostra está entre $(x-h)$ e $(x+h)$, de tal forma que:

$$\begin{aligned} x - h \leq x_i \leq x + h \\ -h \leq x_i - x \leq h \\ -1 \leq \frac{x_i - x}{h} \leq 1 \end{aligned} \tag{2.1.12}$$

Portanto, um estimador não-paramétrico para estimar a densidade pode ser elaborado como segue:

$$\widehat{f}_N(x,h) = \frac{1}{nh} \sum_{i=1}^{n} \Omega\left(\frac{x_i - x}{h}\right), \text{ com } h > 0 \tag{2.1.13}$$

em que h é o parâmetro de suavização (*bandwidth*) e

$$\Omega(x) = \begin{cases} \dfrac{1}{2}, & \text{se} \quad -1 \leq x \leq 1, \\ \\ 0, & \text{caso contrário.} \end{cases} \tag{2.1.14}$$

Analogamente ao histograma, a análise de $\widehat{f}_N(x,h)$ como uma variável aleatória decorre de perceber que

$$\sum_{i=1}^{n} \mathbb{I}_{\{x-h<X_i<x+h\}} \sim B(n, p_{x,h}) \tag{2.1.15}$$

Seja

$$\begin{aligned} p_{x,h} :&= \mathbb{P}[x-h<X<x+h] \\ &= F(x+h) - F(x-h) \end{aligned} \tag{2.1.16}$$

Portanto,

$$\begin{aligned} \mathbb{E}[\hat{f}_N(x;h)] =& \mathbb{E}\left[\frac{\sum_{i=1}^{n} \mathbb{I}_{[x-h,x+h]}(x_i)}{2nh}\right] \\ &= \frac{1}{2nh}\left[\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{I}_{[x-h,x+h]}(x_i)\right]\right] \\ &= \frac{1}{2nh}\left[n\left[F(x+h)-F(x-h)\right]\right] \\ &= \frac{F(x+h)-F(x-h)}{2h} \end{aligned} \tag{2.1.17}$$

e

$$\mathrm{V}[\hat{f}_N(x;h)] = \mathrm{V}\left[\frac{\sum_{i=1}^{n} \mathbb{I}_{[x-h,x+h]}(x_i)}{2nh}\right]$$

$$
\begin{aligned}
&= \frac{1}{4n^2h^2}\,\mathrm{V}\left[\sum_{i=1}^{n}\mathbb{I}_{[x-h,x+h]}(x_i)\right] \\
&= \frac{1}{4n^2h^2}\sum_{i=1}^{n}\mathrm{V}\left[\mathbb{I}_{[x-h,x+h]}(x_i)\right] \\
&= \frac{1}{4nh^2}\,\mathrm{V}\left[\mathbb{I}_{[x-h,x+h]}(x_i)\right] \\
&= \frac{1}{4nh^2}\left\{F(x+h)-F(x-h)\left[1-(F(x+h)-F(x-h))\right]\right\} \\
&= \frac{F(x+h)-F(x-h)}{4nh^2} - \frac{[F(x+h)-F(x-h)]^2}{4nh^2} \qquad (2.1.18)
\end{aligned}
$$

Assim, temos os seguintes resultados:

1. se $h \to 0$
   (a) $\mathbb{E}[\hat{f}_N(x;h)] \to f(x)$, resultando em $f(\xi_{k,h}) \to f(x)$ e, portanto, (2.1.11) torna-se um estimador assintoticamente não-viesado de $f(x)$.
   (b) $\mathrm{V}[\hat{f}_N(x;h)] \approx \frac{f(x)}{2nh} - \frac{f(x)^2}{n} \to \infty$
2. se $h \to \infty$
   (a) $\mathbb{E}[\hat{f}_N(x;h)] \to 0$
   (b) $\mathrm{V}[\hat{f}_N(x;h)] \to 0$
3. a variância converge para zero se $nh \to \infty$. Então, o viés e a variância podem ser reduzidos se $n \to \infty$, $h \to 0$ e $nh \to \infty$ simultaneamente.

Observe que não é necessário preocupar-se com o ponto inicial. O problema dessa abordagem é que $\widehat{f}(x,h)$ não é suave.

2.2 Estimação Kernel para a Densidade Univariada

É notável que o histograma permaneceu como o único estimador de densidade não paramétrico até a década de 1950, quando um progresso substancial e simultâneo foi feito na estimativa da densidade e na estimativa da densidade espectral. Em um relatório técnico pouco conhecido, Fix e Hodges (1951)

introduziram o algoritmo básico de estimativa não paramétrica da densidade. Eles abordaram o problema da discriminação estatística quando a forma paramétrica da densidade de amostragem não era conhecida. Felizmente, este artigo foi reimpresso com comentários de Silverman e Jones (1989). Durante a década seguinte, vários algoritmos gerais e modelos teóricos alternativos de análise foram introduzidos por Rosenblatt (1956), Parzen (1962) e Cencov (1962). Seguiu-se uma segunda onda de artigos importantes e principalmente teóricos de Watson e Leadbetter (1963), Loftsgaarden e Quesenberry (1965), Schwartz (1967), Epanechnikov (1969), Tarter e Kronmal (1970) e Wahba (1971). A generalização multivariada natural foi introduzida por Cacoullos (1966). Finalmente, na década de 1970, surgiram os primeiros artigos enfocando a aplicação prática desses métodos: Scott et al. (1978) e Silverman (1978b). Essas e outras aplicações multivariadas posteriores aguardavam a revolução da computação. O estimador kernel originou-se como uma aproximação numérica da derivada da função de distribuição cumulativa (Rosenblatt, 1956).

Para contornar o problema da falta de suavidade do estimador anterior defina o seguinte estimador em que se substitui Ω por $\mathcal{K}$. O método de *kernel* não possui a necessidade de definir classes. É um procedimento para estimação de uma função que consiste em estabelecer uma média localmente ponderada. Assim,

$$\widehat{f}_x(x,h) = \frac{1}{nh}\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i - x}{h}\right) \tag{2.2.1}$$

em que $\mathcal{K}(\cdot)\colon \mathbb{R} \to \mathbb{R}$ é a função *kernel*. O estimador *kernel* faz a estimação global com ponderação local. Este suavizador se utiliza de ponderação que decai de maneira suave à medida que os valores vão se afastando do ponto em que se está realizando a estimação.

Exemplos de uma função *kernel* incluem

- *Kernel* Retangular: $\frac{1}{2}\mathbb{I}_{[-1,1]}(x)$
- *Kernel* Gaussiano: $\mathcal{K}(x) = \phi(x)$.
- *Kernel* Epanechnikov : $\mathcal{K}(x) = \frac{3}{4}\left(1 - x^2\right)\mathbb{I}_{[-1,1]}(x)$.
- *Kernel* Biweight: $\frac{15}{16}\left(1 - x^2\right)^2\mathbb{I}_{[-1,1]}(x)$

Uma função *kernel* satisfaz as seguintes propriedades:

1. $\int_{-\infty}^{\infty} \mathcal{K}(u)du = 1$
2. um *kernel* não-negativo satisfaz $\mathcal{K}(u) \geq 0$ para todo u
3. os momentos de um *kernel* são $\kappa_j(\mathcal{K}) = \int_{-\infty}^{\infty} u^j \mathcal{K}(u)du$
4. uma função *kernel* simétrica satisfaz $\mathcal{K}(u) = \mathcal{K}(-u)$. Neste caso, todos os momentos ímpares são zero.

A ordem de um *kernel*, ν, é definida como a ordem do primeiro momento diferente de zero. Por exemplo, se $\kappa_1(\mathcal{K}) = 0$ e $\kappa_2(\mathcal{K}) > 0$, então $\mathcal{K}$ é um *kernel* de segunda ordem e $\nu = 2$. A ordem de um *kernel* simétrico é sempre par.

Um *kernel* é um *kernel* de ordem superior se $\nu > 2$. Esses *kernels* terão partes negativas e não são densidades de probabilidade. Eles também são chamados de *bias-reducing kernels*.

Além da fórmula do *kernel*, listamos sua rugosidade $R(\mathcal{K})$. A *roughness* de uma função é

$$R(\mathcal{K}) = \int_{-\infty}^{\infty} \mathcal{K}(u)^2 du \tag{2.2.2}$$

Para o *kernel* retangular, temos $\mathcal{K}(x) = \frac{1}{2}\mathbb{I}_{[-1,1]}(x)$. É possível verificar que

$$\int_{-1}^{1} \frac{1}{2}\mathbb{I}_{[-1,1]}(x)dx = 1 \tag{2.2.3}$$

Também,

$$\begin{aligned} R(\mathcal{K}) &= \int_{-\infty}^{\infty} \left(\frac{1}{2}\mathbb{I}_{[-1,1]}(x)\right)^2 dx \\ &= \frac{1}{4}\int_{-1}^{1} \mathbb{I}_{[-1,1]}(x) \\ &= \frac{1}{2} \end{aligned} \tag{2.2.4}$$

E, além disso,

$$\begin{aligned}\kappa_2(\mathcal{K}) &= \int_{-\infty}^{\infty} x^2 \left(\frac{1}{2}\mathbb{I}_{[-1,1]}(x)\right) dx \\ &= \frac{1}{3}\end{aligned} \tag{2.2.5}$$

Os *kernels* uniforme, *Epanechnikov*, *biweight*, *triweight* e Gaussiano são casos especiais da família polinomial, como segue:

$$\mathcal{K}_s(x) = \frac{(2s+1)!!}{2^{s+1}s!}\left(1-x^2\right)^s \mathbb{I}(|x|\leq 1) \tag{2.2.6}$$

em que o duplo fatorial é dado por $(2s+1)!! = (2s+1)(2s-1)\cdots 5\cdot 3\cdot 1$. O *kernel* gaussiano é obtido quando $s \to \infty$.

Vários tipos de *kernels* são possíveis. O mais popular é o *kernel* gaussiano, embora o *kernel* Epanechnikov produza o histograma móvel como um caso particular. O estimador *kernel* de densidade herda as propriedades de suavidade do *kernel*. Isso significa, por exemplo, que (2.2.1) com um *kernel* gaussiano é infinitamente diferenciável. Mas, com um *kernel* Epanechnikov, (2.2.1) não é diferenciável e, com um *kernel* retangular, o estimador *kernel* nem mesmo é contínuo. No entanto, se uma certa suavidade é garantida (continuidade pelo menos), então a escolha do *kernel* tem pouca importância na prática (pelo menos em comparação com a escolha da *bandwidth*).

2.3 Viés

É útil observar que as esperanças das transformações do *kernel* podem ser escritas como integrais que tomam a forma de uma convolução do *kernel* e da função de densidade. Seja

$$\begin{aligned}\mathbb{E}\left[\widehat{f}(x)\right] &= \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i - x}{h}\right)\right] \\ &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[\mathcal{K}\left(\frac{x_i - x}{h}\right)\right] \text{ [pela independência]}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{h}\mathbb{E}\left[\mathcal{K}\left(\frac{x_i - x}{h}\right)\right] \text{ [pela distribuição idêntica]} \\
&= \frac{1}{h}\int_{-\infty}^{\infty} \mathcal{K}\left(\frac{t-x}{h}\right) f_x(t)\, dt \\
&= \frac{1}{h}\int_{-\infty}^{\infty} \mathcal{K}(s) f_x(x+sh) h\, ds \\
&= \int_{-\infty}^{\infty} \mathcal{K}(s) f_x(x+sh)\, ds \\
&= \int_{-\infty}^{\infty} \mathcal{K}(s) \underbrace{\left(f(x) + f'(x)sh + \frac{s^2h^2}{2}f''(x) + \frac{1}{6}h^3s^3f'''(x)\right)}_{\text{expansão de Taylor de 2ª ordem}} ds \\
&= \underbrace{\int_{-\infty}^{\infty} f(x)\mathcal{K}(s)\, ds + \int_{-\infty}^{\infty} shf'(x)\mathcal{K}(s)\, ds + \int_{-\infty}^{\infty} \frac{s^2h^2}{2}f''(x)\mathcal{K}(s)\, ds + O\left(h^3\right)}_{\text{linearidade da integral}} \\
&= f(x)\underbrace{\int_{-\infty}^{\infty} \mathcal{K}(s)\, ds}_{=1} + hf'(x)\underbrace{\int_{-\infty}^{\infty} s\mathcal{K}(s)\, ds}_{=0 \text{ (simetria)}} + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\, ds + O\left(h^3\right) \\
&= f(x) + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\, ds + O\left(h^3\right) \qquad (2.3.1)
\end{aligned}
$$

em que, com $|f'''(x)| \leq C$,

$$
|R_n| \leq \frac{Ch^3}{6}\int s^3\mathcal{K}(S)ds = o\left(h^2\right). \qquad (2.3.2)
$$

isto é, $\frac{Ch^3}{6}\int s^3\mathcal{K}(S)$ tende para zero mais rápido que h^2.

Como supomos que $f(\cdot)$ é de classe $\mathcal{C}^2$, podemos escrever

$$
\mathbb{E}\left[\widehat{f}(x)\right] = f(x) + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\, ds + o\left(h^2\right) \qquad (2.3.3)
$$

Pela definição de viés, chegamos a:

$$\mathbb{E}\left[\widehat{f}(x)\right] - f(x) = \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} s^2 \mathcal{K}(s)\, ds + o\left(h^2\right) \tag{2.3.4}$$

O viés diminui com h quadraticamente. Além disso,o viés em x é diretamente proporcional a $f''(x)$. Isso tem uma interpretação interessante:

1. O viés é negativo onde f é côncava, ou seja, $\{x \in \mathbb{R} : f''(x) < 0\}$. Essas regiões correspondem a picos e modas de f, onde o estimador de densidade *kernel* subestima f.

2. De forma simétrica, o viés é positivo onde f é convexa, ou seja, $\{x \in \mathbb{R} : f''(x) > 0\}$. Essas regiões correspondem a vales e caudas de f, onde o estimador de densidade *kernel* sobrestima f.

3. Quanto mais "abrupta" for a curvatura de f'', mais difícil será estimar f. As regiões de densidade plana são mais fáceis de estimar do que as regiões de ondulação com alta curvatura (por exemplo, com vários modas).

O viés é causado pela curvatura (segunda derivada) da função densidade. Ou seja, o viés será maior em um ponto em que a função de densidade se curva muito (por exemplo, uma protuberância muito pontiaguda). Isso faz sentido porque para tal estrutura, o estimador de densidade *kernel* tende a suavizar muito, tornando a função de densidade mais suave (menos curva) do que costumava ser.

2.4 Variância

Dado que o estimador *kernel* é linear e $\mathcal{K}\left(\frac{x_i - x}{h}\right)$ é *i.i.d.*, temos que pela definição de variância:

$$\begin{aligned}
\mathrm{V}\left[\widehat{f}(x)\right] &= \mathrm{V}\left[\frac{1}{nh}\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i - x}{h}\right)\right] \\
&= \frac{1}{n^2h^2}\left\{\sum_{i=1}^{n}\mathrm{V}\left[\mathcal{K}\left(\frac{x_i - x}{h}\right)\right]\right\} \quad \text{[pela independência]}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{nh^2}\,\mathrm{V}\left[\mathcal{K}\left(\frac{x_i - x}{h}\right)\right] \quad \text{[pela distribuição idêntica]} \\
&= \frac{1}{nh^2}\left\{\mathbb{E}\left[\mathcal{K}^2\left(\frac{x_i - x}{h}\right)\right] - \left\{\mathbb{E}\left[\mathcal{K}\left(\frac{x_i - x}{h}\right)\right]\right\}^2\right\} \\
&= \frac{1}{nh^2}\left[\int_{-\infty}^{\infty} \mathcal{K}^2\left(\frac{t-x}{h}\right) f_x(t)\,dt - \left(\int_{-\infty}^{\infty} \mathcal{K}\left(\frac{t-x}{h}\right) f_x(t)\,dt\right)^2\right] \\
&= \frac{1}{nh^2}\left(\int_{-\infty}^{\infty} \mathcal{K}^2(s)\,(f(x) + o(1))h\,ds - \left[\int_{-\infty}^{\infty} \mathcal{K}(s)\,(f(x) + o(1))h\,ds\right]^2\right) \\
&= \frac{1}{nh^2}\left(h\int_{-\infty}^{\infty} \mathcal{K}^2(s)\,(f(x) + o(1))\,ds - h\left[\int_{-\infty}^{\infty} \mathcal{K}(s)\,(f(x) + o(1))h^{1/2}\,ds\right]^2\right) \\
&= \frac{1}{nh}\left(\int_{-\infty}^{\infty} \mathcal{K}^2(s)\,(f(x) + o(1))\,ds - \left[\int_{-\infty}^{\infty} \mathcal{K}(s)\,(f(x) + o(1))h^{1/2}\,ds\right]^2\right) \\
&= \frac{1}{nh}\left[f(x)\int_{-\infty}^{\infty} \mathcal{K}^2(s)ds + O\left(h\int \mathcal{K}^2(s)ds\right) - O(h)\right] \\
&= \frac{1}{nh}\left(f(x)\int_{-\infty}^{\infty} \mathcal{K}^2(s)ds + O(h)\right) \\
&= \frac{1}{nh}\left(f(x)\int_{-\infty}^{\infty} \mathcal{K}^2(s)ds + o(1)\right) \qquad (2.4.1)
\end{aligned}
$$

A variância depende diretamente de $f(x)$. Curiosamente, a variância diminui como um fator de $(nh)^{-1}$, uma consequência de nh desempenhar o papel do tamanho efetivo da amostra para a estimativa. O tamanho efetivo da amostra pode ser considerado como a quantidade de dados na vizinhança de x que é empregada para estimar f.

2.5 Erro Quadrático Médio

Uma medida comum e conveniente de precisão de estimativa é o erro quadrático médio. Por definição,

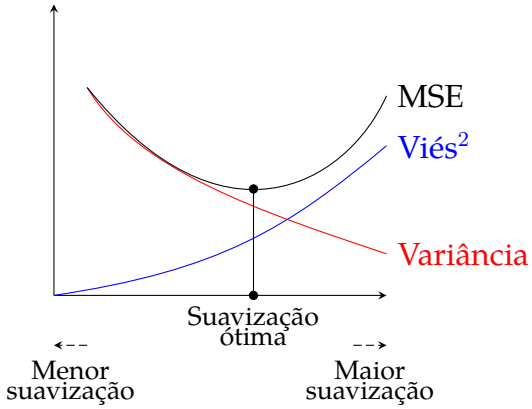
$$\text{MSE}\left[\hat{f}(x)\right] := \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right]$$
$$\text{MSE}\left[\widehat{f}(x)\right] := \text{Viés}\left(\widehat{f}(x)\right)^2 + \text{V}\left[\widehat{f}(x)\right] \tag{2.5.1}$$

Vamos considerar o erro quadrático médio integrado (integrando no suporte comum e em relação a x):

$$\begin{aligned}
\text{MISE} &= \frac{h^4}{4}\int_{-\infty}^{\infty}\left(f''(x)\right)^2 dx\left(\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\,ds\right)^2 + o\left(h^4\right) + \\
&+ \frac{1}{nh}\int_{-\infty}^{\infty} f(x)dx\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds + \frac{1}{nh}o(1) \\
&= \frac{h^4}{4}\int_{-\infty}^{\infty}\left(f''(x)\right)^2 dx\left(\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\,ds\right)^2 + o\left(h^4\right) + \frac{1}{nh}\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds + \\
&+ \frac{1}{nh}o(1)
\end{aligned} \tag{2.5.2}$$

Assim, o valor absoluto do viés de $\widehat{f}$ aumenta e a variância de $\widehat{f}$ decresce quando h aumenta.

Figura 2.5.1 – Trade-off Viés-Variância



Os termos abaixo compõem o AMISE (erro quadrático médio assintótico integrado):

$$\text{AMISE} = \frac{h^4}{4} \int_{-\infty}^{\infty} \left(f''(x)\right)^2 dx \left(\int_{-\infty}^{\infty} s^2 \mathcal{K}(s)\, ds \right)^2 + \frac{1}{nh} \int_{-\infty}^{\infty} \mathcal{K}^2(s)\, ds \qquad (2.5.3)$$

em que $\int_{-\infty}^{\infty} \left(f''(x)\right)^2 dx$ mede a curvatura global.

2.6 Seleção da Bandwidth

Como vimos anteriormente, o estimador de densidade *kernel* depende criticamente da *bandwidth* empregada. O objetivo desta seção é apresentar seletores de *bandwidth* objetivos e automáticos que tentam minimizar o erro de estimativa da densidade alvo f.

Para a escolha da *bandwidth*, sejam os seguintes pressupostos:

1. $f(\cdot)$ é $\mathcal{C}^2$
2. $\mathcal{K}(\cdot)$ é uma densidade simétrica ao redor de zero com as seguintes propriedades:
   (a) $\int \mathcal{K}^2(y)dy < +\infty$
   (b) $\int y^2\mathcal{K}(y)dy < +\infty$
   (c) Suporte comum em $[-1, 1]$
3. $h = h_n \to 0$ quando $n \to \infty$
4. $nh_n \to \infty$ quando $n \to \infty$
5. $x_1, \ldots, x_n$ é uma sequência i.i.d.

Derivando o AMISE com relação à h e igualando a zero, encontramos a *bandwidth* ótima:

$$h^3 \int_{-\infty}^{\infty} \left(f''(x)\right)^2 dx \left(\int_{-\infty}^{\infty} s^2 \mathcal{K}(s) ds \right)^2 - \frac{1}{nh^2} \int_{-\infty}^{\infty} \mathcal{K}^2(s)\, ds = 0$$

$$h^5 = n^{-1} \left(\frac{\int_{-\infty}^{\infty} \mathcal{K}^2(s)\, ds}{\left(\int_{-\infty}^{\infty} s^2 \mathcal{K}(s)\, ds \right)^2 \int_{-\infty}^{\infty} \left(f''(x) \right)^2 dx} \right)$$

$$h^* = n^{-1/5} \left(\frac{\int_{-\infty}^{\infty} \mathcal{K}^2(s)\, ds}{\left(\int_{-\infty}^{\infty} s^2 \mathcal{K}(s)\, ds \right)^2 \int_{-\infty}^{\infty} \left(f''(x) \right)^2 dx} \right)^{1/5} \tag{2.6.1}$$

Observe que r_n, a taxa de convergência, de h^* é $n^{-1/5}$.

No entanto, h^* para a *bandwidth* ótima não pode ser usado na prática porque envolve a quantidade desconhecida $\int_{-\infty}^{\infty} \left(f''(x) \right)^2 dx$. Assim, escolher h é um problema não resolvido em estatística e é conhecido como seleção de *bandwidth*. A maioria das abordagens de seleção de *bandwidth* propõe uma estimativa de AMISE e então minimizar o AMISE estimado ou usar uma estimativa da curvatura e escolhendo h^* de acordo.

Uma solução simples para estimar f'' é supor que f é a densidade de uma distribuição normal[5] $\mathcal{N}(\mu, \sigma^2)$. Assim, temos que: $\int \left(f''(x) \right)^2 dx = \frac{3}{8\sqrt{\pi}} \frac{1}{\sigma^5}$. Ao fazer isso, aproximamos a curvatura de uma densidade arbitrária por meio da curvatura de uma normal e temos que

$$h^* = \sigma \left[\frac{8\sqrt{\pi} R(\mathcal{K})}{3\mu_2^2(\mathcal{K}) N} \right]^{1/5} \approx 1.06 n^{-1/5} \sigma \tag{2.6.2}$$

Curiosamente, a *bandwidth* é diretamente proporcional ao desvio padrão da densidade alvo. Substituir σ por uma estimativa produz o seletor da *bandwidth* da escala normal, que denotamos por

[5] Usamos apenas uma suposição paramétrica para estimar a curvatura de f no h_{AMISE}, não para estimar diretamente a próprio f.

$$\hat{h}_{NS} = \hat{\sigma} \left[\frac{8\sqrt{\pi} R(\mathcal{K})}{3\mu_2^2(\mathcal{K}) N} \right]^{1/5} \tag{2.6.3}$$

para enfatizar sua aleatoriedade.

A estimativa de $\hat{\sigma}$ pode ser escolhida como o desvio padrão s, ou, a fim de evitar os efeitos de potenciais *outliers*, como o intervalo interquantil padronizado

$$\hat{\sigma}_{IQR} := \frac{X_{([0.75n])} - X_{([0.25n])}}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)} \tag{2.6.4}$$

ou

$$\hat{\sigma} = \min(s, \hat{\sigma}_{IQR}) \tag{2.6.5}$$

Agora voltamos nossa atenção para uma filosofia diferente de estimativa de *bandwidth*. Em vez de tentar minimizar o AMISE plugando estimativas para o termo de curvatura desconhecido, tentamos minimizar o MISE diretamente. A ideia é usar a amostra duas vezes: uma para calcular o estimador de densidade *kernel* e outra para avaliar seu desempenho na estimativa de f. Para evitar a dependência clara da amostra, fazemos esta avaliação de uma forma de validação cruzada: os dados usados para calcular o estimador de densidade *kernel* não são usados para sua avaliação.

Começamos com o erro quadrático médio integrado:

$$\begin{aligned} \mathrm{MISE}[\hat{f}(\cdot; h)] &= \mathbb{E}\left[\int \left(\hat{f}(x; h) - f(x) \right)^2 dx \right] \\ &= \mathbb{E}\left[\int \hat{f}(x; h)^2 dx \right] - 2\mathbb{E}\left[\int \hat{f}(x; h) f(x) dx \right] + \int f(x)^2 dx \end{aligned} \tag{2.6.6}$$

Dado que o último termo não depende de h, minimizar o MISE é equivalente a minimizar

$$\mathbb{E}\left[\int \hat{f}(x; h)^2 dx \right] - 2\mathbb{E}\left[\int \hat{f}(x; h) f(x) dx \right] \tag{2.6.7}$$

Esta é uma quantidade desconhecida, mas pode ser estimada não-viesadamente por

$$LSCV(h) := \int \hat{f}(x;h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i;h) \tag{2.6.8}$$

em que

$$\hat{f}_{-i}(x;h) = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} K_h(x - X_j) \tag{2.6.9}$$

Assim,

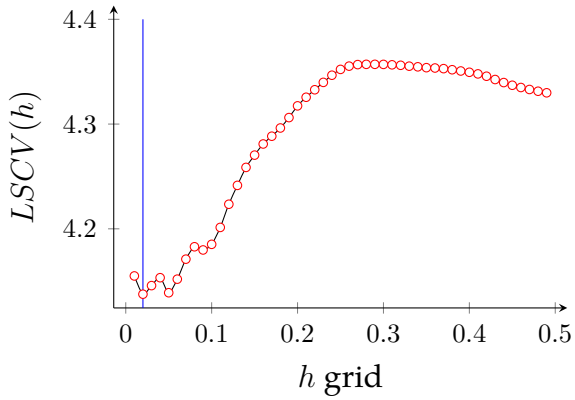
$$\hat{h}_{LSCV} := \arg\min_{h>0} LSCV(h) \tag{2.6.10}$$

A otimização numérica é necessária para obter h, ao contrário dos métodos por meio de *plug-in*, e há pouco controle sobre o formato da função objetivo. A otimização numérica da função LSCV pode ser desafiadora. Na prática, vários mínimos locais são possíveis, e a rugosidade da função objetivo pode variar notavelmente dependendo de n e de f. Como consequência, as rotinas de otimização podem ficar presas em soluções espúrias. Para estar no lado seguro, é sempre aconselhável verificar (quando possível) a solução plotando $LSCV(h)$ para um intervalo de h, ou realizar uma pesquisa em uma grade de *bandwidth*: $\hat{h}_{LSCV} \approx \arg\min_{h_1,\ldots,h_G} LSCV(h)$.

Uma justificativa para o uso da *cross-validation* é dado pelo seguinte teorema:

Teorema 2.6.1 (Teorema de Stone)**.** *Suponha que $f(x)$ é limitada. Seja $\hat{f}_h(x)$ o estimador kernel com bandwidth h e seja $\hat{h}$ a bandwidth escolhida por CV. Então,*

$$\frac{\int \left(f(x) - \hat{f}_{\hat{h}}(x)\right)^2 dx}{\inf_h \int \left(f(x) - \hat{f}_{\hat{h}}(x)\right)^2 dx} \xrightarrow{a.s.} 1 \tag{2.6.11}$$



FONTE.— Cameron Blevins, Paper Trails: The US Post and the Making of the American West. New York: Oxford University Press, 2021. Conjunto de dados históricos e espaciais de 166.140 correios que operaram nos Estados Unidos entre 1639–2000.

2.7 Intervalo de Confiança

As regiões de confiança da função de densidade são intervalos aleatórios $C_{1-\alpha}(x)$ derivados da amostra, de modo que $C_{1-\alpha}(x)$ cobre o valor verdadeiro de $f(x)$ com probabilidade de pelo menos $1-\alpha$. Com base nessa noção, existem dois tipos comuns de regiões de confiança:

1. Intervalo de confiança: para um dado x, o conjunto $C_{1-\alpha}(x)$ satisfaz

$$\mathbb{P}[f(x) \in C_{1-\alpha}(x)] \geq 1-\alpha \tag{2.7.1}$$

2. Banda de confiança: o intervalo $C_{1-\alpha}(x)$ satisfaz

$$\mathbb{P}[f(x) \in C_{1-\alpha}(x),\ \forall x \in K] \geq 1-\alpha \tag{2.7.2}$$

Ou seja, os intervalos de confiança são regiões de confiança apenas com cobertura local e as bandas de confiança são regiões de confiança com cobertura simultânea. Se um intervalo/banda de confiança tiver apenas cobertura $1-\alpha+o(1)$, será denominado intervalo/banda de confiança $1-\alpha$ assintoticamente válido.

Para um dado ponto x

$$\sqrt{nh}\left\{\hat{f}_n(x) - \mathbb{E}\left[\hat{f}_n(x)\right]\right\} = \sqrt{nh}\mathscr{E}_n \xrightarrow{d} \mathcal{N}\left(0, \sigma_f^2(x)\right) \tag{2.7.3}$$

uma banda de confiança usando normalidade assintótica pelo método do *plug-in* pode ser dada por

$$C_{1-\alpha}(x) = \left(\hat{f}_n(x) - z_{1-\alpha/2}\sqrt{\frac{\mu_K \hat{f}_n(x)}{nh}}, \hat{f}_n(x) + z_{1-\alpha/2}\sqrt{\frac{\mu_K \hat{f}_n(x)}{nh}}\right) \tag{2.7.4}$$

em que $\mu_K = \int \mathcal{K}^2(x)dx$. Observe que esse é um intervalo de confiança para $\mathbb{E}\left[\hat{f}_n(x)\right]$ e não para $f(x)$.

Lembre-se da análise de viés, o viés é da ordem de $\mathrm{O}(h^2)$. Assim, se h for fixo ou h convergir para o lentamente, a cobertura de CI será menor que a cobertura nominal (isso é chamado de *undercoverag*). Ou seja, mesmo se construirmos um IC de 99%, a chance deste IC cobrir o valor real da densidade pode ser de apenas 1% ou até menor.

Chamamos esse método de "método *plug-in*" porque plugamos o estimador de variância para construir um intervalo de confiança. Quando $h \to 0$, $nh \to \infty$, $\hat{f}_n(x)$ é um estimador consistente de $f(x)$. Como resultado,

$$\mathbb{P}\left[\mathbb{E}(\hat{f}_n(x)) \in C_{1-\alpha}(x)\right] = 1 - \alpha + o(1) \tag{2.7.5}$$

Um método alternativo é estimar a variância assintótica usando *bootstrap* (Efron, 1979). Em mais detalhes, usamos o *bootstrap* empírico para gerar a amostra de *bootstrap* $X_1^*, \ldots, X_n^*$. Então, aplicamos o estimador de densidade *kernel* a amostra de *bootstrap*. Quando repetimos o *bootstrap* B vezes, nós temos B estimadores de densidade *kernel* por *bootstrap* $\hat{f}_n^{(1)}(x), \ldots, \hat{f}_n^{(B)}(x)$. Assuma,

$$\hat{\sigma}_f^2(x) = \frac{1}{B-1}\sum_{j=1}^{B}\left[\hat{f}_n^{(j)}(x) - \overline{f}_n^*(x)\right]^2 \tag{2.7.6}$$

em que $\overline{f}_n^*(x) = \frac{1}{B}\sum_{j=1}^{B} \hat{f}_n^{(j)}(x)$ é a média amostral dos estimadores de densidade *kernel* por *bootstrap*. Um intervalo de confiança por *bootstrap* de $1-\alpha$ é dado por

$$C_{1-\alpha}(x) = \left(\hat{f}_n(x) - z_{1-\alpha/2}\hat{\sigma}_f^2(x), \hat{f}_n(x) + z_{1-\alpha/2}\hat{\sigma}_f^2(x)\right) \tag{2.7.7}$$

Como o estimador de variância *bootstrap* $\hat{\sigma}_f^2(x)$ converge para $\frac{\sigma_f^2(x)}{nh}$ no sentido de que

$$\frac{\hat{\sigma}_f^2(x)}{\sigma_f^2(x)/nh} \xrightarrow{P} 1 \tag{2.7.8}$$

o estimador de variância *bootstrap* é consistente, então o intervalo de confiança também será consistente:

$$\mathbb{P}\left[\mathbb{E}(\hat{f}_n(x)) \in C_{1-\alpha}(x)\right] = 1 - \alpha + o(1) \tag{2.7.9}$$

Embora todos esses intervalos sejam válidos para cada ponto determinado, não há garantia de que cobrirão toda a função de densidade simultaneamente. Assim, apresentamos métodos de construção de bandas de confiança (regiões de confiança com cobertura simultânea). Dito de outo modo, uma banda de confiança é um conjunto de confiança para uma função como um todo.

Agora, apresentamos métodos de construção de bandas de variabilidade. A ideia chave é aproximar a distribuição do erro uniforme[6] $\sup_x|\hat{f}_n(x) - f(x)|$ e então convertê-lo em uma banda de confiança. Para ser mais específico, seja $G(t) = \mathbb{P}\left[\sup_x|\hat{f}_n(x) - f(x)| < t\right]$ a função de distribuição acumulada do erro, e seja $\overline{c}_{1-\alpha} = G^{-1}(1-\alpha)$ o quantil $1-\alpha$. Então, pode ser mostrado que o conjunto

$$\overline{C}(x) = \left[\hat{f}_n(x) - \overline{c}_{1-\alpha}, \hat{f}_n(x) + \overline{c}_{1-\alpha}\right] \tag{2.7.10}$$

[6] É uma métrica do erro (também conhecido como L_∞) e é dado pela diferença máxima entre $\hat{f}_n$e f, isto é, $\sup_x|\hat{f}_n(x) - f(x)|$.

é uma banda de confiança, isto é,

$$\mathbb{P}\left[f(x) \in \overline{C}(x)\ \forall x \in \mathbb{K}\right] = 1 - \alpha \tag{2.7.11}$$

Portanto, desde que tenhamos uma boa aproximação da distribuição $G(t)$, podemos converter a aproximação em uma banda de confiança.

Um enfoque intuitivo é derivar a distribuição assintótica de $\sup_x |\hat{f}_n(x) - f(x)|$ diretamente e então invertê-la dentro de uma banda de confiabilidade. Bickel e Rosenblatt (1973) provaram que a função perda uniforme converge para uma distribuição de valor extremo no sentido de que

$$\mathbb{P}\left[\sqrt{-2\log h}\left(\sqrt{nh^d}\sup_x \frac{|\hat{f}_n(x) - \mathbb{E}(\hat{f}_n(x))|}{\sqrt{f(x)\mu_K}} - O\left(\sqrt{-2\log h}\right)\right) < t\right] \to \mathrm{e}^{-2\mathrm{e}^{-t}} \tag{2.7.12}$$

em que

$$O\left(\sqrt{-2\log h}\right) = \sqrt{2\delta\log n} + \frac{1}{\sqrt{2\delta\log n}}\left[\log\frac{K_1(\mathcal{K})}{\sqrt{\pi}} + \frac{1}{2}\left(\log\delta + \log\log n\right)\right] \tag{2.7.13}$$

com $0 < \delta < 1$ e $K_1(\mathcal{K}) = \dfrac{\mathcal{K}^2(A) + \mathcal{K}^2(-A)}{2\int K^2(s)ds}$ se $K_1(\mathcal{K}) > 0$ e

$$O\left(\sqrt{-2\log h}\right) = \sqrt{2\delta\log n} + \frac{1}{\sqrt{2\delta\log n}}\log\frac{1}{\pi}\left(\frac{K_2(\mathcal{K})}{2}\right)^{1/2} \tag{2.7.14}$$

com $K_2(\mathcal{K}) = \dfrac{1}{2}\dfrac{\int(\mathcal{K}'(s))^2 ds}{\int K^2(s)ds}$ se $K_1(\mathcal{K}) < 0$. E $[-A, A]$ é o suporte no qual o *kernel* está definido.

Seja $E_{1-\alpha} = -\log\left(-\dfrac{\log\alpha}{2}\right)$ a função de distribuição acumulada à direita do quantil $1 - \alpha$. E seja

$$c_{1-\alpha} = \sqrt{\frac{f(x)\mu_K}{nh^d}}\left(O\left(\sqrt{-2\log h}\right) + \frac{E_{1-\alpha}}{\sqrt{-2\log h}}\right) \tag{2.7.15}$$

Para construir uma banda de confiança substituímos $f(x)$ por $\hat{f}(x)$. Assim, uma banda de confiança de nível $1-\alpha$ é dado por

$$C_{1-\alpha} = \left(\hat{f}_n(x) - c_{1-\alpha}, \hat{f}_n(x) + c_{1-\alpha}\right) \tag{2.7.16}$$

Embora a equação acima seja uma banda de confiança assintoticamente válida, a convergência para a distribuição de valores extremos na equação (2.7.12) é muito. Assim, precisamos de uma amostra de tamanho grande para garantir que a banda de confiança acima seja assintoticamente válida.

2.8 Curse of Dimensionality

Um problema que ocorre com os métodos de suavização é a maldição da dimensionalidade, termo geralmente atribuído a Bellman (1961). Grosso modo, isso significa que a estimativa fica mais difícil muito rapidamente à medida que a dimensão das observações aumenta.

Existem pelo menos duas versões dessa maldição. O primeiro é a maldição computacional da dimensionalidade. Isso se refere ao fato de que a carga computacional de alguns métodos pode aumentar exponencialmente com a dimensão. Nosso foco aqui, entretanto, é com a segunda versão, que chamamos de maldição estatística da dimensionalidade: se os dados têm dimensão d, então precisamos de um tamanho de amostra n que cresce exponencialmente com d. De forma geral, o erro quadrático médio pode ser escrito como

$$\text{MSE} \approx \frac{c}{n^{4/4+d}} \tag{2.8.1}$$

para alguma constante $c > 0$. Se quisermos que o MSE seja igual a um pequeno número δ, podemos definir $\text{MSE} = \delta$ e resolver para n, o que resulta em

$$n \propto \left(\frac{c}{\delta}\right)^{d/4} \tag{2.8.2}$$

que cresce exponencialmente com a dimensão d.

A razão para esse fenômeno é que a suavização envolve estimar uma função $f(x)$ usando pontos de dados em uma vizinhança local de x. Mas em um problema de alta dimensão, os dados são muito esparsos, portanto, as vizinhanças locais contêm muito poucos pontos.

2.9 Script R

```
## DGP ##

set.seed(1991)
n = 500
media = 1
desvio = 2
x = rnorm(n,media,desvio)
alpha = 0.05
delta = 0.4 # 1/5<delta<1/2 (corolário p.1072 Biockel and Rosenblatt)
Rk = 3/5

## Kernels ##

kr = function(x) 0.5*(abs(x)<=1)
ke = function(x) 0.75*(1-x^2)*(abs(x)<=1)
kb = function(x) 0.9375*(1-x^2)^2*(abs(x)<=1)

## Estimador para densidade ##

f.w = function(x,h,k){
   n = length(x)
   f1 = rep(0,n)
   for(i in 1:n){
      f1[i] = sum(k((x-x[i])/h))/(n*h)
   }
   f1
}

h = 1
fr = f.w(x,h,k=kr) # kernel retangular (uniforme)
fe = f.w(x,h,k=ke) # kernel epanechnikov
fb = f.w(x,h,k=kb) # kernel biweight
fp = 1/(desvio*sqrt(2*pi))*exp(-0.5*((x-media)/desvio)^2) # kernel
    paramétrico
```

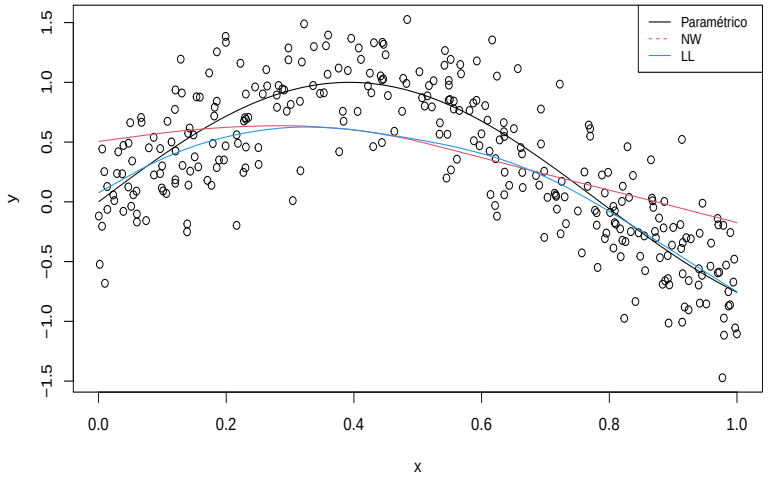
```
plot(sort(x),fp[order(x)],type="l",col=1, ylab="Densidade",xlab="x")
lines(sort(x),fr[order(x)],col=2)
lines(sort(x),fe[order(x)],col=3)
lines(sort(x),fb[order(x)],col=4)
legend("topright",legend=c("Nornal","Kernel Retangular","Kernel
    Epanechnikov","Kernel Biweight"),col=c(1,4,2),lty=1,cex=0.8)

## Bandwidth selection: plug-in (normal distribution) ##

iqr = diff(quantile(x, c(0.25,0.75)))/diff(qnorm(c(0.25, 0.75)))
h.pg = 1.06*n^(-1/5)*min(sd(x), iqr)
f.pg = f.w(x,h.pg,k=dnorm)
dev.off()
plot(sort(x),fp[order(x)],type="l",col=1, ylab="Densidade",xlab="x")
lines(sort(x),f.pg[order(x)],col=2)
legend("topright",legend=c("Nornal", "Kernel
    Plug-in"),col=c(1,2),lty=1,cex=0.8)

## CI CLT ##

fe = f.w(x,h=h.pg,k=ke)
var_kde_hat <- fe*Rk/(n*h.pg)

z_alpha2 <- qnorm(1-alpha/2)
ci_low_clt <- fe - z_alpha2*sqrt(var_kde_hat)
ci_up_clt <- fe + z_alpha2*sqrt(var_kde_hat)

dev.off()
plot(sort(x), fe[order(x)], type="l", col=1, ylab="Densidade", xlab="x",
    ylim=c(min(ci_low_clt), max(ci_up_clt)))
lines(sort(x),type="l",ci_low_clt[order(x)],col=2)
lines(sort(x),type="l",ci_up_clt[order(x)],col=4)
legend("topright",legend=c("Density", "CLT CI Upper","CLT CI
    Lower"),col=c(1,4,2),lty=1,cex=0.8)

## CI Biockel and Rosenblat ##
```

```
e_alpha <- -log(-log(alpha)/2)
sqrt_h <- sqrt(-2*log(h.pg))
aux <- 2*delta*log(n)
o <- sqrt(aux)+1/(aux)*(log(1/(sqrt(pi))))+0.5*(log(delta)+log(log(n)))
var_br <- sqrt(fe*Rk/(n*h.pg))*(o+e_alpha/sqrt_h)
ci_low_br <- fe - var_br
ci_up_br <- fe + var_br

dev.off()
plot(sort(x),fe[order(x)],type="l",col=1, ylab="Densidade",xlab="x",ylim
    = c(min(ci_low_br),max(ci_up_br)))
lines(sort(x),type="l",ci_low_br[order(x)],col=2)
lines(sort(x),type="l",ci_up_br[order(x)],col=4)
legend("topright", legend=c("Density", "BR CI Upper","BR CI
    Lower"),col=c(1,4,2), lty=1, cex=0.8)
```

3. Estimação Não-Paramétrica Univariada de Momentos Condicionais

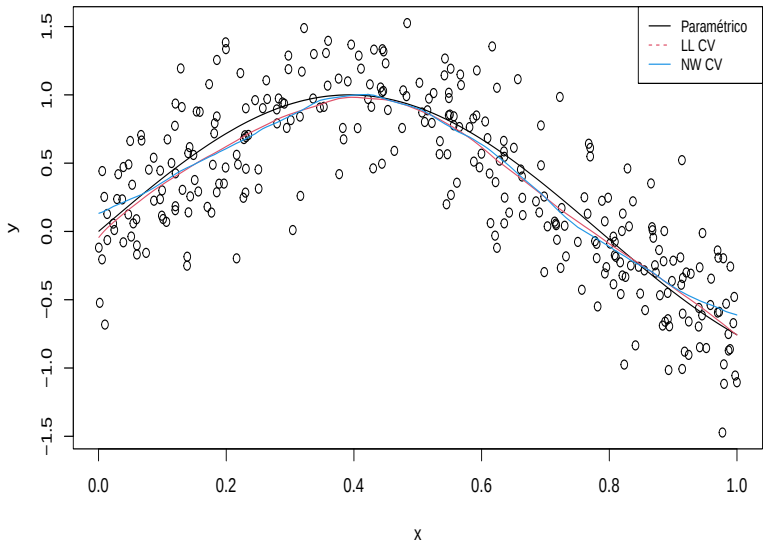
Neste capítulo, descrevemos uma classe de técnicas de regressão que tem flexibilidade na estimativa da função de regressão $f(X)$ sobre o domínio $\mathbb{R}^p$ ajustando um modelo diferente, mas simples, separadamente em cada ponto x_i. Isso é feito usando apenas as observações próximas ao ponto alvo x para ajustar o modelo simples, e de tal forma que a função estimada resultante $\widehat{f}(x)$ seja suave em $\mathbb{R}^p$. Essa localização é obtida por meio de uma função de ponderação ou *kernel* $\mathcal{K}_h(x, x_i)$, que atribui um peso a x_i com base em sua distância de x. Os *kernels* $\mathcal{K}_h$ são normalmente indexados por um parâmetro h. Abaixo segue o exemplo de dois estimadores que vamos estudar.

Figura 3.0.1 – Estimador NW e LL com h Ad Hoc



Nota: No painel, 300 pares (x_i, y_i) são gerados aleatoriamente como $Y = \sin(4X) + \varepsilon$, em que $X \sim U[0,1]$ e $\varepsilon \sim N\left(0, 1/3\right)$.

Figura 3.0.2 – Estimador NW e LL com h Ótimo Selecionado por CV



Nota: No painel, 300 pares (x_i, y_i) são gerados aleatoriamente como $Y = \sin(4X) + \varepsilon$, em que $X \sim U[0,1]$ e $\varepsilon \sim N\left(0, 1/3\right)$.

3.1 Estimador Nadaraya-Watson

Nosso objetivo é estimar uma função de regressão $m: \mathbb{R} \to \mathbb{R}$ não-parametricamente. Assim, seja o seguinte modelo

$$Y_i = m(X_i) + \varepsilon_i, \quad (X_i, Y_i) \text{ é i.i.d.} \tag{3.1.1}$$

e tome o valor esperado

$$\begin{aligned} \mathbb{E}(Y_i|X_i = x) &= m(x) + \underbrace{\mathbb{E}(\varepsilon_i|X_i = x)}_{=0} \\ &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\, dy \\ &= \int_{-\infty}^{\infty} y \frac{f_{Y,X}(y,x)}{f_X(x)}\, dy \end{aligned} \tag{3.1.2}$$

A expressão (3.1.2) mostra um ponto interessante: a função de regressão pode ser calculada a partir da densidade conjunta $f_{Y,X}(x,y)$ e da marginal $f_X(x)$. Portanto, dada uma amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$ uma estimativa não paramétrica de m pode ser obtida substituindo-se as densidades anteriores por seus estimadores de densidade kernel:

$$\begin{aligned} \widehat{m}(x) &= \frac{1}{\widehat{f}_X(x)} \int_{-\infty}^{\infty} y \frac{1}{nh_xh_y} \sum_{i=1}^{n} \mathcal{K}_x\left(\frac{x_i - x}{h_x}\right) \mathcal{K}_y\left(\frac{y_i - y}{h_y}\right) dy \\ &= \frac{1}{\widehat{f}_X(x) nh_xh_y} \sum_{i=1}^{n} \mathcal{K}_x\left(\frac{x_i - x}{h_x}\right) \int_{-\infty}^{\infty} y \mathcal{K}_y\left(\frac{y_i - y}{h_y}\right) dy \\ &= \frac{1}{\widehat{f}_X(x) nh_xh_y} \sum_{i=1}^{n} \mathcal{K}_x\left(\frac{x_i - x}{h_x}\right) \int_{-\infty}^{\infty} (y_i + sh_y) \underbrace{\mathcal{K}_y(-s)}_{=\mathcal{K}_y(s)} h_y\, ds \end{aligned}$$

$$= \frac{1}{\widehat{f}_X(x) n h_x h_y} \sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right) y_i h_y \underbrace{\int_{-\infty}^{\infty} \mathcal{K}_y(s)\, ds}_{=1} + h_y^2 \underbrace{\int_{-\infty}^{\infty} s \mathcal{K}_y(s)\, ds}_{=0}$$

$$= \frac{1}{\widehat{f}_X(x) n h_x h_y} \sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right) y_i h_y \tag{3.1.3}$$

Sabemos que:

$$\widehat{f}_X(x) = \frac{1}{n h_x} \sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right) \tag{3.1.4}$$

Assim,

$$\begin{aligned} \widehat{m}(x)_{NW} &= \frac{n h_x}{n h_x \sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right)} \sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right) y_i \\ &= \frac{\sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right) y_i}{\sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right)} \\ &= \sum_{i=1}^{n} \Omega_i(x) y_i \end{aligned} \tag{3.1.5}$$

em que $\Omega_i(x) = \dfrac{\mathcal{K}_x \left(\frac{x_i - x}{h_x} \right)}{\sum_{i=1}^{n} \mathcal{K}_x \left(\frac{x_i - x}{h_x} \right)}$, $\Omega_i(x) \geq 0$ e $\sum_{i=1}^{n} \Omega_i = 1$.

Observe que o estimador não depende de h_y; em vez disso, depende de h_x, a largura de banda empregada para suavizar x.

Podemos ver esse estimador de outra forma também. Seja,

$$
\begin{aligned}
\widehat{m}(x)_{NW} &= \frac{\sum_{i=1}^{n} \mathcal{K}_x\left(\frac{x_i - x}{h_x}\right) y_i}{\sum_{i=1}^{n} \mathcal{K}_x\left(\frac{x_i - x}{h_x}\right)} \\
&= \frac{\sum_{|X_i - x| \leq h} y_i \left(\frac{1}{2}\right)}{\sum_{|X_i - x| \leq h} \left(\frac{1}{2}\right)} \\
&= \frac{\sum_{|X_i - x| \leq h} [m(X_i) + \varepsilon_i]}{\sum_{|X_i - x| \leq h} 1} \\
&= \{\text{média dos } m(X_i)'s + \text{média dos } \varepsilon_i's \text{ com } |X_i - x| \leq h\} \\
&\xrightarrow{p} g(x) + 0 = g(x) \qquad (3.1.6)
\end{aligned}
$$

porque $|g(X_i) - g(x)| = O(h) = o(1)$, uma vez que $g(x)$ tem derivada limitada em x. De fato, $\frac{1}{n} \sum_{|X_i - x| \leq h} [g(X_i) - g(x)] = O_p(h^2)$. A média dos ε_i's em cada intervalo (de tamanho nh) converge para sua média populacional $\mathbb{E}[\varepsilon] = 0$ (em probabilidade) pela lei dos grandes números porque $nh \to \infty$ quando $n \to \infty$.

Observe que para a derivação do estimador Nadaraya–Watson não assumimos nenhuma suposição particular, além da diferenciabilidade (implícita) de m até a ordem p para o estimador polinomial local. As seguintes suposições são os únicos requisitos para realizar a análise assintótica do estimador:

1. m é duas vezes continuamente diferenciável[1].
2. σ^2 é contínuo e positivo[2].
3. a função distribuição de probabilidade marginal de X, é continuamente diferenciável e limitada a partir de zero[3].

[1] Essa suposição requer certa suavidade da função de regressão, permitindo assim que as expansões de Taylor sejam realizadas. Esta suposição é importante na prática: $\hat{m}(\cdot; p, h)$ é infinitamente diferenciável se os kernels $\mathcal{K}$ considerados também o forem

[2] Isso evita que a variável aleatória Y tenha uma distribuição degenerada.

[3] Isso evita a situação degenerada na qual m é estimada em regiões sem observações dos preditores.

4. o kernel $\mathcal{K}$ é uma fdp simétrica e limitada com segundo momento finito e é quadrado integrável.

5. $h = h_n$ é uma sequência determinística de *bandwidths* tal que, quando $n \to \infty$, $h \to 0$ e $nh \to \infty$.

O viés e a variância são estudados em suas versões condicionais na amostra do preditor $X_1, \ldots, X_n$. A razão para analisar as versões condicionais em vez das incondicionais é evitar dificuldades técnicas que a integração em relação à densidade do preditor desconhecido pode representar. Isso está no espírito do que é feito na inferência paramétrica

Para avaliar a consistência de $\widehat{m}_{NW}(x)$, seja

$$\mathbb{E}\left(\widehat{m}_{NW}(x)|X\right) = \frac{\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right) m(x_i)}{\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)} \tag{3.1.7}$$

em que $m(x_i) = m(x) + m'(x)(x_i - x) + \frac{m''(x)}{2}(x_i - x)^2 + o_p\left((x_i - x)^2\right)$.

Assim,

$$\begin{aligned}\mathbb{E}\left(\widehat{m}_{NW}(x)|X\right) = m(x) &+ m'(x)\left[\frac{\frac{1}{nh_n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)(x_i - x)}{\frac{1}{nh_n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)}\right] + \\ &+ \frac{m''(x)}{2}\left[\frac{\frac{1}{nh_n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)(x_i - x)^2}{\frac{1}{nh_n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)}\right] + \\ &+ o_p(1)\left[\frac{\frac{1}{nh_n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)(x_i - x)^2}{\frac{1}{nh_n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h_n}\right)}\right]\end{aligned} \tag{3.1.8}$$

Seja a expressão para a variância condicional:

$$
\begin{aligned}
\mathrm{V}\left[\widehat{m}_{NW}(x)|X\right] &= \mathrm{V}\left[\left.\frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i-x}{h_n}\right)y_i}{\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i-x}{h_n}\right)}\right| X\right] \\
&= \frac{1}{\left[\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i-x}{h_n}\right)\right]^2}\sum_{i=1}^{n}\mathrm{V}\left[\left.\mathcal{K}\left(\frac{x_i-x}{h_n}\right)y_i\right| X\right] \\
&= \frac{1}{\left[\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i-x}{h_n}\right)\right]^2}\frac{1}{nh}\left(f(x)\int_{-\infty}^{\infty}\mathcal{K}^2(s)ds+o(1)\right)\sigma^2(x_i) \\
&= \frac{1}{\left[\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i-x}{h_n}\right)\right]^2}\frac{1}{nh}\left(f(x)\int_{-\infty}^{\infty}\mathcal{K}^2(s)ds+o(1)\right)\left[\sigma^2(x)+o_p(1)\right] \\
&= \frac{1}{nh_n}\frac{\sigma^2(x)}{f(x)}\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds+\frac{1}{nh_n}o_p(1) \qquad (3.1.9)
\end{aligned}
$$

O AMSE é obtido como segue:

$$
\begin{aligned}
\mathrm{AMSE}\left(\widehat{m}_{NW}(x)|X\right) = &\left(\frac{m'(x)h_n^2\mu_2 f'(x)}{f(x)}+\frac{m''(x)}{2}h_n^2\mu_2\right)^2+ \\
&\frac{1}{nh_n}\frac{\sigma^2(x)}{f(x)}\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds \qquad (3.1.10)
\end{aligned}
$$

em que $\mu_2 = \int_{-\infty}^{\infty} s^2\mathcal{K}(s)\,ds$.

E o AMISE é dado por

$$
\mathrm{AMISE}\left(\widehat{m}_{NW}(x)|X\right) = h^4\mu_2^2\int\left(\frac{m'(x)f_x'(x)}{f_x(x)}+\frac{m''(x)}{2}\right)^2 f(x)dx+
$$

$$+ \frac{1}{nh} \int \mathcal{K}^2(s)\, ds \int \sigma^2(x)\, dx \tag{3.1.11}$$

3.2 Estimadores Polinomiais Locais

Existe outro método local, a regressão linear local, que é superior a regressão do *kernel* (Nadaraya-Watson). Em geral, o estimador LL remove o termo de viés do estimador de *kernel*, o que faz com que tenha o MSE menor em todos os pontos. O estimador Nadaraya–Watson pode ser visto como um caso particular de uma classe mais ampla de estimadores não paramétricos, os chamados estimadores polinomiais locais. Especificamente, Nadaraya–Watson é aquele que corresponde a realizar um ajuste constante local. Vamos ver essa classe mais ampla de estimadores não paramétricos e suas vantagens em relação ao estimador Nadaraya-Watson.

Uma motivação para o ajuste polinomial local vem da tentativa de encontrar um estimador $\hat{m}$ de m que "minimize" a soma dos quadrados dos erros sem assumir qualquer forma particular para o m subjacente. Isso não é alcançável diretamente, uma vez que nenhum conhecimento sobre m está disponível.

O método LL é baseado no seguinte problema de minimização:

$$\min_{\beta_0^x, \beta_1^x} \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0^x - \widehat{\beta}_1^x (X_i - x) \right)^2 \mathcal{K} \left(\frac{X_i - x}{h_n} \right) \tag{3.2.1}$$

No estimador *linear local* definimos uma subamostra (se $\mathcal{K}(\cdot)$ tiver suporte compacto), mas os pesos dependem de x. Na regressão *kernel* mais tradicional, a constante β_0 é restrita a ser zero. Aqui,

$$\widehat{m}_{LL}(x) = \widehat{\beta}_0 \tag{3.2.2}$$

Podemos reescrever o problema de otimização como segue

$$\min_{\beta_x} \left(y - X_x \widehat{\beta} \right)' W_x \left(y - X_x \widehat{\beta} \right) \tag{3.2.3}$$

em que

$$X_x = \begin{pmatrix} 1 & (X_1 - x) \\ 1 & (X_2 - x) \\ \vdots & \vdots \\ 1 & (X_n - x) \end{pmatrix}_{n \times (p+1)} \tag{3.2.4}$$

e p é a ordem do polinômio e

$$W_x = \begin{pmatrix} \mathcal{K}_h(X_1 - x) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{K}_h(X_n - x) \end{pmatrix} \tag{3.2.5}$$

O estimador de interesse é dado por:

$$\widehat{\beta}_x = \left(X_x' W_x X_x\right)^{-1} \left(X_x' W_x y\right) \tag{3.2.6}$$

Assim,

$$\widehat{m}_{LL}(x) = e_1' \widehat{\beta}_x \tag{3.2.7}$$

em que $e_1' = \begin{pmatrix} 1 & 0 \end{pmatrix}$ captura $\widehat{\beta}_0$.

Considere o seguinte modelo,

$$\boldsymbol{y} = \boldsymbol{M} + \boldsymbol{\varepsilon} \tag{3.2.8}$$

em que

$$M = \begin{bmatrix} m(X_1) \\ \vdots \\ m(X_n) \end{bmatrix}_{n \times 1} \tag{3.2.9}$$

Quando m não tem parametrização disponível e pode adotar qualquer forma

matemática, é necessária uma abordagem alternativa. O primeiro passo é induzir uma parametrização local em m, utilizando a expansão de Taylor:

$$m(x_i) = m(x) + m'(x)(x_i - x) + \frac{m''(x)}{2}(x_i - x)^2 + o\left[(x_i - x)^2\right] \tag{3.2.10}$$

Escrevendo em formato matricial

$$m(x_i) = \begin{bmatrix} 1 & (X_i - x) \end{bmatrix} \begin{bmatrix} m(x) \\ m'(x) \end{bmatrix} \tag{3.2.11}$$

e

$$M = X_x \begin{bmatrix} m(x) \\ m'(x) \end{bmatrix} \tag{3.2.12}$$

Retome o estimador condicional

$$\begin{aligned} \mathbb{E}\left[\widehat{m}_{LL}(x) \middle| X\right] &= \mathbb{E}\left[\left(e_1' \left(X_x' W_x X_x\right)^{-1} X_x' W_x M + e_1' \left(X_x' W_x X_x\right)^{-1} X_x' W_x \varepsilon \right) \middle| X \right] \\ &= e_1' \left(X_x' W_x X_x\right)^{-1} X_x' W_x M \end{aligned} \tag{3.2.13}$$

Então,

$$\begin{aligned} \mathbb{E}\left[\widehat{m}_{LL}(x) | X\right] = e_1' \underbrace{\left(X_x' W_x X_x\right)^{-1} \left(X_x' W_x X_x\right)}_{=I} \begin{bmatrix} m(x) \\ m'(x) \end{bmatrix} + \\ + \frac{m''(x)}{2} e_1' \left(X_x' W_x X_x\right)^{-1} X_x' W_x \begin{bmatrix} (X_i - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} + o_p \begin{bmatrix} (X_i - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} \end{aligned} \tag{3.2.14}$$

Logo,

$$\mathbb{E}\left[\widehat{m}_{LL}(x)|X\right] = m(x) + \frac{m^{''}(x)}{2} e_1' \left(n^{-1} X_x' W_x X_x\right)^{-1} n^{-1} X_x' W_x \begin{bmatrix} (X_i - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} +$$

$$+ o_p \begin{bmatrix} (X_i - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} \tag{3.2.15}$$

Observe que:

$$n^{-1} X_x' W_x X_x = n^{-1} \begin{bmatrix} 1 & \dots & 1 \\ (X_1 - x) & \dots & (X_n - x) \end{bmatrix} \begin{bmatrix} \mathcal{K}_h(X_1 - x) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{K}_h(X_n - x) \end{bmatrix} \times$$

$$\times \begin{bmatrix} 1 & (X_1 - x) \\ \vdots & \vdots \\ 1 & (X_n - x) \end{bmatrix}$$

$$= n^{-1} \begin{bmatrix} \sum_{i=1}^{n} \mathcal{K}_h(X_i - x) & \sum_{i=1}^{n} \mathcal{K}_h(X_i - x)(X_i - x) \\ \sum_{i=1}^{n} \mathcal{K}_h(X_i - x)(X_i - x) & \sum_{i=1}^{n} \mathcal{K}_h(X_i - x)(X_i - x)^2 \end{bmatrix}$$

$$= n^{-1} \begin{bmatrix} S_0 & S_1' \\ S_1 & S_2 \end{bmatrix} \tag{3.2.16}$$

Também observe que

$$n^{-1} X_x{}' W_x \begin{bmatrix} (X_i - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} = n^{-1} \begin{bmatrix} 1 & \dots & 1 \\ (X_1 - x) & \dots & (X_n - x) \end{bmatrix} \times$$

$$\times \begin{bmatrix} \mathcal{K}_h(X_1 - x) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{K}_h(X_n - x) \end{bmatrix} \begin{bmatrix} (X_1 - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix}$$

$$= n^{-1} \begin{bmatrix} \sum_{i=1}^{n} \mathcal{K}_h(X_i - x)(X_i - x)^2 \\ \sum_{i=1}^{n} \mathcal{K}_h(X_i - x)(X_i - x)^3 \end{bmatrix}$$
$$= n^{-1} \begin{bmatrix} S_2 \\ S_3 \end{bmatrix} \tag{3.2.17}$$

Defina

$$\widehat{S}_\ell(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_h(X_i - x)(X_i - x)^\ell, \text{ com } \ell = 0, 1, 2, \ldots \tag{3.2.18}$$

Tomando o valor esperado para $\ell = 1$, temos

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x_i - x}{h}\right)(x_i - x)\right] &= \frac{1}{nh} n \mathbb{E}\left[\mathcal{K}\left(\frac{x_i - x}{h}\right)(x_i - x)\right] \\
&= \frac{1}{h}\int_{-\infty}^{\infty} \mathcal{K}\left(\frac{t - x}{h}\right)(sh)\, f_x(t)\, dt \\
&= \int_{-\infty}^{\infty} \mathcal{K}(s) s f_x(x + sh) h\, ds \\
&= \int_{-\infty}^{\infty} \mathcal{K}(s) s \left(f(x) + f'(x)hs + o(sh)\right) h\, ds \\
&= hf(x)\int_{-\infty}^{\infty} s\mathcal{K}(s)\, ds + h^2 f'(x)\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\, ds + \\
&+ h^2 o(1)\int_{-\infty}^{\infty} s^2\mathcal{K}(s)\, ds \\
&= h^2 f'(x)\mu_2 + o(h^2)
\end{aligned} \tag{3.2.19}$$

em que $\mu_2 = \int_{-\infty}^{\infty} s^2\mathcal{K}(s)\, ds$.

Tomando o valor esperado para $\ell = 2$, temos

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n}\mathcal{K}\left(\frac{x_i-x}{h}\right)(x_i-x)\right] &= \frac{1}{nh}n\mathbb{E}\left[\mathcal{K}\left(\frac{x_i-x}{h}\right)(t-x)^2\right] \\
&= \frac{1}{h}\int_{-\infty}^{\infty}\mathcal{K}\left(\frac{t-x}{h}\right)(sh)^2 f_x(t)\,dt \\
&= \frac{1}{h}\int_{-\infty}^{\infty}\mathcal{K}(s)(sh)^2 f_x(x+sh)h\,ds \\
&= \int_{-\infty}^{\infty}\mathcal{K}(s)(sh)^2\left(f(x)+f'(x)sh+o(1)\right)\,ds \\
&= h^2 f(x)\int_{-\infty}^{\infty}s^2\mathcal{K}(s)\,ds + h^3 f'(x)\int_{-\infty}^{\infty}s^3\mathcal{K}(s)\,ds+ \\
&+ h^2 o(1)\int_{-\infty}^{\infty}s^2\mathcal{K}(s)\,ds \\
&= h^2 f(x)\mu_2 + o\left(h^2\right) \qquad (3.2.20)
\end{aligned}
$$

Generalizando,

$$
\mathbb{E}\left[\widehat{S}_\ell(x)\right] = \begin{cases} h^\ell \mu^\ell f(x) + o_p(h^\ell), & \text{para } \ell \text{ par (inclui } \ell = 0) \\ h^{\ell+1}\mu^{\ell+1} f'(x) + o_p(h^{\ell+1}), & \text{para } \ell \text{ ímpar} \end{cases} \qquad (3.2.21)
$$

Logo,

$$
n^{-1}X_x{}'W_x X_x = \begin{bmatrix} f(x)+o_p(1) & h^2\mu^2 f'(x)+o_p\left(h^2\right) \\ h^2\mu^2 f'(x)+o_p\left(h^2\right) & h^2\mu^2 f(x)+o_p\left(h^2\right) \end{bmatrix} \qquad (3.2.22)
$$

e

$$
n^{-1}X_x{}'W_x = \begin{bmatrix} h^2\mu^2 f(x)+o_p\left(h^2\right) \\ h^4\mu^4 f'(x)+o_p\left(h^4\right) \end{bmatrix} \qquad (3.2.23)
$$

Lembre que para uma matriz quadrada A, temos que:

$$
\begin{aligned}
A &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\
A^{-1} &= \frac{1}{|A|}\text{Adj}(A) \\
\text{Adj}(A) &= \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \\
A^{-1} &= \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}
\end{aligned} \tag{3.2.24}
$$

Temos que:

$$
ad - bc = (f(x) + o_p(1))\left(h^2\mu^2 f(x) + o_p\left(h^2\right)\right) - \left(h^2\mu^2 f'(x) + o_p\left(h^2\right)\right)^2 \tag{3.2.25}
$$

Portanto,

$$
\begin{aligned}
\left(n^{-1}X_x{}'W_xX_x\right)^{-1} &= \frac{1}{ad-bc}\begin{bmatrix} h^2\mu^2 f(x) + o_p\left(h^2\right) & -h^2\mu^2 f'(x) - o_p\left(h^2\right) \\ -h^2\mu^2 f'(x) - o_p\left(h^2\right) & f(x) + o_p(1) \end{bmatrix} \\
&= \begin{bmatrix} f^{-1}(x) + o_p(1) & -\dfrac{f'(x)}{f^2(x)} + o_p(1) \\ -\dfrac{f'(x)}{f^2(x)} + o_p(1) & \dfrac{1}{h^2\mu_2 f(x)} + o_p(1) \end{bmatrix}
\end{aligned} \tag{3.2.26}
$$

e

$$
\begin{aligned}
e_1'\left(n^{-1}X_x{}'W_xX_x\right)^{-1}\left(n^{-1}X_x'W_x\right)\begin{bmatrix} (X_1-x)^2 \\ \vdots \\ (X_n-x)^2 \end{bmatrix} &= \left[f^{-1}(x) + o_p(1) \quad -\frac{f'(x)}{f^2(x)} + o_p(1)\right] \times \\
&\quad \times \begin{bmatrix} (X_1-x)^2 \\ \vdots \\ (X_n-x)^2 \end{bmatrix}
\end{aligned} \tag{3.2.27}
$$

Disso decorre que

$$\mathbb{E}\left[\widehat{m}_{LL}(x)|X\right] - m(x) = \frac{h^2\mu^2 m^{''}(x)}{2} + o_p\left(h^2\right) \tag{3.2.28}$$

Assim, o estimador linear local é livre de viés de projeto, é livre de f. Nos pontos de fronteira, o estimador *kernel* de Nadaraya-Watson tem um viés assintótico de ordem h_n, enquanto o estimador linear local tem viés de ordem h_n^2. Nesse sentido, a estimativa linear local elimina o viés nos limites.

3.3 Seleção da Bandwidth

A expressão da variância é dada por:

$$\mathrm{V}\left[\widehat{m}_{LL}(x)|X\right] = e_1'\left(n^{-1}X_x{}'W_xX_x\right)^{-1} n^{-1}X_x{}'W_xVW_xX_x\left(n^{-1}X_x{}'W_xX_x\right)e_1 \tag{3.3.1}$$

em que $\mathrm{V} = \begin{bmatrix} \sigma^2(X_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2(X_n) \end{bmatrix}$.

Semelhante ao resultado anterior,

$$\begin{aligned} n^{-1}X_x{}'W_xVW_xX_x =& n^{-1}\sum_{i=1}^{n}\mathcal{K}_h^2(X_i-x)\sigma^2(X_i)\begin{bmatrix} 1 & (X_i-x) \\ (X_i-x) & (X_i-x)^2 \end{bmatrix} \\ =& \begin{bmatrix} \dfrac{f(x)}{h}\displaystyle\int_{-\infty}^{\infty}\mathcal{K}(s)ds + o_p(h^{-1}) & O_p(h) \\ O_p(h) & O_p\left(h^2\right) \end{bmatrix} \end{aligned} \tag{3.3.2}$$

Portanto,

$$\mathrm{V}\left[\widehat{m}_{LL}(x)|X\right] = \frac{1}{nh}\frac{1}{f(x)}\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds\sigma^2(x) + o_p\left(\frac{1}{nh}\right) \tag{3.3.3}$$

Combinando esses resultados, temos:

$$\text{AMISE} = \frac{h^4\mu^4}{4}\int_{-\infty}^{\infty}\left(m^{''}(x)\right)^2 f(x)\,dx + \frac{1}{nh}\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds\int_{-\infty}^{\infty}\sigma^2(x)\,dx \tag{3.3.4}$$

e assim, otimizando essa expressão e derivando com relação a h, obtemos:

$$h^3\mu^4\int_{-\infty}^{\infty}\left(m^{''}(x)\right)^2 f(x)\,dx - \frac{1}{nh^2}\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds\int_{-\infty}^{\infty}\sigma^2(x)\,dx = 0$$

$$h = \left(\frac{\displaystyle\int_{-\infty}^{\infty}\mathcal{K}^2(s)\,ds\int_{-\infty}^{\infty}\sigma^2(x)\,dx}{\mu^4\displaystyle\int_{-\infty}^{\infty}\left(m^{''}(x)\right)^2 f(x)\,dx}\right)^{1/5} n^{-1/5} \tag{3.3.5}$$

em que

$$\int_{-\infty}^{\infty}\sigma^2(x)\,dx = \widehat{\sigma}^2(\max(x) - \min(x)) \tag{3.3.6}$$

A seleção da *bandwidth*, assim como a estimativa da densidade do *kernel*, é de importância fundamental para a estimativa da regressão não-paramétrica. Vários seletores de *bandwidth* foram propostos para a regressão não-paramétrica seguindo ideias de *plug-in* e validação cruzada. Como no caso anterior,

$$CV(h) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_{-1}(X_i; p, h)\right)^2 \tag{3.3.7}$$

e escolher

$$\hat{h}_{CV} := \arg\min_{h>0} CV(h) \tag{3.3.8}$$

3.4 Intervalo de Confiança

Defina $\ell(x)' = e_1' \left(X_x' W_x X_x\right)^{-1} X_x' W_x$. Vamos construir um intervalo de confiança para $r(x)$, em que $r(x)$ assume o papel de m no estimador Nadaraya-Watson e no linear local (vale para qualquer suavizador linear como até um spline). Tipicamente, essas bandas podem ser escritas como

$$\hat{r}_n(x) \pm c \, \text{se}(x) \tag{3.4.1}$$

em que se(x) é uma estimativa do desvio padrão de $\hat{r}_n(x)$ e $c > 0$ é uma constante. Bandas de confiança como (3.4.1) não são realmente bandas de confiança para $r(x)$, em vez disso são bandas de confiança para $\overline{r}_n(x) = \mathbb{E}\left(\hat{r}_n(x)\right)$.

Sejam $\overline{r}_n(x)$ e $s_n(x)$ a média e o desvio padrão de $\hat{r}_n(x)$, respectivamente. Então,

$$\begin{aligned} \frac{\hat{r}_n(x) - r(x)}{s_n(x)} &= \frac{\hat{r}_n(x) - \overline{r}_n(x)}{s_n(x)} + \frac{\overline{r}_n(x) - r(x)}{s_n(x)} \\ &= Z_n(x) + \frac{\text{Viés}\left(\hat{r}_n(x)\right)}{\sqrt{\text{V}\left(\hat{r}_n(x)\right)}} \end{aligned} \tag{3.4.2}$$

O primeiro termo, $Z_n(x)$, converge para uma normal padrão segundo o Teorema Central do Limite de Lindeberg. O segundo termo é o viés dividido pelo desvio-padrão. Na inferência paramétrica, o viés é geralmente menor do que o desvio padrão do estimador, portanto, esse termo vai para zero à medida que o tamanho da amostra aumenta. Na inferência não paramétrica, vimos que a suavização ideal corresponde a equilibrar o viés e o desvio padrão. O segundo termo não desaparece mesmo com amostras grandes. A presença desse segundo termo, que não se anula, introduz um viés no limite Normal. O resultado é que o intervalo de confiança não será centrado em torno da verdadeira função r devido ao viés de suavização $\overline{r}_n(x) - r(x)$.

Existem várias coisas que podemos fazer a respeito desse problema. A primeira é: conviva com isso. Em outras palavras, apenas aceite o fato de que a banda de confiança é para $\overline{r}_n$ e não r. Não há nada de errado com isso, desde que sejamos cuidadosos ao relatar os resultados para deixar claro que as inferências

são para $\overline{r}_n$ e não r. Uma segunda abordagem é estimar a função viés $\overline{r}_n(x) - r(x)$. Isso é difícil de fazer. Na verdade, o termo principal da tendência é $r''(x)$ e estimar a segunda derivada de r é muito mais difícil do que estimar r. Isso requer a introdução de condições de suavidade extra que, então, colocam em questão o estimador original que não usou essa suavidade extra. Isso tem uma certa circularidade desagradável. Uma terceira abordagem é suavizar. Se suavizarmos menos do que a quantidade ideal, o viés diminuirá assintoticamente em relação à variância. Infelizmente, não parece haver uma regra simples e prática para escolher a quantidade certa de suavização. Adotaremos a primeira abordagem e nos contentaremos em encontrar uma banda de confiança para $\overline{r}_n$.

Assuma que $\hat{r}_n(x)$ é um suavizador linear, então $\hat{r}_n(x) = \sum_{i=1}^{n} Y_i \ell_i(x)$. Então,

$$\overline{r}(x) = \mathbb{E}\left(\hat{r}_n(x)\right) = \sum_{i=1}^{n} \ell_i(x) r(x_i) \tag{3.4.3}$$

Vamos assumir homocedasticidade. Então, $\sigma^2(x) = \sigma^2 = \mathrm{V}\,\varepsilon_i$ é constante. Desse modo,

$$\mathrm{V}\left[\hat{r}_n(x)\right] = \sigma^2 \|\ell(x)\|^2 \tag{3.4.4}$$

Vamos considerar bandas de confiança para $\overline{r}_n(x)$ da forma

$$C(x) = (\hat{r}_n(x) - c\hat{\sigma}\|\ell(x)\|, \hat{r}_n(x) + c\hat{\sigma}\|\ell(x)\|) \tag{3.4.5}$$

para algum $c > 0$ e $a \leq x \leq b$.

Sendo σ conhecido, então,

$$\begin{aligned}
\mathbb{P}\left[\overline{r}(x) \in C(x) \text{ para algum } x \in [a,b]\right] &= \mathbb{P}\left[\max_{x\in[a,b]} \frac{|\hat{r}_n(x) - \overline{r}_n(x)|}{\sigma\|\ell(x)\|} > c\right] \\
&= \mathbb{P}\left[\max_{x\in[a,b]} \frac{\left|\sum_i \varepsilon_i \ell_i(x)\right|}{\sigma\|\ell(x)\|} > c\right] \\
&\left[\right.
\end{aligned}$$

$$= \mathbb{P}\left[\max_{x\in[a,b]} |W(x)| > c\right] \tag{3.4.6}$$

em que $W(x) = \sum_{i=1}^{n} Z_i T_i(x)$, $Z_i = \frac{\varepsilon_i}{\sigma} \sim \mathcal{N}(0,1)$ e $T_i(x) = \frac{\ell_i(x)}{\|\ell(x)\|}$. Desse modo, $W(x)$ é um processo Gaussiano[4]. Para encontrar c é necessário computar a distribuição do máximo de um processo Gaussiano. É bem conhecido que

$$\mathbb{P}\left[\max_{x}\left|\sum_{i=1}^{n} Z_i T_i(x)\right| > c\right] \approx 2\left(1-\Phi(c)\right) + \frac{\kappa_0}{\pi}\mathrm{e}^{-c^2/2} \tag{3.4.7}$$

para c grande, em que

$$\kappa_0 = \int_a^b \|T'(x)\|\, dx \tag{3.4.8}$$

com $T'(x) = (T_1'(x), \ldots, T_n'(x))$ e $T_i'(x) = \frac{\partial T_i(x)}{\partial x}$ é o tamanho da curva $\left\{\frac{\ell(x)}{\|\ell(x)\|}; a \le x \le b\right\}$. Se escolhermos c para resolver

$$2\left(1-\Phi(c)\right) + \frac{\kappa_0}{\pi}\mathrm{e}^{-c^2/2} = \alpha \tag{3.4.9}$$

então obtemos as bandas de confiança simultâneas desejadas. Se σ é desconhecido, nós usamos uma estimativa de $\hat{\sigma}$.

Sun and Loader sugerem substituir o lado direito da fórmula tubo com

$$\mathbb{P}[|T_m| > c] + \frac{\kappa_0}{\pi}\left(1+\frac{c^2}{m}\right)^{-m/2} \tag{3.4.10}$$

em que T_m têm distribuição t−Student com $m = n - tr(L)$ graus de liberdade. Para n grande, continua a ser uma aproximação adequada.

Agora suponhamos que $\sigma(x)$ seja uma função de x. Então,

[4] Isto significa que é uma função aleatória cujo vetor $(W(x_1), \ldots, W(x_k))$ tem uma distribuição multivariada normal para um conjunto finito de pontos $x_1, \ldots, x_k$.

$$\mathrm{V}\left(\widehat{r}_n(x)\right) = \sum_{i=1}^{n} \sigma^2(x_i)\ell_i^2(x)\cdot \tag{3.4.11}$$

Neste caso escolhemos que

$$I(x) = \widehat{r}_n(x) \pm cs(x) \tag{3.4.12}$$

sendo que

$$s(x) = \sqrt{\sum_{i=1}^{n} \widehat{\sigma}^2(x_i)\ell_i^2(x)} \tag{3.4.13}$$

em que $\hat{\sigma}(x)$ é um estimador de $\sigma(x)$ e c é a constante definida acima. Caso $\hat{\sigma}(x)$ varie lentamente com x, então $\sigma(x_i) \approx \sigma(x)$ para aqueles i tais que $\ell_i(x)$ é grande e então

$$s(x) \approx \widehat{\sigma}(x)\|\ell(x)\| \tag{3.4.14}$$

Assim, uma banda de confiança aproximada é

$$I(x) = \widehat{r}_n(x) \pm c\widehat{\sigma}(x)\|\ell(x)\| \tag{3.4.15}$$

3.5 Script R

```
## Kernels ##

kr = function(x) 0.5*(abs(x)<=1)
ke = function(x) 0.75*(1-x^2)*(abs(x)<=1)
kb = function(x) 0.9375*(1-x^2)^2*(abs(x)<=1)

## Estimador Nadaraya-Watson ##

m.nw = function(x,y,h1,k){
   n = length(x)
   result = rep(0,n)
   for(i in 1:n){
      result[i] = sum(k((x-x[i])/h1)*y)/sum(k((x-x[i])/h1))
   }
   result
}

## Estimador Local-Linear ##

m.ll = function(x,y,h1,k){
   n = length(x)
   result = rep(0,n)
   for(i in 1:n){
      x0 = x[i]
      X.x = cbind(rep(1,n),(x-x0))
      W.x = diag(k((x-x0)/h1))
      beta.h = solve(t(X.x)%*%W.x%*%X.x)%*%(t(X.x)%*%W.x%*%y)
      result[i] = beta.h[1]
   }
   result
}

m.nw.x0 = function(x0,x,y,h1,k){
   result = sum(k((x-x0)/h1)*y)/sum(k((x-x0)/h1))
   result
```

```
}

m.ll.x0 = function(x0,x,y,h1,k){
   n = length(x)
   X.x = cbind(rep(1,n),(x-x0))
   W.x = diag(k((x-x0)/h1))
   beta.h = solve(t(X.x)%*%W.x%*%X.x)%*%(t(X.x)%*%W.x%*%y)
   result = beta.h[1]
   result
}

## Cross-Validation Function ##

f.cv = function(h1,x,y,k,estimador=c("LL","NW")){
   n = length(x)
   result = rep(0,n)
   if(estimador=="LL") m1 = m.ll.x0
   if(estimador=="NW") m1 = m.nw.x0
   for(i in 1:n){
      x1 = x[-i]
      y1 = y[-i]
      aux.i = m1(x0=x[i],x1,y1,h1,k)
      result[i] = aux.i
   }
   sum((y - result)^2)
}

## DGP ##

set.seed(1991)
n = 500
x = runif(n,0,1)
m.x = sin(4*x)
e = rnorm(n,0,1/3)
y = m.x + e

dev.off()
plot(x,y,pch=20)
```

```
lines(sort(x),m.x[order(x)],col=1)

## Plot com h fixo ##

h.nw = 0.5
h.ll = 0.5
m.hat.x1 = m.nw(x,y,h1=h.nw,k=ke)
m.hat.x2 = m.ll(x,y,h1=h.ll,k=ke)
lines(sort(x),m.hat.x1[order(x)],col=2)
lines(sort(x),m.hat.x2[order(x)],col=4)
legend("topright", legend=c("Paramétrico", "NW","LL"),col=c(1,2,4),
    lty=1, cex=0.8)

## Bandwidth Selection LL Estimator ##

h.grid = seq(from=0.1,to=0.5,by=0.005)
nh = length(h.grid)
obj.cv = rep(0,nh)
for(i in 1:n){
   obj.cv[i] = f.cv(h1=h.grid[i],x,y,k=ke,estimador="LL")
}
pos.h.ot = which.min(obj.cv)
h.cv.ll = h.grid[pos.h.ot]
m.h.ll = m.ll(x,y,h1=h.cv.ll,k=ke)

## Bandwidth Selection NW Estimator ##

h.grid = seq(from=0.1,to=0.5,by=0.005)
nh = length(h.grid)
obj.cv = rep(0,nh)
for(i in 1:n){
   obj.cv[i] = f.cv(h1=h.grid[i],x,y,k=ke,estimador="NW")
}
pos.h.ot = which.min(obj.cv)
h.cv.nw = h.grid[pos.h.ot]
m.h.nw = m.nw(x,y,h1=h.cv.nw,k=ke)

## Plot Fucntions ##
```

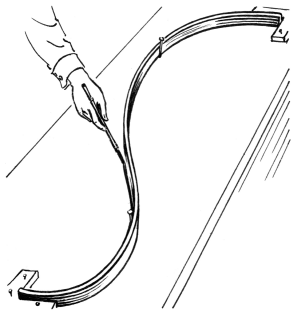
```
dev.off()
plot(x,y,pch=20)
lines(sort(x),m.x[order(x)],col=1)
lines(sort(x),m.h.ll[order(x)],col=2)
lines(sort(x),m.h.nw[order(x)],col=4)
legend("topright", legend=c("Paramétrico","LL", "NW"),col=c(1,2,4),
    lty=1, cex=0.8)

## Banda de Confiança ##

f.w = function(x,h,k){
   n = length(x)
   f1 = rep(0,n)
   for(i in 1:n){
      f1[i] = sum(k((x-x[i])/h))/(n*h)
   }
   f1
}

iqr = diff(quantile(x, c(0.25,0.75)))/diff(qnorm(c(0.25, 0.75)))
h.pg = 1.06*n^(-1/5)*min(sd(x), iqr)
fe = f.w(x,h=h.pg,k=ke)

f.erro = function(x,h,k,estimador=c("LL","NW")){
   n = length(x)
   f1 = rep(0,n)
   if(estimador=="LL") erro = (m.h.ll-y)^2
   if(estimador=="NW") erro = (m.h.nw-y)^2

   if(estimador=="LL") h = h.cv.ll
   if(estimador=="NW") h = h.cv.nw

   for(i in 1:n){
      f1[i] = sum(k((x-x[i])/h)*erro)/(sum(k((x-x[i])/h)))
   }
   f1
}
```

```
erro.ll = f.erro(x,h=h.pg,k=ke,estimador="LL")
erro.nw = f.erro(x,h=h.pg,k=ke,estimador="NW")

var_ll <- erro.ll*Rk/(fe*n*h.pg)
var_nw <- erro.nw*Rk/(fe*n*h.pg)

lim.inf.ll <- m.h.ll - 2*sqrt(var_ll)
lim.sup.ll <- m.h.ll + 2*sqrt(var_ll)

lim.inf.nw <- m.h.nw - 2*sqrt(var_nw)
lim.sup.nw <- m.h.nw + 2*sqrt(var_nw)

dev.off()
plot(x,y,pch=20)
lines(sort(x),m.h.ll[order(x)],col=1,ylim =
    c(min(lim.inf.ll),max(lim.sup.ll)))
lines(sort(x),type="l",lim.inf.ll[order(x)],col=2)
lines(sort(x),type="l",lim.sup.ll[order(x)],col=4)
legend("topright", legend=c("LL", "CLT CI Upper LL","CLT CI Lower
    LL"),col=c(1,4,2), lty=1, cex=0.8)

dev.off()
plot(x,y,pch=20)
lines(sort(x),m.h.nw[order(x)],col=1,ylim =
    c(min(lim.inf.nw),max(lim.sup.nw)))
lines(sort(x),type="l",lim.inf.nw[order(x)],col=2)
lines(sort(x),type="l",lim.sup.nw[order(x)],col=4)
legend("topright", legend=c("NW", "CLT CI Upper NW","CLT CI Lower
    NW"),col=c(1,4,2), lty=1, cex=0.8)
```

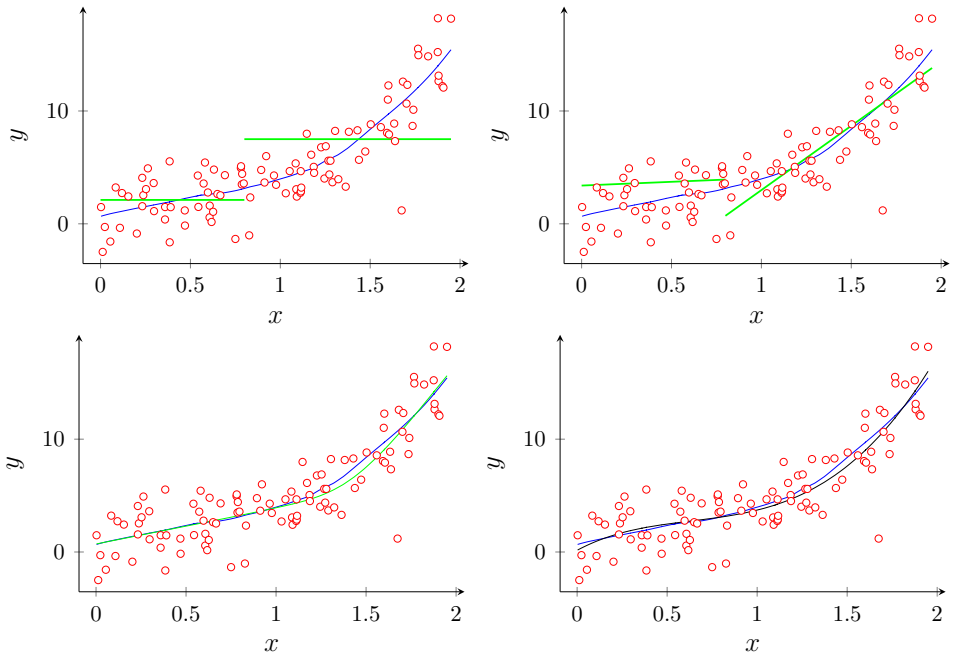
4. Regressão Semiparamétrica

A regressão semiparamétrica pode ser de valor substancial na solução de problemas científicos complexos. O mundo real é muito complicado para a mente humana compreender em grandes detalhes. Modelos de regressão semiparamétricos reduzem conjuntos de dados complexos a resumos que podemos entender. Aplicados corretamente, eles retêm características essenciais dos dados enquanto descartam detalhes sem importância e, portanto, auxiliam na tomada de decisão sólida.

Podemos trabalhar com um *spline*. O nome *spline* vem de uma ferramenta simples usada por artesãos para desenhar curvas suaves, que era uma tira fina de um material flexível como uma madeira macia (ver Figura abaixo). Dobrar esse material consome energia – quanto mais rígido o material, mais energia precisa ser dispendida e, assim, mais reta a curva será entre os pontos. Para suavizar splines, usar um material mais rígido corresponde a aumentar λ.



Abaixo temos exemplos de regressões por splines.



Nota: Geramos 100 pares (x_i, y_i) aleatoriamente como $Y = 3\sin(X) + \varepsilon$ para $x < 1$ e $Y = 2X^3 + \varepsilon$ para $x \geq 1$, em que $X \sim U[0, 2]$ e $\varepsilon \sim N(1, 2)$. A curva azul é obtida por meio do estimador linear local (com $h = 0.05$). O plot da esquerda (acima) é o caso de um *piecewise constant*; do lado direito (acima), *piecewise linear*. O plot da esquerda (abaixo) é o caso de um *cubic spline smoothing*; do lado direito (abaixo), a curva preta é um spline penalizado.

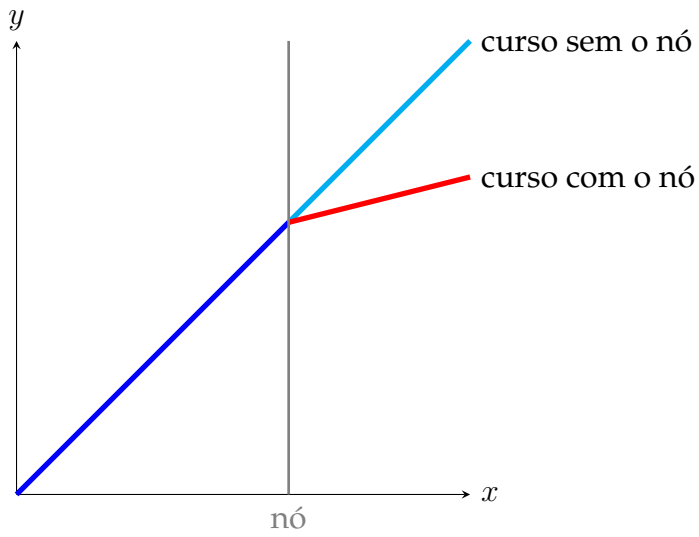
4.1 Splines

4.1.1 Spline Linear

Vamos começar com a seguinte situação:

$$y = \begin{cases} \beta_1 x, \text{ se } x < \text{knot} \\ \beta_1 x + \beta_2 (x - \text{knot}) \text{ se } x \geq \text{knot} \end{cases} \tag{4.1.1}$$

Figura 4.1.1 – Ilustração de um Spline Linear com Um Nó



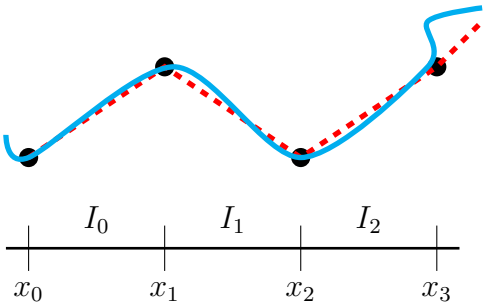
Presumimos até agora que X é unidimensional. Uma função polinomial por partes $f(X)$ é obtida dividindo o domínio de X em intervalos contíguos e representando f por um polinômio separado em cada intervalo. Por exemplo, com três funções básicas:

$$h_1(X) = I(X < \xi_1) \tag{4.1.2}$$
$$h_2(X) = I(\xi_1 \leq X < \xi_2) \tag{4.1.3}$$
$$h_3(X) = I(X \geq \xi_2) \tag{4.1.4}$$

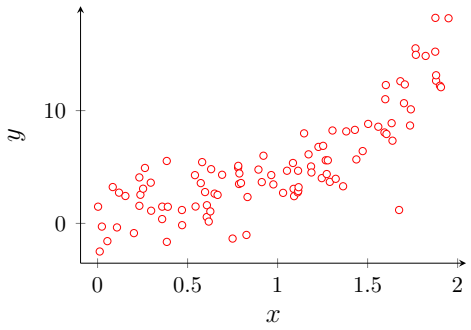
Uma vez que são regiões positivas disjuntas, a estimativa de mínimos quadrados ordinários do modelo $f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$ equivale a $\hat{\beta}_m = \overline{Y}_m$, a média de Y na m-ésima região.

O objetivo é interpolar os dados, como na Figura 4.1.2.

Figura 4.1.2 – Interpolação de Dados



Suponha que tenhamos o seguinte conjunto de dados, cujo comportamento é distinto para $x \leq 1$ e $x > 1$.



Assim, escrevemos o seguinte modelo de regressão

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11} \underbrace{(x_i - 1)_+}_{u_t} + \varepsilon_i \tag{4.1.5}$$

do qual distinguimos dois casos:

1. se $x_i \leq 1 \implies y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. se $x_i > 1 \implies y_i = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 1) + \varepsilon_i = \beta_0 - \beta_{11} + (\beta_1 + \beta_{11})x_i + \varepsilon_i$

Aqui,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & (x_1 - 1) \\ 1 & x_2 & (x_2 - 1) \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - 1) \end{bmatrix} \tag{4.1.6}$$

cujas colunas formam uma base.

O ajuste do modelo pode ser visto por $\widehat{\boldsymbol{y}} = \boldsymbol{S}_\lambda \boldsymbol{y}$, em que $\boldsymbol{S}_\lambda = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X} + \lambda^{2p}\boldsymbol{D}\right)^{-1}\boldsymbol{X}'$ (equivalente à matriz de projeção). Portanto,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda^{2p}\boldsymbol{D}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{4.1.7}$$

Com isso, temos a seguinte regra de seleção do parâmetro de seleção:

$$CV(\lambda) = \sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - S_{\lambda,ii}}\right)^2 \tag{4.1.8}$$

e

$$GCV(\lambda) = \sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - n^{-1}\text{tr}\left(S_\lambda\right)}\right)^2 \tag{4.1.9}$$

Temos diferentes tipos de splines e para isso temos diferentes parâmeros de penalização, variando entre dois casos degenerados:

1. se $\lambda = 0$, temos o estimador de mínimos quadrados ordinários
2. Se $\lambda = \infty$, temos o modelo polinomial (sem *knots*)

Sabemos que o comportamento dos polinômios ajustados aos dados tende a ser errático próximo aos limites, e a extrapolação pode ser perigosa. Esses problemas são exacerbados com splines. Os polinômios ajustados além dos nós de limite se comportam de forma ainda mais selvagem do que os polinômios globais correspondentes naquela região. Isso pode ser convenientemente resumido em termos da variação pontual das funções spline ajustadas por mínimos quadrados.

Para determinarmos o número de knots, usamos a seguinte regra

$$k = \min\left(\frac{n}{4}, 35\right) \tag{4.1.10}$$

4.1.2 Cubic Smoothing Spline

A função objetivo é

$$\min \left[\sum_{i=1}^{n} \left(y_i - \widehat{f}(x_i) \right)^2 + \lambda^3 \int \left\{ \widehat{f}(x)^2 \right\}^2 dx \right], \quad \lambda > 0 \tag{4.1.11}$$

Uma variação são os splines cúbicos naturais. A linearidade é obtida fazendo-se $f'' = f''' = 0$. Os splines cúbicos naturais são chamados de "naturais" porque surgem como a solução de um problema de otimização. No entanto, as restrições à esquerda do primeiro e à direita do último nó não têm nenhuma interpretação estatística natural da qual estamos cientes.

4.1.3 Spline Generalizado

Retomando o exemplo:

$$y_i = (\beta_0 - \beta_{11} \times 1) + (\beta_1 + \beta_{11}) x_i + \varepsilon_i \tag{4.1.12}$$

Seja:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.1.13}$$

em que

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_1 & (x_1 - 1) \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - 1) \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix} \tag{4.1.14}$$

Portanto,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{4.1.15}$$

Generalizando, com $\kappa_1, \kappa_2, \ldots, \kappa_p$ nós a base se torna:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & (x_1-\kappa_1)_+ & \dots & (x_1-\kappa_k)_+ \\ 1 & x_2 & (x_2-\kappa_1)_+ & \dots & (x_2-\kappa_k)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n-\kappa_1)_+ & \dots & (x_n-\kappa_k)_+ \end{bmatrix} \tag{4.1.16}$$

Dado que o ajuste pode não ser suave podemos usar um polinômio de maior ordem, tal que:

$$y_i = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{p} \beta_{pk}(x-\kappa_k)_+^p + \varepsilon_i \tag{4.1.17}$$

e a base se torna

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \dots & x_1^p & (x_1-\kappa_1)_+ & \dots & (x_1-\kappa_k)_+ & \dots & (x_1-\kappa_1)_+^p & \dots & (x_1-\kappa_k)_+^p \\ 1 & x_2 & \dots & x_2^p & (x_2-\kappa_1)_+ & \dots & (x_2-\kappa_k)_+ & \dots & (x_2-\kappa_1)_+^p & \dots & (x_2-\kappa_k)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^p & (x_n-\kappa_1)_+ & \dots & (x_n-\kappa_k)_+ & \dots & (x_n-\kappa_1)_+^p & \dots & (x_n-\kappa_k)_+^p \end{bmatrix}$$

Portanto,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda^{2p}\boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{4.1.18}$$

em que $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{0}_{p+1}, \boldsymbol{1}_K)$.

4.1.4 Spline Penalizado

Impõe-se uma penalização aos coeficientes como segue para se reduzir a rugosidade do *smoothing* devido ao grande número de nós:

$$\sum_{i=1}^{k} \beta_{2k}^2 \leq c \tag{4.1.19}$$

$$\sum_{i=1}^{k} |\beta_{2k}| \leq c \tag{4.1.20}$$

$$\max \|\beta_{2k}\| \leq c \tag{4.1.21}$$

em que alguns β's deixam de ser significativos.

O problema de otimização se torna:

$$\min \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{sujeito a} \quad \lambda^{2p} \boldsymbol{\beta}' \boldsymbol{D} \boldsymbol{\beta}, \quad \lambda \geq 0 \tag{4.1.22}$$

em que $\boldsymbol{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{(k+p+1)\times(k+p+1)} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times(k+p)} \\ \mathbf{0}_{(k+p)\times 2} & \boldsymbol{I}_{(k+p)\times(k+p)} \end{bmatrix}.$

A solução é:

$$\boldsymbol{\beta} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda^{2p}\boldsymbol{D}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} \tag{4.1.23}$$

4.1.5 Inferência

As perguntas naturais são:

- Qual é o desvio padrão estimado de $\hat{f}(x)$?
- O que é um intervalo de confiança de 95% para $f(x)$?

Esses problemas caem no domínio da inferência estatística para a quantidade desconhecida $f(x)$ e estão simplesmente de acordo com aqueles usados rotineiramente na modelagem paramétrica. No entanto, nesse contexto, há uma série de novas questões de inferência global que surgem, tais como:

- f é linear ou não linear?
- uma mudança em $\hat{f}$ "realmente existe"?
- f é monotonicamente crescente?

As bandas de variabilidade são fáceis de calcular se a suavização é linear no vetor $\boldsymbol{y}$, como é o caso de um parâmetro de suavização fixo no spline penalizado ou nas estimativas polinomiais locais.

Seja x um valor qualquer do preditor. Então a estimativa em $f(x)$ é dada por

$$\hat{f}(x) = \boldsymbol{\ell}_x' \boldsymbol{y} \tag{4.1.24}$$

para algum vetor $\boldsymbol{\ell}_x$ de dimensão $n \times 1$ (equivale à linha da matriz de suavização $\boldsymbol{S}$).

Ignorando, por enquanto, a dependência de $\boldsymbol{\ell}_x$ dos parâmetros de suavização estimados, temos

$$\mathrm{V}\left(\hat{f}\,(x)\right) = \boldsymbol{\ell}_x' \mathrm{Cov}(\boldsymbol{y}) \boldsymbol{\ell}_x \tag{4.1.25}$$

Assumindo homocedasticidade, isto é,

$$\mathrm{Cov}(\boldsymbol{y}) = \sigma_\varepsilon^2 \boldsymbol{I} \tag{4.1.26}$$

então,

$$\widehat{\mathrm{dp}}\left\{\hat{f}(x)\right\} = \hat{\sigma}_\varepsilon \|\boldsymbol{\ell}_x\| \tag{4.1.27}$$

Portanto, o intervalo de confiança corresponde a

$$\hat{f}(x) \pm 2\widehat{\mathrm{dp}}\left\{\hat{f}(x)\right\} \tag{4.1.28}$$

Para valores fixados de $\boldsymbol{\ell}_x$, temos que

$$\hat{f}(x) \sim \mathcal{N}\left(\mathbb{E}\left[\hat{f}(x)\right], \sigma_\varepsilon^2 \|\boldsymbol{\ell}_x\|^2\right) \tag{4.1.29}$$

e, então,

$$\frac{\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]}{\sigma_\varepsilon \|\boldsymbol{\ell}_x\|} \sim \mathcal{N}(0,1) \tag{4.1.30}$$

Podemos fazer uma aproximação, assim como no caso paramétrico, tal que

$$\frac{\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]}{\hat{\sigma}_\varepsilon \|\boldsymbol{\ell}_x\|} \overset{approx}{\sim} t_{[df_{residuos}]} \tag{4.1.31}$$

em que $[z]$ é o inteiro mais próximo de z.

Os resultantes intervalos de confiança são

$$\hat{f}(x) \pm \begin{cases} t\left(1 - \frac{\alpha}{2}; df_{residuos}\right) \hat{\sigma}_\varepsilon \|\boldsymbol{\ell}_x\|, & \text{para } n \text{ pequeno} \\ z\left(1 - \frac{\alpha}{2}\right) \hat{\sigma}_\varepsilon \|\boldsymbol{\ell}_x\|, & \text{para } n \text{ grande} \end{cases} \tag{4.1.32}$$

Observe que esses intervalos cobrem $\mathbb{E}\left[\hat{f}(x)\right]$ com $100(1-\alpha)\%$ de confiança em vez de $f(x)$. A interpretação deles como intervalos de confiança aproximados para $f(x)$ requer o não-viés de $\hat{f}(x)$. Frequentemente, a plausibilidade do não-viés pode ser avaliada da inspeção da curva ajustada ao gráfico de dispersão, mas em situações de alto ruído e configurações mais complexas isso pode ser difícil. A teoria de regressão polinomial local mostra que o viés é inerente à regressão não paramétrica quando a quantidade de suavização é ótima e que tende a ser maior em picos e vales na curva de regressão.

4.2 Mixed Models

Nos modelos lineares estamos tentando cumprir dois objetivos: estimar os valores dos parâmetros do modelo e estimar as variâncias apropriadas. Por exemplo, no modelo de regressão mais simples, $y = \alpha + \beta x + \varepsilon$, estimamos os valores para α e β e também a variância de ε. Nós, também podemos estimar $\varepsilon_i = y_i - (\alpha + \beta x_i)$.

Observe que α e β são parâmetros fixos que estamos tentando estimar (fatores fixos ou efeitos fixos), enquanto os valores ε_i são extraídos de alguma distribuição

de probabilidade (normalmente normal com média 0 e variância σ_ε^2). Os ε_i são efeitos aleatórios.

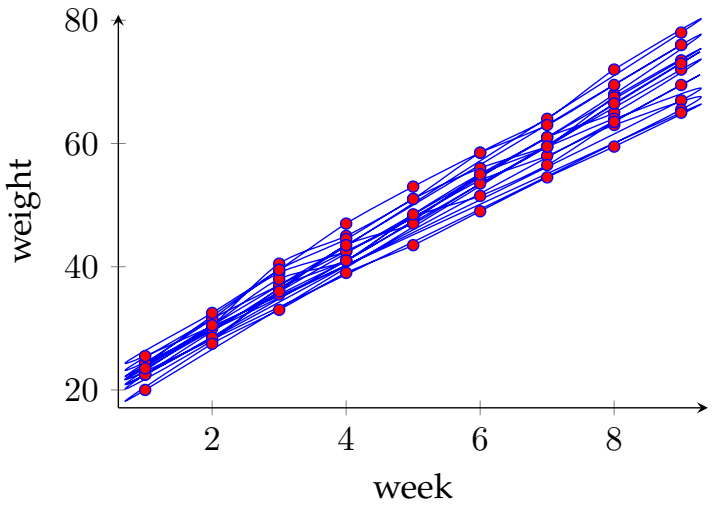
Essa distinção entre efeitos fixos e aleatórios é extremamente importante em termos de como analisamos um modelo. Se um parâmetro é uma constante fixa que desejamos estimar, é um efeito fixo. Se um parâmetro é extraído de alguma distribuição de probabilidade e estamos tentando fazer inferências sobre a distribuição e/ou realizações específicas dessa distribuição, é um efeito aleatório.

Considere o seguinte exemplo de crescimento de porcos:

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{week}_j + \varepsilon_{ij}, \quad j = 1, \ldots, 9 \text{ e } i = 1, \ldots, 48, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{4.2.1}$$

em que i indexa o porco e j indexa a semana.

Abaixo temos uma visualização gráfica do problema (para os 10 primeiros porcos) com os dados de Ruppert, Wand e Carroll (2003):



Fonte.— Consideramos o conjunto de dados longitudinal usado por Ruppert, Wand e Carroll (2003) e Diggle et al. (2002), consistindo em medidas de peso de 48 porcos em nove semanas sucessivas. O gráfico apresenta as curvas de crescimento para os primeiros 10 porcos.

Parece claro que cada porco experimenta uma tendência linear de crescimento e que as medidas gerais de peso variam de porco para porco. Como não estamos realmente interessados nesses 48 porcos em particular, nós os tratamos como uma amostra aleatória de uma população maior e modelamos a variabilidade entre

os porcos como um efeito aleatório ou, equivalentemente, como um termo de intercepto aleatória ao nível do porco. Assim, desejamos ajustar o seguinte modelo:

$$\text{weight}_{ij} = \alpha_i + \beta_1 \text{week}_j + \varepsilon_{ij} \tag{4.2.2}$$

Aqui, temos um intercepto para cada indivíduo. Mas essa solução apresenta os seguintes problemas:

1. 49 parâmetros a serem estimados: $\alpha_i's$ e β_1.
2. $\widehat{\alpha}_i$ depende muito da amostra.
3. O número de $\alpha's$ aumenta com o número de indivíduos observados.

Podemos propor a seguinte alternativa ao modelo (4.2.2):

$$\text{weight}_{ij} = \beta_0 + u_i + \beta_1 \text{week}_j + \varepsilon_{ij} \tag{4.2.3}$$

em que

$u \sim \mathcal{N}(0, \sigma_u^2)$ com σ_u^2 sendo o componente de variância (*variance component*)

$\beta_0 + \beta_1 \text{week}_j$ é o componente fixo (*fixed component*)

u_i é o componente aleatório (*random component*)

Generalizando, seja o seguinte modelo linear:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \tag{4.2.4}$$

Retornando ao exemplo inicial teríamos:

$$\boldsymbol{y} = \begin{bmatrix} \text{weight}_{1,1} \\ \vdots \\ \text{weight}_{1,9} \\ \vdots \\ \text{weight}_{48,1} \\ \vdots \\ \text{weight}_{48,9} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & \text{week}_1 \\ \vdots & \vdots \\ 1 & \text{week}_9 \\ \vdots & \vdots \\ 1 & \text{week}_1 \\ \vdots & \vdots \\ 1 & \text{week}_9 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1}_{9\times 1} & \mathbf{0}_{9\times 1} & \dots & \mathbf{0}_{9\times 1} \\ \mathbf{0}_{9\times 1} & \mathbf{1}_{9\times 1} & \dots & \mathbf{0}_{9\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{9\times 1} & \mathbf{0}_{9\times 1} & \dots & \mathbf{1}_{9\times 1} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_{48} \end{bmatrix}$$

Reescreva o modelo como segue:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \end{aligned} \tag{4.2.5}$$

em que $\mathbb{E}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$, $\text{Cov}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{bmatrix}$, $\boldsymbol{G} = \sigma_u^2 \mathbb{I}_{48}$ e $\boldsymbol{R} = \sigma_\varepsilon^2 \mathbb{I}_{48\times 9}$.

Note que:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}^*\right] = 0 \tag{4.2.6}$$

e

$$\begin{aligned} \text{Cov}(\boldsymbol{\varepsilon}^*) &= \text{Cov}(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}) \\ &= \mathbb{E}\left[(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon})(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon})'\right] \\ &= \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{Z}' + \boldsymbol{Z}\boldsymbol{u}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{u}'\boldsymbol{Z}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \\ &= \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} \equiv \boldsymbol{V} \end{aligned} \tag{4.2.7}$$

Nesse caso, seguindo Rao (1973)

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{4.2.8}$$

e é algumas vezes referido como mínimos quadrados generalizados (GLS).

Segue também que

$$\text{Cov}(\tilde{\boldsymbol{\beta}}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \tag{4.2.9}$$

de modo que a expressão natural para para o desvio-padrão do i-ésimo $\hat{\beta}_i$ é

$$\widehat{\text{dp}}\left(\hat{\beta}_i\right) = \sqrt{i - \text{ésima entrada de } \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}} \tag{4.2.10}$$

Também podemos obter as estimativas de erro-padrão por meio da matriz de covariância:

$$\text{Cov}\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{u}}\end{bmatrix}\middle|\boldsymbol{u}\right) = \sigma_\varepsilon^2\left(\boldsymbol{C}'\boldsymbol{C} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\boldsymbol{D}\right)^{-1}\boldsymbol{C}'\boldsymbol{C}\left(\boldsymbol{C}'\boldsymbol{C} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\boldsymbol{D}\right)^{-1} \tag{4.2.11}$$

4.2.1 Best Linear Prediction (BLP)

Os modelos mistos contêm efeitos fixos, efeitos aleatórios e parâmetros da matriz de covariância. Os efeitos fixos são β_0 e β_1, os efeitos aleatórios são $u_1, \ldots, u_{48}$ e os parâmetros da matriz de covariância são σ_u^2 e σ_ε^2. Os parâmetros no modelo são $\beta_0, \beta_1, \sigma_u^2, \sigma_\varepsilon^2$, e sua estimativa pode ser alcançada usando abordagens estatísticas comuns, como máxima verossimilhança. No entanto, máxima verossimilhança não se aplica a efeitos aleatórios. Em vez disso, podemos fazer previsões de $u_1, \ldots, u_{48}$. A diferença entre previsão e estimativa é que o alvo é aleatório para o primeiro, mas determinístico (não aleatório) para o último.

Assim, seja o modelo

$$y = v + \varepsilon \tag{4.2.12}$$

Observamos apenas y. Com base nessa observação, qual é uma boa previsão para o valor de v? O melhor preditor (BP) de v é definido como sendo o $\tilde{v}$ para o qual o erro quadrático médio

$$\mathbb{E}\left[(\tilde{v} - v)^2\right] \tag{4.2.13}$$

é minimizado. Em geral, a solução é

$$\tilde{v} \equiv BP(v) = \mathbb{E}\left[v|y\right] \tag{4.2.14}$$

Em geral, se $\boldsymbol{y}$ é um vetor de dados observados e $\boldsymbol{v}$ é um vetor aleatório, então a melhor previsão corresponde à minimização de

$$\mathbb{E}\left[\|\tilde{\boldsymbol{v}} - \boldsymbol{v}\|^2\right] \tag{4.2.15}$$

e a solução é

$$\tilde{\boldsymbol{v}} \equiv BP(\boldsymbol{v}) = \mathbb{E}\left[\boldsymbol{v}|\boldsymbol{y}\right] \tag{4.2.16}$$

O único pressuposto necessário aqui é conhecer a distribuição de $\boldsymbol{v}|\boldsymbol{y}$.

O melhor preditor não é necessariamente uma função linear de $\boldsymbol{y}$. Uma simplificação comum é restringir a família de preditores a ser linear. Isso é,

$$\tilde{\boldsymbol{v}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b} \tag{4.2.17}$$

para alguma matriz $\boldsymbol{A}$ e um vetor $\boldsymbol{b}$.

A solução é chamada melhor preditor linear (BLP) e é dada por

$$\tilde{\boldsymbol{v}} \equiv BLP\left(\boldsymbol{v}\right) = \mathbb{E}\left[\boldsymbol{v}\right] + \boldsymbol{C}\boldsymbol{V}^{-1}\left[\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}]\right] \tag{4.2.18}$$

em que $\boldsymbol{C} = \mathbb{E}\left[\left[\boldsymbol{v} - \mathbb{E}(\boldsymbol{v})\right]\left[\boldsymbol{y} - \mathbb{E}(\boldsymbol{y})\right]\right]$ e $\boldsymbol{V} = \mathrm{Cov}\left(\boldsymbol{y}\right)$.

Se

$$\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{y} \end{bmatrix} \tag{4.2.19}$$

é uma normal multivariada, então BP e BLP coincidem. O único pressuposto necessário aqui é conhecer o primeiro e o segundo momento de $\boldsymbol{u}$ e de $\boldsymbol{y}$.

Então,

$$\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{u} \tag{4.2.20}$$

$$\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \tag{4.2.21}$$

e lembrando que $\mathbb{E}[\boldsymbol{Z}\boldsymbol{u}] = 0$ e $\mathbb{E}[\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}] = 0$, temos que

$$\begin{aligned} \boldsymbol{C} &= \mathbb{E}\left[[\boldsymbol{Z}\boldsymbol{u} - \mathbb{E}[\boldsymbol{Z}\boldsymbol{u}]]\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \mathbb{E}[\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}]'\right]\right] \\ &= \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{u}\left(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}\right)'\right] \\ &= \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{Z}' + \boldsymbol{Z}\boldsymbol{u}\boldsymbol{\varepsilon}'\right] \\ &= \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' \end{aligned} \tag{4.2.22}$$

Então,

$$\begin{aligned} BLP\left(\boldsymbol{Z}\boldsymbol{u}\right) &= \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \\ \boldsymbol{Z}\tilde{u} &= \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \quad \left(\times \boldsymbol{Z}^{-1}\right) \\ \tilde{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \\ \tilde{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}'\left[\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}\right]^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \end{aligned} \tag{4.2.23}$$

Na prática, $\boldsymbol{\beta}$ é substituído por um estimador como $\tilde{\boldsymbol{\beta}}$ em (4.2.8), e os parâmetros em $\boldsymbol{G}$ e $\boldsymbol{V}$ precisariam ser estimados.

No nosso exemplo inicial, tínhamos:

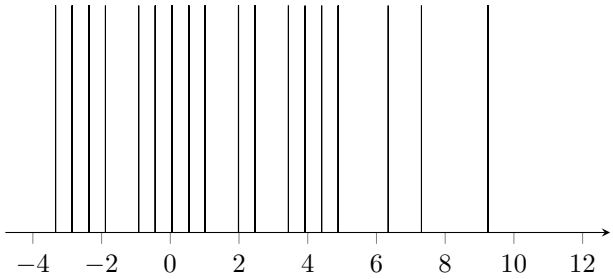
$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_{48} \end{bmatrix} \quad \text{e} \quad \boldsymbol{y} = \begin{bmatrix} \text{weight}_{1,1} \\ \vdots \\ \text{weight}_{1,9} \\ \vdots \\ \text{weight}_{48,1} \\ \vdots \\ \text{weight}_{48,9} \end{bmatrix}$$

Então,

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \tag{4.2.24}$$

é uma normal multivariada e para dados β_0, β_1, σ_u^2 e σ_ε^2, o melhor estimador de u_i se reduz a:

$$\begin{aligned} BLP(\tilde{u}_i) &= \mathbb{E}\left[\tilde{u}\right] + \text{Cov}(\tilde{u}, y)[\text{V}(y)]^{-1}[y - \mathbb{E}[y]] \\ &= 0 + \sigma_u^2 \frac{1}{\sigma_u^2 + \dfrac{\sigma_\varepsilon^2}{n}} \left(\bar{y}_i - \beta_0 - \beta_1 \bar{x}_i\right) \\ &= \frac{n_i \sigma_u^2}{n_i \sigma_u^2 + \sigma_\varepsilon^2} \left(\bar{y}_i - \beta_0 - \beta_1 \bar{x}_i\right) \\ &= \frac{9\sigma_u^2}{9\sigma_u^2 + \sigma_\varepsilon^2} \left(\overline{\text{weight}}_i - \widehat{\beta}_0 - \widehat{\beta}_1 \overline{\text{week}}\right) \end{aligned} \tag{4.2.25}$$



Predições dos Efeitos Aleatórios

4.2.2 Best Linear Unbiaesd Prediction (BLUP)

Uma maneira mais sofisticada de chegar aos resultados das seções anteriores é por meio da noção da melhor previsão linear não-viesada. Aqui, o único pressuposto requerido é conhecer $\boldsymbol{V}$ e $\boldsymbol{C}$. Uma maneira simples (embora um tanto ad hoc) é a justificativa de Henderson, que faz as suposições distributivas.

Assume-se que

$$\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}), \quad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$$

e maximiza a verossimilhança de $(\boldsymbol{y}, \boldsymbol{u})$ ao longo das incógnitas $\boldsymbol{\beta}$ e $\boldsymbol{u}$. Nesse caso, isto mostra que a estimação BLUP de $(\boldsymbol{\beta}, \boldsymbol{u})$ envolve mínimos quadrados generalizados com um termo de penalização, de tal modo que maximizar a verossimilhança de $\boldsymbol{y}|\boldsymbol{u}$ é equivalente a minimizar

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}\boldsymbol{G}^{-1}\boldsymbol{u} \tag{4.2.26}$$

Desse modo,

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{B}\right)^{-1} \boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{y} \tag{4.2.27}$$

em que $\boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix}$ e $\boldsymbol{B} \equiv \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{bmatrix}$.

4.2.3 Estimação da Matriz de Covariância

Estimamos $\boldsymbol{V}$ supondo que $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$. Assim, a função de log-verossimilhança é dada por:

$$\ell(\boldsymbol{\beta}, \boldsymbol{V}) = -\frac{1}{2}\left(n\log(2\pi) + \log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{-1} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \tag{4.2.28}$$

em que $\tilde{\boldsymbol{\beta}}$, para um $\boldsymbol{V}$ qualquer fixo, é dado por

$$\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{V})}{\partial \boldsymbol{\beta}} = 0 \quad \Longleftrightarrow \quad \tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{4.2.29}$$

Substituindo (4.2.29) em (4.2.28), temos a *profile likelihood* para $\boldsymbol{V}$:

$$\ell_p(\boldsymbol{V}) = -\frac{1}{2}\left(\log|\boldsymbol{V}| + \boldsymbol{y}'\boldsymbol{V}^{-1}\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)\boldsymbol{y}\right) - \frac{n}{2}\log(2\pi) \tag{4.2.30}$$

No nosso caso,

$$\boldsymbol{V} = \sigma_u^2 \boldsymbol{Z}'\boldsymbol{Z} + \sigma_\varepsilon^2 \boldsymbol{I} \tag{4.2.31}$$

Então, maximizando ℓ_p vamos obter $\widehat{\sigma}_u$ e $\widehat{\sigma}_\varepsilon$ e, assim, $\hat{\boldsymbol{V}}$, $\hat{\boldsymbol{R}}$ e $\hat{\boldsymbol{G}}$.

Podemos ainda incorporar os graus de liberdade para o modelo de efeitos fixos:

$$\ell_R(\boldsymbol{V}) = \ell_p(\boldsymbol{V}) - \frac{1}{2}\log\left|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right| \tag{4.2.32}$$

No exemplo dos porcos, temos que ao colocar que os *random effects* ocorrem no nível dos porcos, obtemos que $\hat{\beta}_0 = 19,30$ e $\hat{\beta}_1 = 6,21$, com $\sigma_u^2 = 14,81$ e $\sigma_\varepsilon^2 = 4,38$. Quando usamos a verossimilhança restrita, obtemos $\sigma_u^2 = 15,14$ e $\sigma_\varepsilon^2 = 4,39$.

4.2.4 Formulação BLUP para Splines Penalizados

Seja o modelo

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n \tag{4.2.33}$$

em que $f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u\,(x_i - \kappa_k)_+ + \varepsilon_i$ e estamos supondo que os erros satisfazem $\text{Cov}(\varepsilon) = \sigma_\varepsilon^2 \boldsymbol{I}$.

Definindo:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} u\left(x_1 - \kappa_1\right)_+ & \dots & u\left(x_1 - \kappa_k\right)_+ \\ \vdots & \ddots & \vdots \\ u\left(x_n - \kappa_1\right)_+ & \dots & u\left(x_n - \kappa_k\right)_+ \end{bmatrix} \tag{4.2.34}$$

temos que (4.2.26) pode ser reescrita como segue:

$$\frac{1}{\sigma_\varepsilon^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right) + \frac{1}{\sigma_u^2}\boldsymbol{u}'\boldsymbol{u} = \frac{1}{\sigma_\varepsilon^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|^2 + \left(\frac{\lambda^2}{\sigma_\varepsilon^2}\right)\|\boldsymbol{u}\|^2 \tag{4.2.35}$$

com $\text{Cov}(\boldsymbol{u}) = \sigma_u^2\boldsymbol{I}$ em que $\sigma_u^2 = \dfrac{\sigma_\varepsilon^2}{\lambda^2}$.

Assim, podemos escrever um mixed models por meio de splines penalizados como

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \tag{4.2.36}$$

com

$$\text{Cov}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_u^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\varepsilon^2\boldsymbol{I} \end{bmatrix} \tag{4.2.37}$$

No caso de *splines* temos:

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \left(\boldsymbol{C}'\boldsymbol{C} + \lambda^2\boldsymbol{D}\right)^{-1}\boldsymbol{C}'\boldsymbol{y} \tag{4.2.38}$$

em que $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix}$, $\boldsymbol{D} = \text{diag}(0, 0, 1, 1, \dots, 1)$, e λ pode ser estimado por ML ou REML.

4.3 Script R

```
## install.packages("matrixcalc")
## Spline ##

f.splines = function(x,y,la,p1,nk="automatico"){
   n = length(y)
   if(nk=="automatico") nk = min((n/4),35)
   z = seq(from=min(x),to=max(x),length=nk+2)
   z = z[2:(nk+1)]
   Z = rep(1,n)%o%z
   x.aux = matrix(x,nrow=n,ncol=ncol(Z),byrow=F) # desnecessario, mas
       prudente.
   xz = (x.aux-Z>=0)*(x.aux-Z)
   x1 = NULL
   for(i in 1:p1){
      x2 = x^i
      x1=cbind(x1,x2)
   }
   X = cbind(rep(1,n),x1,xz^p1)
   dimnames(X)=NULL
   D1 = diag(c(rep(0,(p1+1)),rep(1,nk)))
   S1 = X%*%solve(t(X)%*%X + la^(2*p1)*D1)%*%t(X)
   y.h = S1%*%y
   list("y.h"=y.h,"la"=la,"n.grid"=nk,"grau.p"=p1,"matrix.suavizacao"=S1)
}

## DGP ##

set.seed(1991)
n = 500
x = runif(n,-1,1)
mx = sin(pi*x)
e = rnorm(n,0,0.7)
y = mx + e

## CV lambda ##
```

```
n.la = 20
la.c = seq(from=0.1,to=3,length=n.la)
erro.la = rep(0,n.la)
for(i in 1:n.la){
   reg.sp.cv = f.splines(x,y,la=la.c[i],p1=2,nk="automatico")
   d.S = diag(reg.sp.cv$mat)
   yh.cv = reg.sp.cv$y.h
   erro.la[i] = sum(((y-yh.cv)/(1-d.S))^2)
}

plot(la.c,erro.la,type="l")
pos.la = which.min(erro.la)
la.ot = la.c[pos.la]

## GCV lambda ##

n.la = 20
la.c = seq(from=0.1,to=3,length=n.la)
erro.la.g = rep(0,n.la)

for(i in 1:n.la){
   reg.sp.cv = f.splines(x,y,la=la.c[i],p1=2,nk="automatico")
   d.S = diag(reg.sp.cv$mat)
   yh.gcv = reg.sp.cv$y.h
   erro.la.g[i] = sum(((y-yh.gcv)/(1-n^(-1)*sum(d.S)))^2)
}

plot(la.c,erro.la.g,type="l")
pos.la.g = which.min(erro.la.g)
la.ot.g = la.c[pos.la.g]

## Local Linear Smoothing ##

m.ll = function(x,y,h1,k1){
   n = length(x)
   result = rep(0,n)
   for(i in 1:n){
```

```
    x0 = x[i]
    x.x0 = x-x0
    s0 = n^(-1)*sum((k1(x.x0/h1)/h1))
    s1 = n^(-1)*sum((k1(x.x0/h1)/h1)*(x.x0))
    s2 = n^(-1)*sum((k1(x.x0/h1)/h1)*(x.x0^2))
    num = n^(-1)*sum((s2-s1*x.x0)*(k1(x.x0/h1)/h1)*y)
    den = s2*s0 - s1^2
    result[i] = num/den
  }
  result
}

m.ll.x0 = function(x0,x,y,h1,k1){
  n = length(x)
  x.x0 = x-x0
  s0 = n^(-1)*sum((k1(x.x0/h1)/h1))
  s1 = n^(-1)*sum((k1(x.x0/h1)/h1)*(x.x0))
  s2 = n^(-1)*sum((k1(x.x0/h1)/h1)*(x.x0^2))
  num = n^(-1)*sum((s2-s1*x.x0)*(k1(x.x0/h1)/h1)*y)
  den = s2*s0 - s1^2
  num/den
}

## CV bandwidth ##

hc = seq(from=0.01,to=0.5,by=0.005)
nh = length(hc)
erro.cv = rep(0,nh)
for(j in 1:nh){
  m.cv = rep(0,n)
  for(i in 1:n){
    x0=x[i]
    x1=x[-i]
    y1=y[-i]
    m.cv[i] = m.ll.x0(x0,x1,y1,hc[j],k1=dnorm)
  }
  erro.cv[j] = sum((y-m.cv)^2)
}
```

```
dev.off()
plot(hc,erro.cv,type="l")
pos.h = which.min(erro.cv)
h.ot = hc[pos.h]
reg.m.ll = m.ll(x,y,h.ot,k1=dnorm)

## Plot comparing LL and Spline (CV) ##

reg.sp = f.splines(x,y,la=la.ot,p1=2,nk="automatico")$y.h
dev.off()
plot(x,y,pch=20)
lines(sort(x),mx[order(x)],col=1)
lines(sort(x),reg.m.ll[order(x)],col=4)
lines(sort(x),reg.sp[order(x)],col=2)
legend("topright", legend=c("mx", "LL","Spline CV"),col=c(1,4,2), lty=1,
    cex=0.8)

## Spline CI ##

reg.sp = f.splines(x,y,la=la.ot,p1=2,nk="automatico")

S.est = reg.sp$mat
L.x = apply(S.est^2,2,sum)
sig = sd(y)
lim.inf = reg.sp$y.h - 2*sig*sqrt(L.x)
lim.sup = reg.sp$y.h + 2*sig*sqrt(L.x)

plot(x,y,type="n")
polygon(c(sort(x),rev(sort(x))), c(lim.inf[order(x)],
    rev(lim.sup[order(x)])), col="gray", border=F)
points(x,y)
lines(sort(x),reg.sp$y.h[order(x)])

## Plot comparing LL and Spline (GCV) ##

reg.sp.g = f.splines(x,y,la=la.ot.g,p1=2,nk="automatico")$y.h
dev.off()
```

```
plot(x,y,pch=20)
lines(sort(x),mx[order(x)],col=1)
lines(sort(x),reg.m.ll[order(x)],col=4)
lines(sort(x),reg.sp.g[order(x)],col=2)
legend("topright",legend=c("mx", "LL","Spline
    GCV"),col=c(1,4,2),lty=1,cex=0.8)

## Plot Spline GCV and Spline (CV) ##

reg.sp = f.splines(x,y,la=la.ot,p1=2,nk="automatico")$y.h
reg.sp.g = f.splines(x,y,la=la.ot.g,p1=2,nk="automatico")$y.h
dev.off()
plot(x,y,pch=20)
lines(sort(x),reg.sp[order(x)],col=2)
lines(sort(x),reg.sp.g[order(x)],col=4)
legend("topright", legend=c("Spline CV","Spline GCV"),col=c(2,4), lty=1,
    cex=0.8)

## Spline GCV CI ##

reg.sp.g = f.splines(x,y,la=la.ot.g,p1=2,nk="automatico")

S.est = reg.sp.g$mat
L.x = apply(S.est^2,2,sum)
sig = sd(y)
lim.inf = reg.sp.g$y.h - 2*sig*sqrt(L.x)
lim.sup = reg.sp.g$y.h + 2*sig*sqrt(L.x)

plot(x,y,type="n")
polygon(c(sort(x),rev(sort(x))), c(lim.inf[order(x)],
    rev(lim.sup[order(x)])), col="gray",border=F)
points(x,y)
lines(sort(x),reg.sp.g$y.h[order(x)])

## Mixed Models ##

f.sp.mx = function(x1,x,y,sig2.e,sig2.u,p1,nk="automatico"){
   n = length(y)
```

```
   p1=1
   if(nk=="automatico") nk = min((n/4),35)
   z = seq(from=min(x),to=max(x),length=nk+2)
   z = z[2:(nk+1)]
   z2 = rep(1,n)%o%z
   x.aux = matrix(x,nrow=n,ncol=ncol(z2),byrow=F)
   Z = (x.aux-z2>=0)*(x.aux-z2)
   X = cbind(rep(1,n),x1,x)
   dimnames(X)=NULL
   C1 = cbind(X,Z)
   D1 = diag(c(rep(0,(ncol(X))),rep(1,nk)))
   la = sig2.e/sig2.u
   b.u = solve(t(C1)%*%C1 + la^(2*p1)*D1)%*%t(C1)%*%y
   y.h = C1%*%b.u
   list("y.h"=y.h,"la"=la,"n.grid"=nk,"grau.p"=p1,"beta.u"=b.u)
}

## Matrix Var-Covar ##

sig.mv = function(x1,x,y,nk="automatico",mv=c("ML","REML")){
   n = length(y)
   p1=1
   if(nk=="automatico") nk = min((n/4),35)
   z = seq(from=min(x),to=max(x),length=nk+2)
   z = z[2:(nk+1)]
   z2 = rep(1,n)%o%z
   x.aux = matrix(x,nrow=n,ncol=ncol(z2),byrow=F)
   Z = (x.aux-z2>=0)*(x.aux-z2)
   X = cbind(rep(1,n),x1,x)
   f.profile = function(a1){
      sig2.e = a1[1]
      sig2.u = a1[2]
      In = diag(1,nrow=n,ncol=n)
      V = sig2.u*Z%*%t(Z) + sig2.e*In
         if(mv=="ML"){
            det.v = determinant(V,logarithm=T)$mod
            M1 = X%*%solve(t(X)%*%solve(V)%*%X)%*%t(X)%*%solve(V)
            lp = -0.5*det.v -0.5*t(y)%*%solve(V)%*%(In-M1)%*%y
```

```
        }
        if(mv=="REML"){
           det.v = determinant(V,logarithm=T)$mod
           det.r = -0.5*determinant(t(X)%*%solve(V)%*%X,logarithm=T)$mod
           M1 = X%*%solve(t(X)%*%solve(V)%*%X)%*%t(X)%*%solve(V)
           lp = -0.5*det.v -0.5*t(y)%*%solve(V)%*%(In-M1)%*%y + det.r
        }
      -lp
   }
   reg1 = optim(par=c(0.01,0.01),fn=f.profile,method="BFGS")
   list("sig2.e"=reg1$par[1],"sig2.u"=reg1$par[2])
}

install.packages("SemiPar")
require(SemiPar)
data(onions)

reg1 = f.sp.mx(x=dens,x1=location,y=log(yield),sig2.e=0.1,sig2.u=0.2)

sig2.ot=sig.mv(x=dens,x1=location,y=log(yield),nk="automatico",mv="REML")
reg2=f.sp.mx(x=dens,x1=location,y=log(yield),sig2.e=sig2.ot$sig2.e,sig2.u=sig2.ot$sig2.u)

sig2.ot=sig.mv(x=dens,x1=location,y=log(yield),nk="automatico",mv="ML")
reg3=f.sp.mx(x=dens,x1=location,y=log(yield),sig2.e=sig2.ot$sig2.e,sig2.u=sig2.ot$sig2.u)
```

5. Modelos Aditivos

5.1 Modelos Multivariados

O capítulo anterior mostrou como construir modelos de regressão flexíveis para um único preditor contínuo modelado como uma função suave. No entanto, muitos problemas de regressão envolvem várias covariáveis contínuas que podem ter relações não lineares com a resposta. Uma vez que a única suposição feita é a de aditividade, eles são referidos como modelos aditivos.

Seja o seguinte modelo

$$m\left(X_{1i},\ldots,X_{Di}\right)=\alpha_0+m_1(X_{1i})+\ldots+m_D(X_{Di})+\varepsilon_i,\quad 1\leq i\leq n, 1\leq j\leq D \tag{5.1.1}$$

Assumimos que

$$\mathbb{E}[Y]=\alpha_0 \tag{5.1.2}$$
$$\mathbb{E}[m_1(X_1)]=\mathbb{E}[m_2(X_2)]=\ldots=\mathbb{E}[m_D(X_D)]=0 \tag{5.1.3}$$
$$\mathbb{E}[\varepsilon_i]=0 \tag{5.1.4}$$
$$\mathbb{E}[\varepsilon_i^2]=\sigma^2 \tag{5.1.5}$$

isto é, um modelo homocedástico padrão.

O pressuposto $\mathbb{E}[m_1(X_1)]=\mathbb{E}[m_2(X_2)]=\ldots=\mathbb{E}[m_D(X_D)]=0$ e a inclusão de uma constante garante a identificação das funções $m_1, m_2, \ldots, m_D$.

Vamos supor dois regressores no modelo. Assim, temos:

$$Y_i = \alpha_0 + m_1(X_i) + m_2(Z_i) + \varepsilon_i \tag{5.1.6}$$

Disso decorre que:

$$\mathbb{E}[Y_i|X_i = x] = \alpha_0 + m_1(X) + \mathbb{E}\left[m_2(Z_i)|X_i = x\right] \tag{5.1.7}$$
$$\mathbb{E}[Y_i|Z_i = z] = \alpha_0 + \mathbb{E}\left[m_1(X_i)|Z_i = z\right] + m_2(Z) \tag{5.1.8}$$

Como consequência,

$$\widehat{m}_1(X) = \mathbb{E}\left[Y_i - \widehat{\alpha}_0 - m_2(Z_i)|X = x\right] = S_1^*(\boldsymbol{y} - \overline{\boldsymbol{y}} - \hat{m}_2) \tag{5.1.9}$$
$$\widehat{m}_2(Z) = \mathbb{E}\left[Y_i - \widehat{\alpha}_0 - m_1(X_i)|Z = z\right] = S_2^*(\boldsymbol{y} - \overline{\boldsymbol{y}} - \hat{m}_1) \tag{5.1.10}$$

A partir disso, procedemos ao algoritmo empregado para estimar o modelo, o *backfitting*:

1: Inicialmente, fazemos $\widehat{\alpha}_0 = \overline{Y}$ e $\widehat{m_2}^{(0)} = \mathbb{E}[Y|Z]$.

2: Estimamos as seguintes quantidades:

2a: $\mathbb{E}\left[Y - \widehat{\alpha}_0 - \widehat{m}_2^{(\ell-1)}|X\right] = \widehat{m}_1^{(\ell)}$

2b: $\mathbb{E}\left[Y - \widehat{\alpha}_0 - \widehat{m}_1^{(\ell-1)}|Z\right] = \widehat{m}_2^{(\ell)}$

2c: $\widehat{m}_1^* = \widehat{m}_1^{(\ell)} - \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \widehat{m}_1^{(\ell)}(X_i)$

2d: $\widehat{m}_2^* = \widehat{m}_2^{(\ell)} - \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \widehat{m}_2^{(\ell)}(Z_i)$

3: Critério de convergência:

3a: CC1: $\dfrac{\displaystyle\sum_{i=1}^{n}\left(\widehat{m}_1^{(\ell)}(X_i) - \widehat{m}_1^{(\ell-1)}(X_i)\right)^2}{\displaystyle\sum_{i=1}^{n}\left(\widehat{m}_1^{(\ell-1)}(X_i)\right)^2} < \varepsilon_1$

3b: CC2: $\dfrac{\sum_{i=1}^{n}\left(\widehat{m}_2^{(\ell)}(Z_i)-\widehat{m}_2^{(\ell-1)}(Z_i)\right)^2}{\sum_{i=1}^{n}\left(\widehat{m}_2^{(\ell-1)}(Z_i)\right)^2}<\varepsilon_2$

Por exemplo, $\varepsilon_1=\varepsilon_2=10\text{e}^{-6}$. Se $CC1<\varepsilon_1$ e $CC2<\varepsilon_2$ finalizar o algoritmo.

Vamos denotar por $\boldsymbol{s}'_{1,x}$ e $\boldsymbol{s}'_{2,z}$, respectivamente, os *kernels* equivalentes para a regressão polinomial local em X_1 e X_2. Assim, estes *kernels* podem ser escrito como

$$\boldsymbol{s}'_{1,x}=\boldsymbol{e}'_1\left(\boldsymbol{X}'_x\boldsymbol{W}_x\boldsymbol{X}_x\right)^{-1}\boldsymbol{X}'_x\boldsymbol{W}_x \tag{5.1.11}$$

$$\boldsymbol{s}'_{2,z}=\boldsymbol{e}'_2\left(\boldsymbol{Z}'_z\boldsymbol{W}_z\boldsymbol{Z}_z\right)^{-1}\boldsymbol{Z}'_z\boldsymbol{W}_z \tag{5.1.12}$$

em que $\boldsymbol{e}'_1=(1,0)$, $\boldsymbol{e}'_2=(1,0)$ e

$$W_x=\text{diag}\left\{\frac{1}{h_1}\mathcal{K}_x\left(\frac{X_1-x}{h_1}\right),\ldots,\frac{1}{h_1}\mathcal{K}_x\left(\frac{X_n-x}{h_1}\right)\right\} \tag{5.1.13}$$

$$W_z=\text{diag}\left\{\frac{1}{h_2}\mathcal{K}_z\left(\frac{Z_1-z}{h_2}\right),\ldots,\frac{1}{h_2}\mathcal{K}_z\left(\frac{Z_n-z}{h_2}\right)\right\} \tag{5.1.14}$$

para alguma função *kernel* $\mathcal{K}_x$ e $\mathcal{K}_z$ e *bandwidths* h_1,h_2 e

$$\boldsymbol{X}_x=\begin{bmatrix}1 & (X_1-x) & \ldots & (X_1-x)^{p_1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & (X_n-x) & \ldots & (X_n-x)^{p_1}\end{bmatrix} \tag{5.1.15}$$

$$\boldsymbol{Z}_z=\begin{bmatrix}1 & (Z_1-z) & \ldots & (Z_1-z)^{p_2}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & (Z_n-z) & \ldots & (Z_n-z)^{p_2}\end{bmatrix} \tag{5.1.16}$$

em que p_1 é a ordem dos polinômios locais para ajustar m_1 e p_2 é a ordem dos polinômios locais para ajustar m_2.

Vamos denotar por $\boldsymbol{S}_1$ e $\boldsymbol{S}_2$ as matrizes de suavização cujas linhas são os *kernels* equivalentes nas observações X e Z, respectivamente:

$$\boldsymbol{S}_1 = \begin{bmatrix} \boldsymbol{s}'_{1,X_1} \\ \vdots \\ \boldsymbol{s}'_{1,X_n} \end{bmatrix}, \quad \boldsymbol{S}_2 = \begin{bmatrix} \boldsymbol{s}'_{2,Z_1} \\ \vdots \\ \boldsymbol{s}'_{2,Z_n} \end{bmatrix} \tag{5.1.17}$$

Definimos o vetor de valores ajustados nos pontos de observação como

$$\hat{\boldsymbol{m}} = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{m}}_1 + \hat{\boldsymbol{m}}_2 \tag{5.1.18}$$

no qual $\hat{\boldsymbol{m}} = \overline{Y}$, e $\hat{\boldsymbol{m}}_1$ e $\hat{\boldsymbol{m}}_2$ são soluções para o conjunto de equações

$$\begin{bmatrix} \mathbb{I} & \boldsymbol{S}_1^* \\ \boldsymbol{S}_2^* & \mathbb{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{m}}_1 \\ \hat{\boldsymbol{m}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1^* \\ \boldsymbol{S}_2^* \end{bmatrix} \boldsymbol{Y} \tag{5.1.19}$$

em que $\boldsymbol{S}_1^* = \left(\mathbb{I} - \dfrac{\mathbf{11}'}{n}\right) \boldsymbol{S}_1$ e $\boldsymbol{S}_2^* = \left(\mathbb{I} - \dfrac{\mathbf{11}'}{n}\right) \boldsymbol{S}_2$.

No caso bivariado, as soluções são:

$$\hat{\boldsymbol{m}}_1 = \left\{\mathbb{I} - \left(\mathbb{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_1^*\right)\right\} \boldsymbol{Y} \equiv \boldsymbol{W}_1 \boldsymbol{Y} \tag{5.1.20}$$

$$\hat{\boldsymbol{m}}_2 = \left\{\mathbb{I} - \left(\mathbb{I} - \boldsymbol{S}_2^* \boldsymbol{S}_1^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_2^*\right)\right\} \boldsymbol{Y} \equiv \boldsymbol{W}_2 \boldsymbol{Y} \tag{5.1.21}$$

$$\hat{\boldsymbol{m}} = \left\{ \frac{\mathbf{11}'}{n} + 2\mathbb{I} - \left(\mathbb{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_1^*\right) - \left(\mathbb{I} - \boldsymbol{S}_2^* \boldsymbol{S}_1^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_2^*\right) \right\} \boldsymbol{Y} \equiv \boldsymbol{W} \boldsymbol{Y} \tag{5.1.22}$$

Tomando os valores esperados, obtemos:

$$\mathbb{E}[\hat{\boldsymbol{m}}_1] = \left\{\mathbb{I} - \left(\mathbb{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_1^*\right)\right\} \left(\boldsymbol{\alpha} + \boldsymbol{m_1} + \boldsymbol{m_2}\right) \tag{5.1.23}$$

$$\mathbb{E}[\hat{\boldsymbol{m}}_2] = \left\{\mathbb{I} - \left(\mathbb{I} - \boldsymbol{S}_2^* \boldsymbol{S}_1^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_2^*\right)\right\} \left(\boldsymbol{\alpha} + \boldsymbol{m_1} + \boldsymbol{m_2}\right) \tag{5.1.24}$$

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{m}}] = &\left\{ \frac{\mathbf{11}'}{n} + 2\mathbb{I} - \left(\mathbb{I} - \boldsymbol{S}_1^* \boldsymbol{S}_2^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_1^*\right) - \left(\mathbb{I} - \boldsymbol{S}_2^* \boldsymbol{S}_1^*\right)^{-1} \left(\mathbb{I} - \boldsymbol{S}_2^*\right) \right\} \times \\ &\times \left(\boldsymbol{\alpha} + \boldsymbol{m_1} + \boldsymbol{m_2}\right) \end{aligned} \tag{5.1.25}$$

Com $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, as variâncias são dadas por

$$\mathrm{V}\left(\hat{m}_1(X_i)\right) = \sigma^2 \boldsymbol{e}_i' \boldsymbol{W}_1 \boldsymbol{W}_1' \boldsymbol{e}_i \tag{5.1.26}$$

$$\mathrm{V}\left(\hat{m}_2(Z_i)\right) = \sigma^2 \boldsymbol{e}_i' \boldsymbol{W}_2 \boldsymbol{W}_2' \boldsymbol{e}_i \tag{5.1.27}$$

$$\mathrm{V}\left(\hat{m}(X_i, Z_i)\right) = \sigma^2 \boldsymbol{e}_i' \boldsymbol{W} \boldsymbol{W}' \boldsymbol{e}_i \tag{5.1.28}$$

```
## Suavização ##

llr=function(x,y,h1,k1=dnorm){
   n=length(x)
   S1=matrix(0,nrow=n,ncol=n)
   for(i in 1:n){
      x0 = x[i]
      x.x0 = x-x0
      s0 = n^(-1)*sum((k1(x.x0/h1)/h1))
      s1 = n^(-1)*sum((k1(x.x0/h1)/h1)*(x.x0))
      s2 = n^(-1)*sum((k1(x.x0/h1)/h1)*(x.x0^2))
      num = n^(-1)*((s2-s1*x.x0)*(k1(x.x0/h1)/h1))
      den = s2*s0 - s1^2
      S1[i,] = num/den
   }
   m.h=S1%*%y
   list("matriz.suavização"=S1,"estimativa"=m.h)
}

## DGP ##

library(MASS)
set.seed(1991)
n = 500
x = runif(n,-1,1)
mx = sin(pi*x)
e = rnorm(n,0,1)
y = mx + e

m1=function(x) 1-6*x + 36*x^2 - 53*x^3 + 22*x^5
m2=function(x) sin(5*pi*x)

z=seq(from=0.001,to=0.99,length=100)
plot(z,m1(z),type="l")
plot(z,m2(z),type="l")
```

```
## Multivariada normal ##

X=matrix(0,nrow=n,ncol=2)
ro=-0.25
V1=matrix(c(1/9,1/9*ro,1/9*ro,1/9),nrow=2,ncol=2)

## Critério de convergência ##

cc=0

while (cc < n){
   g.xy=mvrnorm(n=1,mu=c(1/2,1/2),Sigma=V1)
   if((max(g.xy)<1)&(min(g.xy)>0)){
      X[(cc+1),]=g.xy
      cc=cc+1
   }
}

## Backfitting ##

x1=X[,1]
x2=X[,2]
y=m1(x1)+m2(x2)+e
h1=0.3
h2=0.3
m1.a=0
m2.a=llr(x2,y,h2)$est
a0=mean(y)
cc1=1
cc2=1
ep=0.00001
conta=0

while((cc1>ep)|(cc2>ep)){
   m1.b=llr(x=x1,y=(y-a0-m2.a),h1)$est
   m1.b=m1.b-mean(m1.b)
   cc1=sum(m1.b-m1.a)^2/(sum(m1.a^2)+ep)
```

```
   m1.a=m1.b
   m2.b=llr(x=x2,y=(y-a0-m1.a),h2)$est
   m2.b=m2.b-mean(m2.b)
   cc2=sum(m2.b-m2.a)^2/(sum(m2.a^2)+ep)
   m2.a=m2.b
   conta=conta + 1
}

m1.ve=m1(x1)-mean(m1(x1))
m2.ve=m2(x2)-mean(m2(x2))

plot(sort(x1),m1.ve[order(x1)],type="l")
lines(sort(x1),m1.a[order(x1)],col=4)
plot(sort(x2),m2.ve[order(x2)],type="l")
lines(sort(x2),m2.a[order(x2)],col=4)

## Solução Analítica ##

S1.e1=llr(x1,y,h1)$mat
S2.e2=llr(x2,y,h2)$mat
In=diag(1,nrow=n,ncol=n)
uns=matrix(1,nrow=n,ncol=n)
S1.e=(In-uns/n)%*%S1.e1
S2.e=(In-uns/n)%*%S2.e2
A1=rbind(cbind(In,S1.e),cbind(S2.e,In))
B1=rbind(S1.e,S2.e)
m1.m2=solve(A1)%*%B1%*%y
m1.sl=m1.m2[1:n]
m2.sl=m1.m2[(n+1):(2*n)]

plot(sort(x1),m1.ve[order(x1)],type="l")
lines(sort(x1),m1.sl[order(x1)],col=2)

dev.off()
plot(sort(x2),m2.ve[order(x2)],type="l")
lines(sort(x2),m2.sl[order(x2)],col=2)
```

6. Modelos Generalizados

Os dados de que tratamos nos capítulos anteriores têm a característica de que a variável dependente é contínua. Isso geralmente significa que, possivelmente com a ajuda de uma transformação, os dados podem ser modelados para serem normais e que técnicas de regressão linear (como mínimos quadrados e BLUP) podem ser usadas. No entanto, é comum que a variável dependente não seja contínua, mas categórica ou talvez uma variável de contagem. Os exemplos incluem: um tumor presente ou ausente; o cliente prefere embalagem verde, rosa, laranja ou amarela; o número de admissões de emergência por asma em um determinado dia. Essas variáveis dependente não podem ser tratadas por meio da estrutura de regressão normal. Em muitos campos, as variáveis de resposta categóricas são mais a regra do que a exceção. Alguns dados de variável dependente contínua não podem ser tratados de forma satisfatória na estrutura de erros normais – por exemplo, se eles estiverem muito distorcidos. Os dados distorcidos geralmente podem ser transformados mas uma alternativa é aplicar um modelo Gama aos dados não transformados. Modelos de regressão que visam lidar com variáveis de resposta não gaussianas como essas são geralmente denominados modelos lineares generalizados (GLMs).

6.1 Modelos Lineares Generalizados

Embora o modelo de regressão logística

$$y_i \sim \text{Bernoulli}\left(\frac{\exp\{(\boldsymbol{X}\boldsymbol{\beta})_i\}}{1+\exp\{(\boldsymbol{X}\boldsymbol{\beta})_i\}}\right) \tag{6.1.1}$$

seja o modelo linear generalizado mais comum, existem outros que são

frequentemente usados na prática. Isso inclui o modelo de regressão de Poisson

$$y_i \sim \text{Poisson}[\exp\{(\boldsymbol{X}\boldsymbol{\beta})_i\}] \tag{6.1.2}$$

que é apropriado para dados contáveis, e modelos de regressão do tipo Gamma tais como

$$y_i \sim \text{Gamma}\left[\frac{1}{(\boldsymbol{X}\boldsymbol{\beta})_i}, \phi\right] \tag{6.1.3}$$

e

$$y_i \sim \text{Gamma}\left[\exp\{(\boldsymbol{X}\boldsymbol{\beta})_i\}, \phi\right] \tag{6.1.4}$$

que são apropriados para dados contínuos com cauda pesada à direita.

Um GLM começa com uma família de distribuição exponencial para a variável dependente com densidade da forma

$$f(y;\eta) = \exp\left(\frac{y\eta - b(\eta)}{\phi} + c(y,\phi)\right) \tag{6.1.5}$$

para algumas funções $b(\eta)$ e $c(y,\phi)$. Aqui, ϕ é um parâmetro de dispersão; as distribuições Bernoulli e Poisson não têm parâmetros de dispersão, então para essas distribuições tomamos $\phi \equiv 1$. O parâmetro η é chamado de parâmetro natural. Pode-se mostrar que $E[y] = b'(\eta)$ e $\mathrm{V}(y) = \phi b''(\eta)$, em que $b'(\eta)$ e $b''(\eta)$ são a primeira e a segunda derivadas de b. Em um GLM, assume-se que o parâmetro natural para y_i, η_i, depende de um vetor de variáveis preditoras, $\boldsymbol{x}_i$. Mais explicitamente, assume que para alguma função ψ, $\eta_i = \psi(\boldsymbol{x}_i'\boldsymbol{\beta})$.

A função link[1] é definida pela equação $\mathcal{L}\{\mathbb{E}[y_i]\} = \boldsymbol{x}_i'\boldsymbol{\beta}$. Seja a seguinte notação: $\mu(\cdot) = \mathcal{L}(\cdot)^{-1}$. A inversa da função link converte a predição linear $\boldsymbol{x}_i'\boldsymbol{\beta}$ na expectativa de y_i: $\mu(\boldsymbol{x}_i'\boldsymbol{\beta}) = \mathbb{E}[y_i] = \mu_i$. Para a regressão logística, $b(\eta) = \log(1 + \mathrm{e}^{\eta})$, $\mathcal{L}(\mu) = \text{logit}(u)$ e $\mu(u) = H(u)$, em que H é a função logística.

Assumindo que o parâmetro de dispersão ϕ não depende de i; isto é a

[1] Em particular, a transformação logit – aquela que resulta na probabilidade mais simples – é chamada de função link canônica.

generalização do pressuposto de variância constante do modelo linear. Com isso, a densidade de $\boldsymbol{y}$ é dada por

$$f(\boldsymbol{y};\boldsymbol{\beta}) = \exp\left(\frac{\boldsymbol{y}'\psi(\boldsymbol{X}\boldsymbol{\beta}) - \mathbf{1}'b\{\psi(\boldsymbol{X}\boldsymbol{\beta})\}}{\phi} + \mathbf{1}'c(\boldsymbol{y},\phi)\right) \tag{6.1.6}$$

Na família GLM, se a média $\mathbb{E}[y|\boldsymbol{x}] = \mu(\boldsymbol{x}'\boldsymbol{\beta})$, então a variância é $\mathrm{V}[y|\boldsymbol{x}] = \phi\,\mathrm{V}[\boldsymbol{x}'\boldsymbol{\beta}]$ para alguma função V. Em famílias exponenciais canônicas, pode-se mostrar que a primeira derivada de μ é V, ou seja, $\mu' = V$.

Calcular estimativas de parâmetros em GLMs é particularmente simples e usa um método chamado mínimos quadrados reponderados iterativamente. A ideia básica é a seguinte. Suponha que a estimativa atual seja $\boldsymbol{\beta}^{(t)}$. Constrói-se os pesos $w = \dfrac{1}{\mathrm{V}\left[\boldsymbol{x}'\boldsymbol{\beta}^{(t)}\right]}$. Em seguida, em mínimos quadrados reponderados iterativamente, atualizamos a estimativa atual por MQP. Assim, defina

$$\boldsymbol{W}_{1,\beta} \equiv \mathrm{diag}\{\mu'(\boldsymbol{x}_i'\boldsymbol{\beta})\} \tag{6.1.7}$$

$$\boldsymbol{W}_{2,\beta} \equiv \mathrm{diag}\{V(\boldsymbol{x}_i'\boldsymbol{\beta})\} \tag{6.1.8}$$

Desse modo, o passo $t+1$ é obtido como

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + \left(\boldsymbol{X}'\boldsymbol{W}_{1,\hat{\beta}}\boldsymbol{W}_{2,\hat{\beta}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_{1,\hat{\beta}}\boldsymbol{W}_{2,\hat{\beta}}^{-1}(\boldsymbol{y} - \mu(\boldsymbol{X}\hat{\boldsymbol{\beta}})) \tag{6.1.9}$$

No caso da regressão logística, o algoritmo assume a seguinte forma:

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + \left(\boldsymbol{X}'\,\mathrm{diag}\left\{\frac{\exp[\boldsymbol{X}\hat{\boldsymbol{\beta}}]}{(1+\exp[\boldsymbol{X}\hat{\boldsymbol{\beta}}])^2}\right\}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{y} - \frac{\exp[\boldsymbol{X}\hat{\boldsymbol{\beta}}]}{1+\exp[\boldsymbol{X}\hat{\boldsymbol{\beta}}]}\right) \tag{6.1.10}$$

A estrutura de covariância é dada por

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}'\boldsymbol{W}_{1,\hat{\beta}}\boldsymbol{W}_{2,\hat{\beta}}^{-1}\boldsymbol{W}_{1,\hat{\beta}}\boldsymbol{X}\right)^{-1} \tag{6.1.11}$$

em que

$$\widehat{\mathrm{dp}}(\hat{\beta}_i) = \sqrt{i - \text{ésima entrada diagonal de } \left(\boldsymbol{X}'\boldsymbol{W}_{1,\hat{\beta}}\boldsymbol{W}_{2,\hat{\beta}}^{-1}\boldsymbol{W}_{1,\hat{\beta}}\boldsymbol{X}\right)^{-1}} \tag{6.1.12}$$

Vamos agora entender esses detalhes no caso de um modelo de variável discreta para regressão logística.

Um modelo de regressão logística pode ser escrito como

$$\mathbb{P}[y_i = 1|\boldsymbol{x}_i] = \boldsymbol{\beta}'\boldsymbol{x}_i, \quad i = 1, \ldots, n \tag{6.1.13}$$

A função de log-verossimilhança para esse problema é

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ y_i \left(\boldsymbol{\beta}'\boldsymbol{x}_i\right) - \log(1 + \exp[\boldsymbol{\beta}'\boldsymbol{x}_i]) \right\} \\ &= \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \mathbf{1}' \log\left(\mathbf{1} + \exp[\boldsymbol{X}\boldsymbol{\beta}]\right)\end{aligned} \tag{6.1.14}$$

Diferenciando com relação a $\boldsymbol{\beta}$, temos a função score

$$\boldsymbol{S}(\boldsymbol{\beta}) \equiv \boldsymbol{X}' \left(\boldsymbol{y} - \frac{\exp[\boldsymbol{X}\boldsymbol{\beta}]}{1 + \exp[\boldsymbol{X}\boldsymbol{\beta}]} \right) = 0 \tag{6.1.15}$$

O processo iterativo é da forma

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + \left\{ \boldsymbol{D}\boldsymbol{S}(\boldsymbol{\beta}) \right\}^{-1} \boldsymbol{S}(\boldsymbol{\beta}) \tag{6.1.16}$$

em que $\boldsymbol{D}\boldsymbol{S}(\boldsymbol{\beta})$ é a matriz hessiana: a matriz quadrada com entrada (i, j) igual a

$$\frac{\partial}{\partial \beta_j} \boldsymbol{S}(\boldsymbol{\beta})_i \tag{6.1.17}$$

Aqui,

$$\boldsymbol{D}\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}\boldsymbol{X}' \left(\boldsymbol{y} - \frac{\exp[\boldsymbol{X}\boldsymbol{\beta}]}{1 + \exp[\boldsymbol{X}\boldsymbol{\beta}]} \right)$$

$$
\begin{aligned}
&= -\boldsymbol{X}' \operatorname{diag}\left(\frac{\exp[\boldsymbol{X}\boldsymbol{\beta}]}{(1+\exp[\boldsymbol{X}\boldsymbol{\beta}])^2}\right)\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{W}_{\boldsymbol{\beta}}\boldsymbol{X} \qquad (6.1.18)
\end{aligned}
$$

Perceba que o estimador de máxima verossimilhança de $\boldsymbol{\beta}$ satisfaz

$$
\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = \mathbf{0} \qquad (6.1.19)
$$

Pelo teorema de Taylor, podemos fazer

$$
\begin{aligned}
\mathbf{0} &= \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}\right) \\
&= \boldsymbol{S}\left(\boldsymbol{\beta} + \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\approx \boldsymbol{S}\hat{\boldsymbol{\beta}} + \boldsymbol{D}\boldsymbol{S}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad (6.1.20)
\end{aligned}
$$

Rearranjando termos, obtemos:

$$
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\left\{\boldsymbol{D}\boldsymbol{S}(\boldsymbol{\beta})\right\}^{-1}\boldsymbol{X}'\left(\boldsymbol{y} - \frac{\exp[\boldsymbol{X}\boldsymbol{\beta}]}{1+\exp[\boldsymbol{X}\boldsymbol{\beta}]}\right) \qquad (6.1.21)
$$

Isso implica que

$$
\mathbb{E}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx 0 \qquad (6.1.22)
$$

$$
\operatorname{Cov}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx (\boldsymbol{X}'\boldsymbol{W}_{\boldsymbol{\beta}}\boldsymbol{X})^{-1} \qquad (6.1.23)
$$

o que implica que

$$
\widehat{\mathrm{dp}}(\hat{\beta}_i) = \sqrt{i - \text{ésima entrada diagonal de } \left(\boldsymbol{X}'\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}\boldsymbol{X}\right)^{-1}} \qquad (6.1.24)
$$

6.2 Algoritmo GLM

Pode-se mostrar que $\mu(x)$, para um modelo de escolha discreta, implica, a partir da equação (6.1.6) que:

$$\mathbb{E}(Y|X=x) = b'(\theta(x)) \tag{6.2.1}$$

$$\mathrm{V}(Y|X=x) = a(\phi) + b''(\theta(x)) \tag{6.2.2}$$

$$g(\mu(x)) = x'\beta \Longrightarrow \mu(x) = b'(\theta(x)) \Longrightarrow b^{-1}(\mu(x)) = \theta(x) \tag{6.2.3}$$

$$\binom{\eta_0}{y} P(x)^y (1-P(x))^{1-y} \quad [\text{fdp}] \tag{6.2.4}$$

$$\binom{\eta_0}{y} \exp(y\theta(x) - b(\theta(x))), \quad a=1 \quad \text{e} \quad c=0 \quad [\text{fdp da exponencial}] \tag{6.2.5}$$

$$\theta(x) = \log\left(\frac{P(x)}{1-P(x)}\right) = \log\left(\frac{\mu(x)}{\eta_0 - \mu(x)}\right) \tag{6.2.6}$$

$$b(\theta(x)) = \eta_0 \log[1 + \exp(\theta(x))] \tag{6.2.7}$$

No modelo *logit* (fazendo $\eta_0 = 1$), obtemos então:

$$\begin{aligned} g(t) = \log\left(\frac{t}{\eta_0 - t}\right) \quad &\Longrightarrow \quad x = \log\left(\frac{t}{\eta_0 - t}\right) \\ &\Longrightarrow \quad e^x = \frac{t}{\eta_0 - t} \\ &\Longrightarrow \quad (\eta_0 - t)e^x = t \\ &\Longrightarrow \quad \eta_0 e^x = t + te^x \\ &\Longrightarrow \quad \frac{\eta_0 e^x}{1+e^x} = g^{-1}(x) \\ &\Longrightarrow \quad \frac{e^x}{1+e^x} = g^{-1}(x) \end{aligned} \tag{6.2.8}$$

Assim, procedemos como segue:

1. Escolher β_0:

$$\eta_0 = \left(\eta_1^0, \ldots, \eta_n^0\right)' = (X_1'\beta_0, \ldots, X_n'\beta_0) \tag{6.2.9}$$

$$\mu_0 = \left(\mu_1^0, \ldots, \mu_n^0\right)' = \left(g^{-1}\left(X_1'\beta_0\right), \ldots, g^{-1}\left(X_n'\beta_0\right)\right) \tag{6.2.10}$$

2. Atualização:

   Gerar a variável dependente ajustada:

$$Z_i = \eta_i^0 + \left(Y_i - \mu_i^0\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0 \tag{6.2.11}$$

$$w_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0^2 V_i^0 \tag{6.2.12}$$

   em que $V_i^0 = V(\mu_i^0)$ é a função que caracteriza a variância de $Y|X$.

   Lembre que $\dfrac{\partial \eta_i}{\partial \mu_i} = \dfrac{\partial g(\mu_i)}{\partial \mu_i} = g'(\mu_i)$.

   Assim,

$$g(Y_i) \cong g(\mu_i) + (Y_i - \mu) g'(\mu_i) \tag{6.2.13}$$

3. Por MQP de Z_i sobre Y_i com pesos w_i, obtém-se os novos β_i's.

   Para o modelo *logit*, o algoritmo se torna

$$g^{-1}(x) = \frac{e^x}{1+e^x} \quad \Longrightarrow \quad g(x) = \log\left(\frac{x}{1-x}\right) \tag{6.2.14}$$

$$g'(x) = \frac{1}{\dfrac{x}{1-x}} \frac{1}{1-x^2} = \frac{1}{x(-x)} \tag{6.2.15}$$

$$\mathrm{V}(x) = x(1-x) \tag{6.2.16}$$

$$\mathrm{V}(\mu_i^0) = \mu_i^0(1-\mu_i^0) \tag{6.2.17}$$

$$w^{-1} = \frac{1}{\left(\mu_i^0(1-\mu_i^0)\right)^2}\left(\mu_i^0(1-\mu_i^0)\right) = \frac{1}{\left(\mu_i^0(1-\mu_i^0)\right)} \tag{6.2.18}$$

6.3 Modelos Aditivos Generalizados

Os modelos paramétricos generalizados da seção anterior são não-lineares por causa da função link, mas, ainda assim, são paramétricos e não têm a flexibilidade dos modelos não-paramétricos.

Suponha que observamos pares (x_i, y_i), onde a distribuição condicional de y_i dado x_i é dada pela família exponencial com densidade dada por (6.1.5). Por exemplo, y_i pode ser uma variável binária. Vimos na seção anterior que poderíamos estimar $f(x) = \mathbb{E}[y|x]$ sob um conjunto de suposições paramétricas chamadas de modelo linear generalizado (GLM). Nesta seção, estimamos f assumindo apenas que f é uma função suave. Em outras palavras, a parte "linear" das premissas do GLM será relaxada, mas a estrutura restante do GLM será mantida. Assim, assumimos que y_i dado x_i tem densidade

$$f(y_i; \eta_i) = \exp\left(\frac{y\eta_i - b(\eta_i)}{\phi} + c(y, \phi)\right) \tag{6.3.1}$$

em que η_i depende de x_i. Especificamente, assumimos que $\eta_i = \eta(x_i)$ para uma função suave $\eta(\cdot)$ e usaremos a notação

$$\boldsymbol{\eta} = [\eta(x_1), \ldots, \eta(x_n)]' \tag{6.3.2}$$

Para simplificar, ao longo desta seção, assumiremos uma função link canônica. Suponha também que $\phi \equiv 1$, por exemplo, regressão binária ou Poisson.

O método mais comum para suavizar esses dados é a probabilidade penalizada, que é uma generalização para dados não gaussianos de splines de suavização que minimizam a soma dos quadrados penalizada. Por exemplo, a solução de smoothing spline é

$$\hat{\boldsymbol{f}} = (b')^{-1}(\hat{\boldsymbol{\eta}}) \tag{6.3.3}$$

em que

$$\hat{\boldsymbol{\eta}} = \arg\max_{\eta(\cdot)} \{\boldsymbol{y}'\boldsymbol{\eta} - \mathbf{1}'b(\boldsymbol{\eta})\} - \frac{1}{2}\lambda^3 \int_{-\infty}^{\infty} \eta''(x)^2 dx \tag{6.3.4}$$

Para qualquer espaço de splines, há uma matriz $\boldsymbol{K}$ tal que $\boldsymbol{\eta}'\boldsymbol{K}\boldsymbol{\eta} = \int_{-\infty}^{\infty} \eta''(x)^2 dx$ para todo $\eta(\cdot)$.

A solução do spline linear penalizado, para as funções base

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_i \end{bmatrix}, \quad \boldsymbol{Z} = \Big[(x_i - \kappa_k)_+\Big]_{\substack{1\leq k\leq K \\ 1\leq i\leq n}} \tag{6.3.5}$$

é

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}} \tag{6.3.6}$$

em que

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} & \hat{\boldsymbol{u}} \end{bmatrix} = \arg\max_{\boldsymbol{\beta},\boldsymbol{u}} \{\boldsymbol{y}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}'b(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\} - \frac{1}{2}\lambda^3 \|\boldsymbol{u}\|^2 \tag{6.3.7}$$

6.4 Algoritmo GAM

Procede-se em três passos:

1. Escolher β_0

$$\eta_0 = \left(\eta_1^0, \ldots, \eta_n^0\right)'$$

em que $\eta_i^0 = \alpha^0 + \sum_{d=1}^{D} m_d^0(X_{di})$ e $\mu_0 = (\mu_1^0, \ldots, \mu_n^0)' = \left(g^{-1}\left(\eta_1^0\right), \ldots, g^{-1}\left(\eta_n^0\right)\right)$.

2. Atualizar

3. Fazer o *backfitting* de Z_i sobre X_i com pesos w_i a fim de obterem-se novos $(m_1(\cdot), \ldots, m_j(\cdot))$.

O estimador LL terá $\mathcal{K}_h(X_i - x)w_i$ como ponderador (suavizador).

6.5 Script R

```
## Data ##

install.packages("wooldridge")
library(wooldridge)
data('mroz')
y = mroz$inlf
x1 = mroz$educ
x2 = mroz$exper

## Local Linear Regression ##

llr = function(x,y,h1,k1=dnorm,w=1){
   n = length(x)
   S1 = matrix(0,nrow=n,ncol=n)
   for(i in 1:n){
      x0 = x[i]
      x.x0 = x-x0
      k2 = w*k1(x.x0/h1)/h1
      s0 = n^(-1)*sum((k2))
      s1 = n^(-1)*sum((k2)*(x.x0))
      s2 = n^(-1)*sum((k2)*(x.x0^2))
      num = n^(-1)*((s2-s1*x.x0)*(k2))
      den = s2*s0 - s1^2
      S1[i,] = num/den
   }
   m.h = S1%*%y
   list("matriz.suavizacao"=S1,"estimativa"=m.h)
}

## Nonparametric Regression ##

f.cbe.2x = function(x1,x2,y,h1,h2,k1=dnorm,w=1){
   m1.a = 0
   m2.a = llr(x2,y,h2,k1,w)$est
   m2.a = m2.a - mean(m2.a)
```

```
   a0 = mean(y)
   cc1 = 1
   cc2 = 1
   ep = 0.00001
   conta = 0
   while((cc1>ep)|(cc2>ep)){
      m1.b = llr(x=x1,y=(y-a0-m2.a),h1,k1,w)$est
      m1.b = m1.b-mean(m1.b)
      cc1 = sum((m1.b-m1.a)^2)/(sum(m1.a^2)+ep)
      m1.a = m1.b
      m2.b = llr(x=x2,y=(y-a0-m1.a),h2,k1,w)$est
      m2.b = m2.b - mean(m2.b)
      cc2 = sum((m2.b-m2.a)^2)/(sum(m2.a^2)+ep)
      m2.a = m2.b
      conta = conta+1
   }
   list("m.h" = a0+m1.a+m2.a, "m1.h"=m1.a, "m2.h"=m2.a,
       "n.iteracoes"=conta)
}

## Generalized Linear Model ##

g1 = function(x)  log(x/(1-x))
g1.inv = function(x)  exp(x)/(1+exp(x))
g1.der = function(x) 1/(x*(1-x))
var1 = function(x)   x*(1-x)

n = length(y)
X = cbind(cte=1,"educ"=x1,"exper"=x2,"expersq"=x2^2)
W = matrix(0,nrow=n,ncol=n)
cc = 1
ep = 0.001
conta = 0
beta.0 = solve(t(X)%*%X)%*%t(X)%*%y
while(cc>ep){
   conta=conta+1
   eta.0 = X%*%beta.0
   mi.0 = g1.inv(eta.0)
```

```
   z = eta.0 + (y-mi.0)*g1.der(mi.0)
   w.inv = (g1.der(mi.0)^2)*var1(mi.0)
   diag(W) = 1/w.inv
   beta.1 = solve(t(X)%*%W%*%X)%*%t(X)%*%W%*%z
   cc = sum((beta.1-beta.0)^2)/sum(beta.0^2 + ep)
   beta.0 = beta.1
}

## Generalized Linear Model - funcao pronta ##

reg.logit = glm(y~1+x1+x2+I(x2^2),family="binomial")

## Generalized Additive Model ##

h1 = 2
h2 = 2
cc = 1
ep = 0.001
conta = 0
a0 = g1(mean(y))
m1.0 = 0
m2.0 = 0
while(cc>ep){
   conta=conta+1
   eta.0 = a0+m1.0+m2.0
   mi.0 = g1.inv(eta.0)
   z = eta.0 + (y-mi.0)*g1.der(mi.0)
   w.inv = (g1.der(mi.0)^2)*var1(mi.0)
   reg.np = f.cbe.2x(x1,x2,z,h1,h2,k1=dnorm,w=(1/w.inv))
   m1.1 = reg.np$m1.h
   m2.1 = reg.np$m2.h
   cc = sum((m1.1-m1.0)^2+(m2.1-m2.0)^2)/sum((m1.0)^2+(m2.0)^2 + ep)
   m1.0 = m1.1
   m2.0 = m2.1
}

## Plots ##
```

```
plot(sort(x1),m1.0[order(x1)],type="l")
plot(sort(x2),m2.0[order(x2)],type="l")
```

7. Regressão Quantílica Não-Paramétrica

7.1 Regressão Paramétrica

Podemos denotar a função perda (*loss function*) como:

$$\min_{q\in\mathbb{R}} \mathbb{E}\left[\left(\alpha\mathbb{I}_{[y-q\geq 0]} + (1-\alpha)\mathbb{I}_{[y-q\geq 0]}\right)|y-q|\right] \tag{7.1.1}$$

em que q_α minimiza a função perda.

Podemos pensar o problema em termos da função *check*:

$$\rho_\alpha(u) = \left(\alpha\mathbb{I}_{[u\geq 0]} + (1-\alpha)\mathbb{I}_{[u<0]}\right)|u| \tag{7.1.2}$$

Assim,

1. se $u < 0$, então $|u| = -u$ e, portanto, $\rho_\alpha(u) = (1-\alpha)(-u) = -u + \alpha u$.
2. se $u \geq 0$, então $|u| = u$ e, portanto, $\rho_\alpha(u) = \alpha u$.

Então,

$$\begin{aligned}\rho_\alpha(u) &= \alpha u - u\mathbb{I}_{[u<0]} \\ &= u\left(\alpha - \mathbb{I}_{[u<0]}\right)\end{aligned} \tag{7.1.3}$$

Assim, o estimador para a regressão quantílica é

$$\min_{\beta_0(\alpha),\beta_1(\alpha)} \sum_{i=1}^{n} \rho_\alpha \left(y_i - \beta_0(\alpha) - x_i' \beta_1(\alpha)\right) \tag{7.1.4}$$

7.2 Estimador Linear Local

Estendendo o resultado anterior para o estimador linear local, obtemos:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \rho_\alpha \left(y_i - \beta_0 - \beta_1(x_i - x)\right) \mathcal{K}_h(x_i - x) \tag{7.2.1}$$

Vamos ver o estimador proposto por Yu & Jones (1998), que é um estimador *kernel* para a função de distribuição acumulada. Assim, por definição

$$F_Y(y) = \int_{-\infty}^{y} f_Y(s)\, ds \tag{7.2.2}$$

Fazemos a substituição pela FDA empírica:

$$\begin{aligned}
\widehat{F}_Y(y) &= \int_{-\infty}^{y} \widehat{f}_Y(s)\, ds \\
&= \int_{-\infty}^{y} \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{s - y_i}{h}\right) ds \quad \text{[semelhante à equação (2.2.1)]} \\
&= \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\frac{y - y_i}{h}} \mathcal{K}(\phi) h\, d\phi \\
&= \frac{1}{n} \sum_{i=1}^{n} \Omega\left(\frac{y_i - y}{h}\right)
\end{aligned} \tag{7.2.3}$$

em que se troca a ordem do somatório e da integração pois a função é contínua.

Fez-se a substituição $\phi = \dfrac{s - y_i}{h}$ e trocando-se o limite superior do integrando observado-se o que ocorre quando $h \to 0$.

Lembre que por definição q_α é o α-ésimo percentil da distribuição de y se $P(y \leq q_\alpha) \geq \alpha$ e $P(y \geq q_\alpha) \leq 1 - \alpha$.

E também sabemos que

$$F^{-1}\left[F\left(q_\alpha\right)\right] = F^{-1}(\alpha) = q_\alpha \tag{7.2.4}$$

Assim, se estimarmos $F_{y|x}$ podemos estimar q_α invertendo $F_{y|x}$ avaliada em α. Para tanto, considere:

$$\int_{-\infty}^{y} \frac{1}{h_2}\Omega\left(\frac{y_j - u}{h_2}\right) du = \Omega\left(\frac{y - y_j}{h_2}\right) \tag{7.2.5}$$

É possível mostrar que quando $h_2 \to 0$, obtemos:

$$\begin{aligned}
\mathbb{E}\left[\Omega\left(\frac{y_j - y}{h_2}\right) \middle| X = \alpha\right] &\cong F_{Y|X}(y|x) \\
&= \int_{-\infty}^{\infty} \Omega\left(\frac{y - \alpha}{h_2}\right) f_{Y|X}(\alpha|x)\, d\alpha \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{y} \frac{1}{h_2}\mathcal{K}\left(\frac{\alpha - s}{h_2}\right)\, ds\, f_{Y|X}(\alpha|x)\, d\alpha \\
&= \int_{-\infty}^{y}\int_{-\infty}^{\infty} \frac{1}{h_2}\mathcal{K}\left(\frac{\alpha - s}{h_2}\right) f_{Y|X}(\alpha|x)\, d\alpha\, ds \\
&= \int_{-\infty}^{y}\int_{-\infty}^{\infty} \frac{1}{h_2}\mathcal{K}(\phi) f_{Y|X}(s + \phi h_2|x) h_2\, d\phi\, ds \\
&= \int_{-\infty}^{y}\int_{-\infty}^{\infty} \frac{1}{h_2}\mathcal{K}(\phi)[f_{Y|X}(s|x) + o(1)] h_2\, d\phi\, ds
\end{aligned}$$

$$= \int_{-\infty}^{y} f_{Y|X}(s|x) \underbrace{\int_{-\infty}^{\infty} \mathcal{K}(\phi)\, d\phi}_{=1}\, ds + o(1)$$

$$\approx F_{Y|X}(y|x) \tag{7.2.6}$$

em que $\phi = \dfrac{\alpha - s}{h_2}$, então $\alpha = \phi h_2 + s$ e $d\alpha = h_2\, d\phi$.

Assim, o estimador proposto tem a seguinte função objetivo:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(\Omega\left(\frac{y - y_i}{h_2}\right) - \beta_0 - \beta_1(x_i - x) \right)^2 \mathcal{K}\left(\frac{x_i - x}{h_1}\right) \tag{7.2.7}$$

e o estimador da função de distribuição condicional é dado por

$$\widetilde{F}_{h_1,h_2}(y|x) = \frac{1}{\sum_{j=1}^{n} \Omega_j(x, h_1)} \sum_{j} \Omega_j(x, h_1) \Omega\left(\frac{y - y_j}{h_2}\right) \tag{7.2.8}$$

Assim, o α-ésimo percentil é a solução de:

$$\alpha = \frac{1}{\sum_{j=1}^{n} \Omega_j(x, h_1)} \sum_{j}^{n} \Omega_j(x, h_1) \Omega\left(\frac{\widehat{q}_\alpha(x) - y_j}{h_2}\right) \tag{7.2.9}$$

7.3 Script R

```
## Check Function ##

ro.u = function(u,a) (a - (u < 0))*u

## Quantile - LL ##

Q.ll = function(b,a,x0,x,y,h1,k1){
   b0 = b[1]
   b1 = b[2]
   sum(ro.u(y-b0-b1*(x-x0), a)*k1((x-x0)/h1))
}

## Quantile - Double - kernel ##

Q.db = function(qs,a,x0,x,y,k1,o1,h1,h2){
   s1 = sum(k1((x-x0)/h1)*(x-x0))
   s2 = sum(k1((x-x0)/h1)*(x-x0)^2)
   w.x = k1((x-x0)/h1)*(s2-(x-x0)*s1)
   sum(w.x*o1((qs-y)/h2))/sum(w.x)
}

## DGP ##

set.seed(1991)
n = 500
x = runif(n,0,6)
m.x = exp(sin(x)^3)
e = rnorm(n,0,0.5)
y = m.x + e
plot(x,y,pch=20)
lines(sort(x),m.x[order(x)],col=1)

## Plot Local Linear ##

q.opt.ll = rep(0,n)
```

```
a1 = 0.1
h1.ot = 0.2
for(i in 1:n){
   reg.ot =
       optim(par=c(0,0),fn=Q.ll,a=a1,x0=x[i],x=x,y=y,h1=h1.ot,k1=dnorm,method="BFGS")
   q.opt.ll[i] = reg.ot$par[1]
}

lines(sort(x),q.opt.ll[order(x)],col=2)

## Plot Double Kernel ##

nq = 100
qc = seq(from=min(y),to=max(y),length=nq)
h2.ot = h1.ot
q.x = rep(0,n)
for(i in 1:n){
   F.h = rep(0,nq)
   for(k in 1:nq){
      F.h[k] =
          Q.db(qc[k],a=a1,x0=x[i],x,y,k1=dnorm,o1=punif,h1=h1.ot,h2=h2.ot)
   }
   pos.ot = which.min(abs(F.h-a1))
   q.x[i] = qc[pos.ot]
}

lines(sort(x), q.x[order(x)], col=7)
```

8. Endogeneidade e IV Não-Paramétrica

8.1 Endogeneidade Não-Paramétrica

Uma equação não-paramétrica para IV pode ser escrita como

$$\begin{aligned} Y_i &= g(X_i) + \varepsilon_i \\ \mathbb{E}\left[\varepsilon_i | Z_i\right] &= 0 \end{aligned} \tag{8.1.1}$$

Nesse modelo, alguns elementos de X_i são potencialmente endógenos e Z_i é exógeno.

A extensão para o caso em que g é não-linear não é óbvia como no caso paramétrico tradicional.

A primeira e principal questão é a identificação. O que significa g?

Assuma que

$$\lambda(z) = \mathbb{E}\left[Y_i | Z_i = z\right] \tag{8.1.2}$$

e tome o valor esperado em (8.1.1):

$$\begin{aligned} \lambda(z) &= \mathbb{E}\left[Y_i | Z_i = z\right] \\ &= \mathbb{E}\left[g(X_i) + \varepsilon_i | Z_i = z\right] \\ &= \mathbb{E}\left[g(X_i) | Z_i = z\right] + \mathbb{E}\left[\varepsilon_i | Z_i = z\right] \\ &= \int g(x) f(x|z) dx \end{aligned} \tag{8.1.3}$$

As funções $\lambda(z)$ e $f(x|z)$ são identificadas. A função não-paramétrica desconhecida $g(x)$ é solução da equação integral

$$\lambda(z) = \int g(x) f(x|z) dx \tag{8.1.4}$$

A dificuldade é que a solução $g(x)$ não é necessariamente única. O problema matemático é que a solução g não é necessariamente contínua na função f. A não unicidade de g é chamada de "problema inverso mal colocado".

Uma solução é restringir o espaço de funções permitidas g: por exemplo, o modelo linear $g(x) = x'\beta$ é linear, então a equação acima se reduz a

$$\begin{aligned} \lambda(z) &= \beta' \int x f(x|z) dx \\ &= \beta' \mathbb{E}\left[X_i | Z_i = z\right] \end{aligned} \tag{8.1.5}$$

A identificação de β no modelo linear explora esta relação simples.

8.2 Estimador de Newey-Powell-Vella

Nessa abordagem de equações triangulares simultâneas o modelo é escrito como

$$\begin{aligned} Y_i &= g(X_i) + \varepsilon_i \\ \mathbb{E}(\varepsilon_i | Z_i) &= 0 \end{aligned} \tag{8.2.1}$$

e escrevemos uma equação na forma reduzida para X_i, de modo que

$$\begin{aligned} X_i &= \Xi(Z_i) + u_i \\ \mathbb{E}\left[u_i | Z_i\right] &= 0 \end{aligned} \tag{8.2.2}$$

Portanto, $\Xi(Z_i)$ é a média condicional de X_i dado $Z_i = z$. Os vetores X_i e Z_i podem se sobrepor.

Assim,

$$\begin{aligned}\mathbb{E}\left[Y_i|X_i, Z_i\right] &= \mathbb{E}\left[g(X_i) + \varepsilon_i|X_i, Z_i\right] \\ &= g(X_i) + \mathbb{E}\left[\varepsilon_i|X_i, Z_i\right] \end{aligned} \tag{8.2.3}$$

Dado que X_i é endógeno, a expressão $\mathbb{E}\left[\varepsilon_i|X_i, Z_i\right]$ não é igual a zero. Em geral, isso não pode ser simplificado ainda mais. Mas os autores observaram o seguinte. De (8.2.2), X_i é uma função de Z_i e u_i, então o condicionamento em X_i e Z_i é equivalente ao condicionamento em u_i e Z_i. Consequentemente

$$\mathbb{E}\left[Y_i|X_i, Z_i\right] = g(X_i) + \mathbb{E}\left[\varepsilon_i|u_i, Z_i\right] \tag{8.2.4}$$

Supondo que Z_i é fortemente exógeno, então

$$\mathbb{E}\left[\varepsilon_i|u_i, Z_i\right] = \mathbb{E}\left[\varepsilon_i|u_i\right] = g_2(u_i) \tag{8.2.5}$$

ou seja, condicional em u_i, Z_i não fornece nenhuma informação sobre a média do erro ε_i. Neste caso temos a simplificação

$$\mathbb{E}\left[Y_i|X_i, Z_i\right] = g(X_i) + g_2(u_i) \tag{8.2.6}$$

o que implica

$$\begin{aligned} Y_i &= g(X_i) + g_2(u_i) + \varepsilon_i \\ \mathbb{E}\left[\varepsilon_i|u_i, X_i\right] &= 0 \end{aligned} \tag{8.2.7}$$

Este é um modelo de regressão aditiva, cujo regressor u_i não observado, mas identificável. E seria a função g identificável? Dado que (8.3.1) é uma regressão na forma reduzida, Ξ é identificável e, portanto, u_i é identificável.

As funções g e g_2 são identificadas contanto que u_i e X_i sejam distintos. Trabalhando com a diferença de duas expectativas condicionais, a identificação equivale à afirmação de que uma função aditiva deve ter apenas componentes

constantes. Para ser mais preciso, temos o seguinte teorema.

Teorema 8.2.1. *Se não houver relação funcional entre (x, z) e u, então $g(x)$ é identificada sob uma constante aditiva.*

A qualificação de constante aditiva é exigida em todos os modelos não-paramétricos aditivos. Embora seja uma condição suficiente, a inexistência de uma relação funcional entre (x, z) e u não é uma condição necessária para a identificação. É a inexistência de uma relação funcional aditiva que é uma condição necessária e suficiente. Assim, a identificação ainda pode ocorrer quando há uma relação funcional exata e não aditiva. A estrutura aditiva é tão forte que o modelo pode ser identificado mesmo quando a condição de ordem usual não é satisfeita, ou seja, mesmo que z tenha dimensão menor que (x, z_1).

Os autores propõem o seguinte estimador de série:

1: Estimar $\hat{\Xi}_L(z) = \hat{\theta}_L' Z_{Li}$ usando uma série em Z_i com L termos, digamos

  # 1a: $\hat{\theta}_L = \left(Z_L' Z_L\right)^{-1} Z_L' X$ em que Z_L são funções básicas de Z

  # 1b: Faça $\hat{u}_i = X_i - \hat{\Xi}(X_i)$

2: Criar uma base (transformação) para X_i e uma outra base para $\hat{u}_i$ com K coeficientes

  # 2a: spline para X_i

  # 2b: spline para $\hat{u}_i$

3: Regredir por mínimos quadrados Y_i contra as bases criadas, obtendo $\hat{g}$ e $\hat{g}_2$

As condições para consistência requerem que as funções g e Ξ sejam suficientemente suaves (existam derivadas o suficiente), e que o número de termos K e L tendam para infinito de uma maneira controlada. As condições de regularidade não são particularmente úteis. Não está claro como K e L devem ser selecionados na prática. Uma sugestão razoável é selecionar L por validação cruzada na regressão da forma reduzida e, em seguida, selecionar K por validação cruzada na regressão do segundo estágio. O problema é que os dois estágios não são ortogonais, de modo que o erro quadrático médio do segundo estágio é afetado pelo primeiro estágio, de modo que é improvável que o critério de CV reflita isso corretamente. A matriz de variância-covariância pode ser obtida por GMM.

8.3 Estimador de Newey-Powell

Newey e Powell propõem um método não paramétrico que evita a suposição de exogeneidade forte, mas impõe restrições a função g. Seja o modelo inicial,

$$\begin{aligned} Y_i &= g(X_i) + \varepsilon_i \\ \mathbb{E}(\varepsilon_i | Z_i) &= 0 \end{aligned} \tag{8.3.1}$$

e seja a equação integral

$$\mathbb{E}\left[Y_i | Z_i\right] = \int g(x) f(x|Z_i) dx \tag{8.3.2}$$

Para identificar g, os autores apontam que uma solução é assumir que g está em um espaço compacto. A abordagem é restringir a função verdadeira a ser um elemento de um conjunto compacto de funções, impondo limites em derivadas de ordem superior, o que torna o mapeamento da forma reduzida para a estrutura contínua.

Assim, supõe-se que $g(x)$ pode ser aproximada usando uma aproximação por séries. Isso pode ser escrito como

$$g(x) \approx g_K(x) = \gamma_K' p_K(x) \tag{8.3.3}$$

em que γ_K é um vetor de parâmetros e $p_K(x)$ é um vetor de funções de base. A compacidade de g pode ser imposta ao assumir que γ_K é limitada. Para tanto, se utiliza do fato de que $\gamma_K' W_K \gamma_K \leq C$ em que W_K é uma matriz de peso e C é uma constante pré-determinada.

Assim,

$$\begin{aligned} \mathbb{E}\left[Y_i | Z_i\right] &\approx \gamma_K' \int p_K(x) f(x|Z_i) dx \\ &= \gamma_K' \mathbb{E}\left[p_K(X_i) | Z_i\right] \\ &= \gamma_K' h_K(Z_i) \end{aligned} \tag{8.3.4}$$

em que $h_K(Z_i) = \mathbb{E}[p_K(X_i)|Z_i = z]$.

Portanto, obtemos o seguinte modelo:

$$\begin{aligned} Y_i &= \gamma_K' h_K(Z_i) + v_i \\ \mathbb{E}[v_i|Z_i] &= 0 \end{aligned} \tag{8.3.5}$$

e

$$\begin{aligned} p_K(X_i) &= h_K(z) + \eta_i \\ \mathbb{E}[\eta_i|Z_i] &= 0 \end{aligned} \tag{8.3.6}$$

Os autores propõem o seguinte estimador:

1: Selecionar as funções de base $p_K(x)$

2: Regredir não-paramétrica cada elemento de $p_K(x)$ contra Z_i usando método de séries. As estimativas são coletadas no vetor $\hat{h}_K(z)$.

3: Regredir Y_i contra $\hat{h}_K(z)$ (por mínimos quadrados) para obter $\hat{\gamma}_K$ (impor $\gamma_K' W_K \gamma_K \leq C$ para garantir a compacidade de g e, portanto, de $\hat{g}$).

4: As estimativas de interesse são $\hat{g}(x) = \gamma_K' p_K(x)$.

O estimador $\hat{\gamma}_K$ é um estimador NP2SLS. Portanto, os erros-padrão convencionais para $\hat{\gamma}_K$ e, portanto, para $\hat{g}$ são incorretos.

9. Bootstrap e Jackknife

9.1 Introdução

A estimação por ponto é bastante útil, embora deixe alguma coisa a desejar, isto é, ela não da indicação da precisão a ela associada. No caso em que a função distribuição de probabilidade do estimador por ponto, sob consideração, for contínua a probabilidade de que o estimador seja igual ao valor do parâmetro é zero. Portanto, parece desejável que uma estimativa por ponto deva ser acompanhada por alguma medida do possível erro da estimativa.

Portanto, parece desejável que uma estimativa por ponto deva ser acompanhada por alguma medida do possível erro da estimativa. Por exemplo, uma estimativa por ponto pode ser acompanhada por algum intervalo em torno da estimativa por ponto junto com alguma medida de segurança de que o verdadeiro valor do parâmetro pertença a esse intervalo.

Um estimador por intervalo é uma regra que especifica o método para usar as medidas amostrais para calcular dois números que formam os extremos do intervalo. Idealmente, gostaríamos que o intervalo resultante tivesse duas propriedades, a saber:

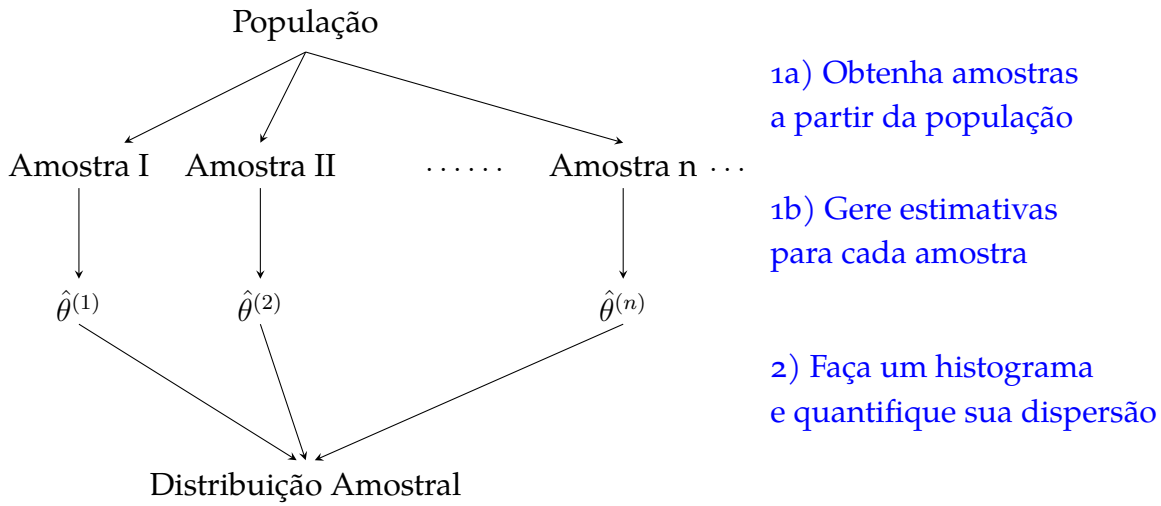
1. que "cobrisse" o verdadeiro valor do parâmetro de interesse, θ.
2. que o intervalo fosse relativamente pequeno.

Existe mais de um intervalo de confiança com o mesmo coeficiente de confiança $(1 - \alpha)$, mas estaremos interessados em obter aquele que tem o menor comprimento dentro de uma certa classe de IC.

Especificamente, aprendemos como a distribuição amostral formaliza o conceito de "incerteza estatística" em termos de um experimento mental: o que

aconteceria se pudéssemos executar a mesma análise em muitos universos paralelos imaginários, onde em cada universo paralelo experimentamos uma única realização do mesmo processo de geração de dados aleatórios? Na Figura 9.1.1 está a imagem para ilustrar essa ideia.

Figura 9.1.1 – Processo de Amostragem



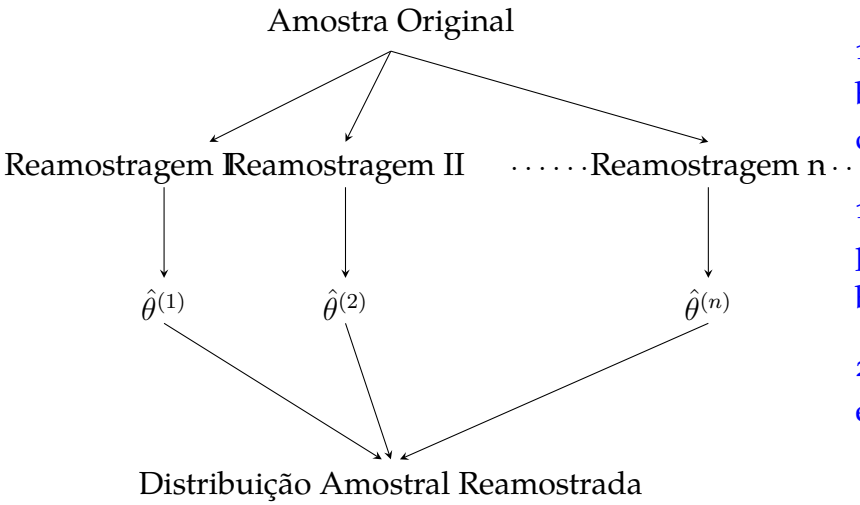
O *bootstrap* e o *jackknife* são métodos não paramétricos para calcular erros padrão e intervalos de confiança. O *jackknife* é menos caro computacionalmente, mas o *bootstrap* tem algumas vantagens estatísticas.

No centro da abordagem de reamostragem para inferência estatística está uma ideia simples. Na maioria das vezes, não podemos pegar amostras repetidas do mesmo processo aleatório que gerou nossos dados, para ver como nossa estimativa muda de uma amostra para outra. Mas podemos repetidamente tirar amostras da própria amostra e aplicar nosso estimador novamente a cada amostra. A variabilidade das estimativas em todas essas reamostras pode ser usada para aproximar a distribuição amostral verdadeira do nosso estimador.

Esse processo – fingir que nossa amostra representa alguma população e tirar amostras repetidas de tamanho N com reposição de nossa amostra original de tamanho N – é chamado de reamostragem de *bootstrap*, ou apenas *bootstrap*.

Por que isso funcionaria? Lembre-se de que a incerteza surge da aleatoriedade inerente ao nosso processo de geração de dados (seja variabilidade de amostragem, erro de medição, o que for). Portanto, se pudermos simular aproximadamente essa aleatoriedade, podemos quantificar aproximadamente nossa incerteza. Esse é o objetivo do *bootstrap*: aproximar a aleatoriedade inerente ao processo de geração

Figura 9.1.2 – Processo de Bootstrap



de dados, para que possamos simular o experimento mental central da inferência estatística.

Na Figura 9.1.2 está a aparência do *bootstrap* em uma imagem.

Cada bloco de N pontos de dados reamostrados é chamado de "amostra de inicialização". Para o *bootstrap*, escrevemos um programa de computador que reamostra repetidamente nossa amostra original e recalcula nossa estimativa para cada amostra de *bootstrap*.

Existem duas propriedades principais do *bootstrap* que fazem essa ideia aparentemente maluca realmente funcionar. Primeiro, cada amostra *bootstrap* deve ser do mesmo tamanho (N) da amostra original. Lembre-se, temos que aproximar a aleatoriedade em nosso processo de geração de dados, e o tamanho da amostra é uma parte absolutamente fundamental desse processo. Se pegarmos amostras *bootstrap* de tamanho $N/2$, ou $N - 1$, ou qualquer outra coisa que não N, estamos simulando o processo de geração de dados "errado".

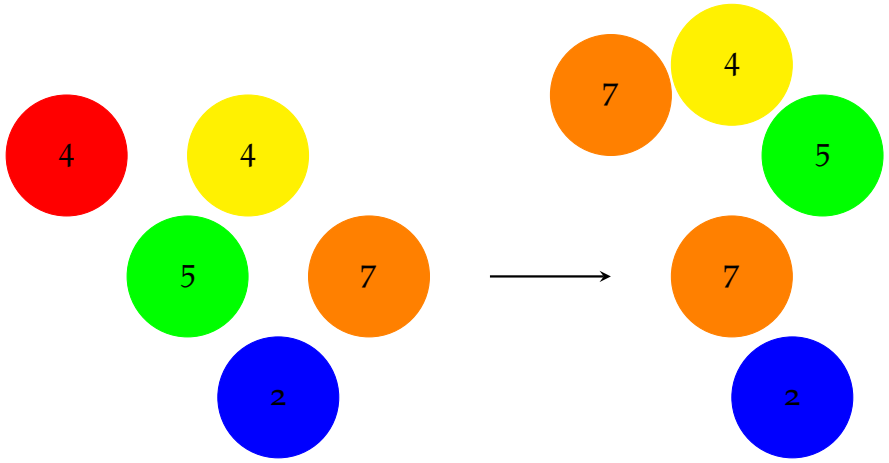
Em segundo lugar, cada amostra *bootstrap* deve ser retirada com reposição da amostra original. A intuição aqui é que cada amostra *bootstrap* terá seu próprio padrão aleatório de duplicatas e omissões em comparação com a amostra original, criando variabilidade de amostragem sintética que se aproxima da verdadeira variabilidade de amostragem.

Para entender a amostragem com reposição, imagine um sorteio de loteria, onde há um grande balde com bolas numeradas. Escolhemos 6 números do balde. Depois de escolhermos uma bola, podemos fazer uma de duas coisas: 1) colocar a

bola de lado, ou 2) registrar o número na bola e depois jogá-la de volta no balde. Se você colocar uma bola de lado depois de ter sido escolhida, ela poderá ser escolhida apenas uma vez; isso é amostragem sem reposição, e é o que acontece em uma loteria real. Mas se, em vez disso, você colocar a bola de volta no balde, ela tem uma pequena chance de ser escolhida novamente e, portanto, ser representada mais de uma vez no conjunto final de 6 números da loteria. Isso é amostragem com substituição, e é o que fazemos quando iniciamos.

Vamos ver como esse processo de "amostragem com substituição" funciona no contexto de *bootstrap* e por que é importante. No desenho abaixo, o painel esquerdo mostra uma amostra hipotética de tamanho $N = 5$, enquanto o painel direito mostra uma amostra *bootstrap* desta amostra original.

Observe as duas propriedades principais de nossa amostra *bootstrap*: (1) a amostra *bootstrap* tem o mesmo tamanho ($N = 5$) da amostra original; e (2) a amostra *bootstrap* foi retirada com reposição e, portanto, apresenta um padrão aleatório de duplicatas e omissões quando comparada com a amostra original. Especificamente, o 4 vermelho foi totalmente omitido, enquanto o 7 laranja foi escolhido duas vezes.



Amostra ($n = 5$) Bootstrap ($n = 5$)

Por que isso importa? Bem, vamos ver o que acontece quando calculamos a média amostral da amostra *bootstrap* versus a amostra original:

$$\text{Média amostral da amostra original} = \frac{2+4+4+5+7}{5} = 4.4$$

$$\text{Média amostral da amostra bootstrap} = \frac{2+4+5+7+7}{5} = 5$$

E este é o fato central a ser observado: quando computamos uma estatística de resumo para a amostra *bootstrap*, não obteremos necessariamente a mesma resposta que obtivemos para a amostra original.

Este é o mecanismo central pelo qual o *bootstrap* funciona: a reamostragem cria variabilidade sintética em nossos resumos estatísticos, de uma forma projetada para aproximar a variabilidade de amostragem real. Se repetirmos esse processo milhares de vezes, alguns resumos serão muito altos, alguns serão muito baixos e alguns serão perfeitos quando comparados com a resposta da amostra original. A questão é que os resumos *bootstrap* diferem uns dos outros – e a quantidade pela qual eles diferem uns dos outros nos fornece uma medida quantitativa de nossa incerteza estatística.

E essa é a ideia básica:

1. Você tem certeza se seus resultados são repetíveis em diferentes amostras do mesmo processo de geração de dados aleatórios.

2. *Bootstrap* permite medir a repetição de seus resultados, aproximando o processo de amostragem aleatória da população mais ampla.

9.2 Jackknife

O *jackknife*, devido a Quenouille (1949), é um método simples para aproximar o viés e a variância de um estimador. Seja $T_n = T(X_1, \ldots, X_n)$ um estimador de alguma quantidade θ e defina viés$(T_n) = \mathbb{E}[T_n] - \theta$. Seja $T_{(-i)}$ a estatística com i-ésima a observação removida. A estimativa do viés é definida por

$$b_{jack} = (n-1)\left(\overline{T}_n - T_n\right) \tag{9.2.1}$$

em que $\overline{T}_n = \dfrac{1}{n}\sum_i T_{(-i)}$. O estimador corrigido é $T_{jack} = T_n - b_{jack}$.

Por que b_{jack} é definido dessa forma? Para muitas estatísticas pode ser mostrado que

$$\text{viés}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \tag{9.2.2}$$

para algum a e b.

Por exemplo, seja $\sigma^2 = \mathrm{V}[X_i]$ e seja $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. De modo que,

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n}X_j + \frac{1}{n^2}\sum_{j=1}^{n}X_j\sum_{k=1}^{n}X_k\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n-2}{n}\mathbb{E}\left[X_i^2\right] - \frac{2}{n}\sum_{j\neq i}\mathbb{E}\left[X_iX_j\right] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\mathbb{E}\left[X_jX_k\right] + \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[X_j^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}\left(\sigma^2+\mu^2\right) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}\left(\sigma^2+\mu^2\right)\right] \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned} \tag{9.2.3}$$

Então, $\text{viés}(\hat{\sigma}_n^2) = -\frac{\sigma^2}{n}$. Portanto, (9.2.2) é válido como $a = -\sigma^2$ e $b = 0$. De modo semelhante,

$$\text{viés}\left(T_{(-i)}\right) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right) \tag{9.2.4}$$

Segue então que

$$\begin{aligned}
\mathbb{E}[b_{jack}] &= (n-1)\left\{\mathbb{E}[\text{viés}(\overline{T}_n)] - \mathbb{E}[\text{viés}(T_n)]\right\} \\
&= (n-1)\left[\frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right) - \frac{a}{n} - \frac{b}{n^2} - O\left(\frac{1}{n^3}\right)\right] \\
&= (n-1)\left[\frac{a}{n(n-1)} + \frac{(2n-1)b}{n^2(n-1)^2} + O\left(\frac{1}{n^3}\right)\right] \\
&= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right)
\end{aligned}$$

$$= \text{viés}(T_n) + O\left(\frac{1}{n^2}\right) \tag{9.2.5}$$

De forma similar,

$$\begin{aligned} \text{viés}(T_{jack}) &= n(T_n - b_{jack}) \\ &= -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{9.2.6}$$

então o viés de T_{jack} é de uma ordem de magnitude menor do que T_n. Assim, T_{jack} pode ser escrito como

$$T_{jack} = \frac{1}{n}\sum_{i=1}^{n} \tilde{T}_i \tag{9.2.7}$$

em que $\tilde{T}_i = nT_n - (n-1)T_{(-i)}$ são chamados de pseudo-valores.

O estimador *jackknife* da variância de T_n é dado por

$$v_{jack} = \frac{\tilde{s}^2}{n} \tag{9.2.8}$$

em que $s^2 = \dfrac{\sum_{i=1}^{n}\left(\tilde{T}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_i\right)^2}{n-1}$ é a variância amostral dos pseudo-valores.

A partir disso, obtemos dois teoremas importantes.

Teorema 9.2.1. *Seja $\mu = \mathbb{E}[X_1]$ e $\sigma^2 = \mathrm{V}[X_1] < \infty$ e suponha que $T_n = g\left(\overline{X}_n\right)$, em que g tem derivada contínua não-nula em μ. Então, $\dfrac{T_n - g(\mu)}{\sigma^2} \rightsquigarrow \mathcal{N}(0,1)$, em que $\sigma_n^2 = n^{-1}\left(g'(\mu)\right)^2\sigma^2$. O estimador jackknife é consistente, o que significa que*

$$\frac{v_{jack}}{\sigma_n^2} \xrightarrow{a.s.} 1 \tag{9.2.9}$$

Teorema 9.2.2 (Efron, 1982)**.** *Se $T(F) = F^{-1}(p)$ é o p-ésimo quantil, então a estimativa jackknife da variância é inconsistente. Para a mediana nós temos que $\dfrac{v_{jack}}{\sigma_n^2} \rightsquigarrow \left(\dfrac{\chi_2}{2}\right)^2$, em que σ_n^2 é a variância assintótica da mediana amostral.*

Nota-se que o erro padrão não é uma estimativa de uma quantidade pertinente a uma população, mas uma medida da incerteza da média amostral vista como uma estimativa da média populacional.

9.3 Bootstrap

O *bootstrap* é um método para estimar a variância e a distribuição de uma estatística $T_n = g(X_1, \ldots, X_n)$. Também podemos usar o *bootstrap* para construir intervalos de confiança. Uma vantagem do *bootstrap* é que esta técnica não depende inteiramente do teorema central do limite, já que, em suas aplicações, medidas de precisão são obtidas diretamente dos dados (Efron & Tibshirani, 1993, p.40). O estimador *bootstrap* é chamado estimador *bootstrap* não paramétrico, já que se baseia em $\hat{F}$, um estimador não paramétrico de F. Um estimador *bootstrap* paramétrico do erro padrão é baseado em um estimador $\hat{F}$ de F derivado de um modelo paramétrico. Por exemplo, ao invés de estimarmos F pela função distribuição empírica $\hat{F}$, podemos assumir que a população tem distribuição normal.

Seja $\mathrm{V}_F[T_n]$ a variância de T_n. Adicionamos o subscrito F para enfatizar que a variância é uma função de F. Se nós conhecemos F, podemos, em princípio, computar a variância. Por exemplo, se $T_n = n^{-1}\sum_{i=1}^{n} X_i$, então

$$\mathrm{V}_F[T_n] = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n} \tag{9.3.1}$$

que é claramente uma função de F.

Com o *bootstrap* nós estimamos $\mathrm{V}_F[T_n]$ a partir de $\mathrm{V}_{\hat{F}_n}[T_n]$. Em outras palavras, nós usamos um estimador de *plug-in* da variância. Visto que $\mathrm{V}_{\hat{F}_n}[T_n]$ pode ser difícil de calcular, nós o aproximamos com uma estimativa de simulação denotada por v_{boot}. Especificamente, seguimos os seguintes passos:

1. Amostre $X_1^*, \ldots, X_n^* \sim \hat{F}_n$
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$
3. Repita os passos 1 e 2 B vezes para obter $T_{n,1}^*, \ldots, T_{n,B}^*$

4. $v_{boot} = \dfrac{1}{B}\sum_{b=1}^{B}\left(T^*_{n,B} - \dfrac{1}{B}\sum_{r=1}^{B} T^*_{n,r}\right)^2$

Pela lei ds grandes números, $v_{boot} \overset{a.s.}{\to} \mathrm{V}_{\hat{F}_n}(T_n)$ quando $B \to \infty$. O erro padrão de T_n é $\widehat{dp}_{boot} = \sqrt{v_{boot}}$. A ideia do *boostratp* pode ser sumarizado como segue:

$$\begin{aligned} \text{Mundo real:} F \quad &\Longrightarrow \quad X_1, \ldots, X_n \quad \Longrightarrow \quad T_n = g(X_1, \ldots, X_n) \\ \text{Mundo do bootstrap:} \hat{F}_n \quad &\Longrightarrow \quad X^*_1, \ldots, X^*_n \quad \Longrightarrow \quad T^*_n = g(X^*_1, \ldots, X^*_n) \end{aligned} \tag{9.3.2}$$

Como simulamos a partir de $\hat{F}_n$? Uma vez que $\hat{F}_n$ atribui probabilidade $1/n$ para cada ponto de dados, extrair n pontos aleatoriamente de $\hat{F}_n$ é o mesmo que extrair uma amostra de tamanho n com reposição dos dados originais. Portanto, a etapa 1 pode ser substituída por:

1. Amostre $X^*_1, \ldots, X^*_n$ com reposição de $X_1, \ldots, X_n$

9.4 Intervalo de Confiança por Bootstrap

Há muitas formas de construir intervalos de confiança por *bootstrap*. Elas variam na forma de cálculo e na acurácia.

1. O mais simples é o intervalo normal

$$T_n \pm z_{\alpha/2}\widehat{dp}_{boot} \tag{9.4.1}$$

em que $\widehat{dp}_{boot}$ é a estimativa por *bootstrap* do erro padrão (apesar der tratado como conhecido) e $z_{\alpha/2} = \Phi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$. O fato de que $z \sim \mathcal{N}(0,1)$ está condicionado ao tamanho da amostra n implica que o intervalo assim encontrado pode ser impreciso. Além disto, características como assimetria e tendenciosidade podem estar presentes na distribuição de z, podendo prejudicar o desempenho do intervalo de confiança. Esse intervalo não é acurado a menos que a distribuição T_n esteja próxima de uma normal.

2. Intervalo pivotal: *bootstrap* básico

   Uma boa propriedade do intervalo de confiança construído pelo método percentil é a invariância a transformações monótonas. Este intervalo transforma a distribuição das réplicas do estimador por subtrair o valor observado da estatística.

   Seja $\theta = T(F)$ e $\hat{\theta}_n = T(\hat{F}_n)$ e defina a quantidade pivotal $R_n = \hat{\theta}_n - \theta$. Assuma que $H(r)$ denote a função de distribuição acumulada da quantidade pivotal

$$H(r) = \mathbb{P}_F[R_n \leq r] \tag{9.4.2}$$

   Seja $C_n^* = (a, b)$ em que

$$a = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{9.4.3}$$

$$b = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right) \tag{9.4.4}$$

   Disso, segue que

$$\begin{aligned}
\mathbb{P}[a \leq \theta \leq b] &= \mathbb{P}\left[\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a\right] \\
&= H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\
&= H\left[H^{-1}\left(1 - \frac{\alpha}{2}\right)\right] - H\left[H^{-1}\left(\frac{\alpha}{2}\right)\right] \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha
\end{aligned} \tag{9.4.5}$$

   Portanto, C_n^* é um intervalo de confiança $1 - \alpha$ exato para θ. Infelizmente, a e b dependem da distribuição H desconhecida, mas podemos formar uma estimativa *bootstrap* de H:

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \leq r) \tag{9.4.6}$$

em que $R^*_{n,b} = \hat{\theta}^*_{n,b} - \hat{\theta}_n$. Denote por r^*_β o quantil amostral β de $(R^*_{n,1}, \ldots, R^*_{n,B})$ e por θ^*_β o quantil amostral β de $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$. Note que $r^*_\beta = \theta^*_\beta - \hat{\theta}_n$. Segue que um intervalo de confiança $1-\alpha$ aproximado é $C_n = (\hat{a}, \hat{b})$ em que

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r^*_{1-\alpha/2} = 2\hat{\theta}_n - \theta^*_{1-\alpha/2} \tag{9.4.7}$$

$$\hat{b} = \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r^*_{\alpha/2} = 2\hat{\theta}_n - \theta^*_{\alpha/2} \tag{9.4.8}$$

O intervalo de confiança pivotal de nível $1-\alpha$ obtido por *bootstrap* é

$$C_n = \left(2\hat{\theta}_n - \hat{\theta}^*_{(1-\alpha/2)B}, 2\hat{\theta}_n - \hat{\theta}^*_{(\alpha/2)B}\right) \tag{9.4.9}$$

Segue o seguinte teorema.

Teorema 9.4.1. *Se $F(T)$ é Hadamard diferenciável e C_n é dado por (9.4.9), então $\mathbb{P}[T(F) \in C_n] \to 1-\alpha$.*

3. t-*Bootstrap*

   O procedimento *t-bootstrap*, também conhecido como método pivotal, segundo Efron & Tibshirani (1993, p.161) é uma generalização do usual método t de Student, sendo particularmente aplicável a estatísticas de localização, como a média amostral, a mediana, ou percentis amostrais. Estes autores citam que, pelo menos em sua forma tradicional, o método *t-bootstrap* não é bom para a construção de intervalos para outras estatísticas, como por exemplo, o coeficiente de correlação. Nestes casos, é necessário o uso de transformações. O *t-bootstrap* não usa distribuição t-Student como referência, mas usa a distribuição amostral de uma estatística (studenized) gerada por reamostragem.

   Defina

$$Z_n = \frac{T_n - \theta}{\widehat{se}_{boot}} \tag{9.4.10}$$

   e

$$Z_{n,b}^* = \frac{T_{n,b}^* - \theta}{\widehat{se}_b^*} \tag{9.4.11}$$

em que $\widehat{se}_b^*$ é uma estimativa do erro padrão de $T_{n,b}^*$ e não de T_n. Agora raciocinamos como no intervalo central. Os quantis amostrais das quantidades de *bootstrap* $Z_{n,1}^*, \ldots, Z_{n,B}^*$ deve aproximar os quantis verdadeiros da distribuição de Z_n. Seja z_α^* o α-ésimo quantil da amostra de $Z_{n,1}^*, \ldots, Z_{n,B}^*$, então $\Pr(Z_n \leq z_\alpha^*) \approx \alpha$. Seja

$$C_n = \left(T_n - z_{1-\alpha/2}^* \widehat{se}_{boot}, T_n - z_{\alpha/2}^* \widehat{se}_{boot}\right) \tag{9.4.12}$$

Então,

$$\begin{aligned}
\mathbb{P}[\theta \in C_n] &= \mathbb{P}\left[T_n - z_{1-\alpha/2}^* \widehat{se}_{boot} \leq \theta \leq T_n - z_{\alpha/2}^* \widehat{se}_{boot}\right] \\
&= \mathbb{P}\left[z_{\alpha/2}^* \leq \frac{T_n - \theta}{\mathrm{se}_{boot}} \leq z_{1-\alpha/2}^*\right] \\
&= \mathbb{P}\left[z_{\alpha/2}^* \leq Z_n \leq z_{1-\alpha/2}^*\right] \\
&\approx 1 - \alpha
\end{aligned} \tag{9.4.13}$$

Esse intervalo tem maior precisão do que todos os intervalos discutidos até agora, mas há um problema: você precisa calcular $\widehat{se}_b^*$ para cada amostra de *bootstrap*. Isso pode exigir uma segunda inicialização dentro de cada inicialização

O intervalo pivotal de *t-bootstrap* é

$$\left(T_n - z_{1-\alpha/2}^* \widehat{se}_{boot}, T_n - z_{\alpha/2}^* \widehat{se}_{boot}\right) \tag{9.4.14}$$

Esse intervalo requer alguma explicação. Para cada replicação de *bootstrap* calculamos $\hat{\theta}^*$ e também precisamos do erro padrão $\widehat{se}^*$ de $\hat{\theta}^*$. Poderíamos fazer um *bootstrap* dentro do *bootstrap* (chamado de *bootstrap* duplo), mas isso

é caro do ponto de vista computacional. Em vez disso, calculamos $\widehat{se}^*$ usando o método não paramétrico delta aplicado à amostra de *bootstrap*.

4. *Bootstrap* percentílico

   Esta abordagem utiliza a distribuição empírica das réplicas *bootstrap* como a distribuição de referência.

   O intervalo é definido como

$$C_n = \left(T^*_{(B\alpha/2)}, T^*_{(B(1-\alpha/2))}\right) \tag{9.4.15}$$

   isto é, usa-se somente os quantis $\alpha/2$ e $1-\alpha/2$ da amostra de *bootstrap*. A justificativa para este intervalo é a seguinte. Suponha que exista uma transformação monótona $U = m(T)$ tal que $U \sim \mathcal{N}(\phi, c^2)$, em que $\phi = m(\theta)$. Não supomos que conheçamos a transformação, apenas que existe uma. Seja $U^*_b = m(T^*_b)$. Observe que $U^*_{(B\alpha/2)} = m\left(T^*_{(B\alpha/2)}\right)$ uma vez que uma transformação monótona preserva os quantis. Dado que $U \sim \mathcal{N}(\phi, c^2)$, o $\alpha/2$ quantil de U é $\phi - z_{\alpha/2}c$. Portanto, $U^*_{(B\alpha/2)} = \phi - z_{\alpha/2}c \approx U - z_{\alpha/2}c$ e $U^*_{(B(1-\alpha/2))} \approx U + z_{\alpha/2}c$. Logo,

$$\begin{aligned}
\mathbb{P}\left[T^*_{B\alpha/2} \leq \theta \leq T^*_{B(1-\alpha/2)}\right] &= \mathbb{P}\left[m\left(T^*_{(B\alpha/2)}\right) \leq m(\theta) \leq m\left(T^*_{B(1-\alpha/2)}\right)\right] \\
&= \mathbb{P}\left[U^*_{(B\alpha/2)} \leq \theta \leq U^*_{B(1-\alpha/2)}\right] \\
&\approx \mathbb{P}\left[U - z_{\alpha/2}c \leq \phi \leq U + z_{\alpha/2}c\right] \\
&= \mathbb{P}\left[-z_{\alpha/2} \leq \frac{U-\phi}{c} \leq z_{\alpha/2}\right] \\
&= 1 - \alpha
\end{aligned} \tag{9.4.16}$$

   Surpreendentemente, nunca precisamos conhecer m. Infelizmente, uma transformação de normalização exata raramente existirá, mas pode haver transformações de normalização aproximadas. Isso levou a um amplo conjunto de métodos de percentis ajustados, sendo o mais popular o intervalo BCa (corrigido por viés e acelerado).

5. BCa

O melhor intervalo de confiança do *bootstrap* é chamado BCa para "viés corrigido" e "ajustado para aceleração". Intervalos BCa são uma versão modificada de intervalos percentuais que têm melhores propriedades teóricas e melhor desempenho na prática. Para um intervalo de confiança de $100(1-\alpha)\%$, os quantis habituais $\alpha/2$ e $1-\alpha/2$ são ajustados por dois fatores: uma correção para viés e uma correção para assimetria. A correção de viés é denotada z_0 e o ajuste de assimetria ou "aceleração" é dado por a.

Um intervalo *bootstrap* BCa de confiança de $100(1-\alpha)\%$ é calculado por

$$\alpha_1 = \Phi^{-1}\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1+\hat{a}(\hat{z}_0 + z_{\alpha/2})}\right) \tag{9.4.17}$$

$$\alpha_2 = \Phi^{-1}\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1+\hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right) \tag{9.4.18}$$

em que

$$z_\alpha = \Phi^{-1}(\alpha) \tag{9.4.19}$$

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\left\{\hat{\theta}_b < \hat{\theta}\right\}\right) \tag{9.4.20}$$

$$\hat{a} = \frac{1}{6}\frac{\sum_{i=}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \hat{\theta}_{-i}\right)^3}{\left[\sum_{i=}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \hat{\theta}_{-i}\right)^2\right]^{3/2}} \tag{9.4.21}$$

no qual $\hat{z}_0$ mensura o viés mediano de $\hat{\theta}_b$, isto é, a diferença entre a mediana de $\hat{\theta}_b$ e $\hat{\theta}$. Segundo Efron & Tibshirani (1993, p.186), $\hat{z}_0$ representa a magnitude da tendenciosidade mediana de $\hat{\theta}_b$ em uma escala normalizada. Obteremos $\hat{z}_0$ igual a zero se exatamente a metade das medidas de $\hat{\theta}_b$ forem menores iguais a $\hat{\theta}$. E $\hat{a}$ mensura a taxa de mudança do erro padrão com relação ao verdadeiro parâmetro e é obtida por *jackknife*.

Assim, os limites sãos quantis empíricos das réplicas *bootstrap* e o intervalo BCa é $\left(\hat{\theta}_{\alpha_1}, \hat{\theta}_{\alpha_2}\right)$.

O parâmetro de aceleração $\hat{a}$ que aparece na fórmula BCa (9.4.21) parece misterioso. Sua ideia envolve uma transformação idealizada para a normalidade que não será conhecida na prática. Felizmente, $\hat{a}$ tem uma relação simples com a função score de Fisher, que o torna fácil de estimar. O valor ótimo de a equivale a um sexto da função score[1].

9.5 Teoremas Adicionais

Sob certos condições, $\hat{G}_n^*$ é um estimador consistente de $G_n(t) = \mathbb{P}(T_n \leq t)$. Para tornar isso mais preciso, seja $\mathbb{P}_{\hat{F}_n}(\cdot)$ a probabilidade de $\hat{F}_n$, tratando os dados originais $X_1, \ldots, X_n$ como fixos. Assumimos que $T_n = T(\hat{F}_n)$ é algum funcional de $\hat{F}_n$. Então,

$$\hat{G}_n^*(t) = \mathbb{P}_{\hat{F}_n}\left[T\left(\hat{F}_n^*\right)\right] = \mathbb{P}_{\hat{F}_n}\left[\sqrt{n}\left(T\left(\hat{F}_n^*\right) - T(F)\right) \leq u\right] \tag{9.5.1}$$

em que $u = \sqrt{n}\,(t - T(F))$. A consistência do *bootstrap* pode ser expressa como segue.

Teorema 9.5.1. *Suponha que* $\mathbb{E}\left[X_1^2\right] < \infty$. *Seja,* $T_n = g(\overline{X}_n)$ *em que* g *é continuamente diferenciável em* $\mu = \mathbb{E}(X_1)$. *Então,*

$$\sup_u \left|\mathbb{P}_{\hat{F}_n}\left[\sqrt{n}\left(T\left(\hat{F}_n^*\right) - T\left(\hat{F}_n\right) \leq u\right)\right] - \mathbb{P}_F\left[\sqrt{n}\left(T\left(\hat{F}_n\right) - T\left(F\right) \leq u\right)\right]\right| \xrightarrow{a.s.} 0 \tag{9.5.2}$$

Teorema 9.5.2. *Suponha que* $T(F)$ *é Hadamard diferenciável com respeito a* $d(F,G) = \sup_x |F(x) - G(x)|$ *e que* $0 < \int L_F^2(x)dF(x) < \infty$. *Então,*

$$\sup_u \left|\mathbb{P}_{\hat{F}_n}\left[\sqrt{n}\left(T\left(\hat{F}_n^*\right) - T\left(\hat{F}_n\right) \leq u\right)\right] - \mathbb{P}_F\left[\sqrt{n}\left(T\left(\hat{F}_n\right) - T\left(F\right) \leq u\right)\right]\right| \xrightarrow{P} 0 \tag{9.5.3}$$

[1] Sendo $\ell(\theta)$ a função de verossimilhança, $\dfrac{d\ell(\theta)}{d\theta}$ é a função socre e $-\dfrac{d^2\ell(\theta)}{d\theta^2}$ é a função de informação de Fisher.

Também pode ser mostrado que a estimativa de variância *bootstrap* é consistente com algumas condições em T. Geralmente, as condições de consistência do *bootstrap* são mais fracas do que para o *jackknife*. Por exemplo, a estimativa *bootstrap* da variância da mediana é consistente, mas a estimativa *jackknife* da variância da mediana não é consistente.