Introdução à Econometria Semiparamétrica

Aula 1 - Estimação Não Paramétrica Moderna

Luis A. F. Alvarez

4 de outubro de 2024

Estimando uma densidade

- Suponha que você, pesquisador, possui uma amostra aleatória $\{X_i\}_{i=1}^n$ de uma distribuição contínua F, cuja densidade denotamos por f.
  - Por exemplo, você possui dados da riqueza não financeira de uma amostra de indivíduos.
- Você está interessado em estimar a densidade f.
- Como você pode fazer isso?

Estimação paramétrica

- Um caminho consiste em realizar uma hipótese de forma funcional acerca de f.
- Por exemplo, se nossa população de interesse está restrita à cauda superior da distribuição de riqueza, a literatura sugere que deve valer a "lei de Pareto" (Benhabib e Bisin, 2018), segundo a qual:

$$f(x) = g(x|x_0, \alpha) = \alpha \frac{x_0^\alpha}{x^{\alpha+1}} \mathbf{1}\{x \geq x_0\}, \quad x_0, \alpha > 0$$

  - Distribuição compatível com modelos de agentes heterogêneos com mercados incompletos e ativos arriscados (Achdou et al., 2022).
- Sob a hipótese de forma funcional, a estimação de f se resume à estimação de dois parâmetros, $(x_0, \alpha) \in \mathbb{R}^2$.
- Por exemplo, podemos estimá-los através de máxima verossimilhança:

$$(\hat{x}_0, \hat{\alpha}) \in \text{argmax}_{(z,a)\in\mathbb{R}^2} \frac{1}{n} \sum_{i=1}^{n} \log g(X_i|z, a)$$

Forma funcional e estimação não paramétrica

- E se a lei de Pareto não for válida?
  - Nesse caso, a densidade estimada não recuperará a distribuição correta, mesmo em amostras grandes.
- Como podemos estimar a distribuição correta sem requerer a hipótese de forma funcional?
- Para isso, necessitamos de métodos não paramétricos, que dispensem da hipótese de forma funcional.

Histograma

- O estimador não paramétrico mais simples é o histograma.
- Em um histograma, o suporte da distribuição é dividio em J células (*bins*) de comprimento $2h$, e a densidade para os pontos na j-ésima célula, $[a_j, b_j)$, é estimada como constante igual a:

$$\hat{f}_h(x_j) = \frac{1}{n2h} \sum_{i=1}^{n} \mathbf{1}\left\{|X_i - x_j| \leq h\right\} \ ,$$

onde o $x_j = \frac{a_j + b_j}{2}$ é o ponto intermediário da célula j.

Estimador de *kernel* para densidade

- O estimador $\hat{f}_h(x_j)$ usado na construção do histograma é um caso particular de um estimador de *kernel* para uma densidade em um ponto x, que é dado por:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right),$$

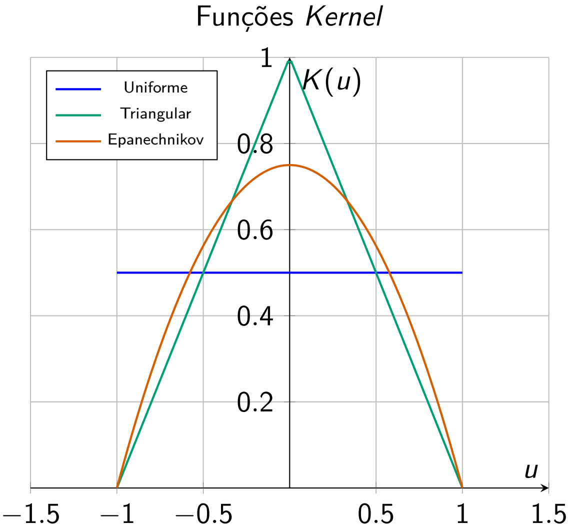
onde $K : \mathbb{R} \mapsto \mathbb{R}$ é um *kernel*, i.e. uma função satisfazendo $\int_{-\infty}^{\infty} K(u)du = 1$ e simétrica em torno do zero, $K(u) = K(-u)$.

- No que segue, vamos supor que K é uma função não negativa com suporte no intervalo $[-1, 1]$, embora essas hipóteses não sejam necessárias para a maioria dos resultados.
- Sob a hipótese anterior, estimador de *kernel* propõe estimar $f(x)$ "contando" os pontos da amostra com distância a menos de h de x. K nos dá então o peso relativo de cada ponto dentro de essa região.

Exemplos de *kernel*

Kernel	**Fórmula**
Uniforme	$K(u) = \frac{1}{2} \cdot \mathbf{1}(\lvert u \rvert \leq 1)$
Triangular	$K(u) = (1 - \lvert u \rvert) \cdot \mathbf{1}(\lvert u \rvert \leq 1)$
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2) \cdot \mathbf{1}(\lvert u \rvert \leq 1)$

Exemplos de *kernel*



Calculando o viés do estimador de *kernel*

- Vamos analisar as propriedades estatísticas do estimador de *kernel*.
- Considere um ponto x do suporte da distribuição tal que $f(\cdot)$ seja duas vezes continuamente diferenciável numa vizinhança de x. Neste caso, temos que, à medida que $h \to 0$:

$$\mathbb{E}[\hat{f}_h(x)] = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right] =$$

$$\frac{1}{h}\int_{x-h}^{x+h} K\left(\frac{s-x}{h}\right) f(s)ds \overset{u=\frac{s-x}{h}}{=} \int_{-1}^{1} K(u)f(x+hu)du =$$

$$f(x)\underbrace{\int_{-1}^{-1} K(u)du}_{=1} + f'(x)h\underbrace{\int_{-1}^{-1} uK(u)du}_{=0} + \int_{-1}^{1}(uh)^2 f''(\tilde{x}_{hu})K(u)du =$$

$$f(x) + h^2 f''(x)\int_{-1}^{1} u^2K(u)du + o(h^2)$$

onde, da segunda para a terceira linha, aplicamos o teorema do valor médio de segunda ordem, e a última passagem usa continuidade (uniforme) da segunda derivada numa vizinhança de x.

Calculando a variância do estimador de *kernel*

- Do mesmo modo:

$$\mathbb{V}[\hat{f}_h(x)] = \frac{1}{nh^2}\mathbb{V}\left[K\left(\frac{X_1 - X}{h}\right)\right]$$

- Mas:

$$\mathbb{E}\left[K\left(\frac{X_1 - X}{h}\right)^2\right] = h\int_{-1}^{1} f(x)K^2(u)du + O(h^3)$$

e

$$\mathbb{E}\left[K\left(\frac{X_1 - X}{h}\right)\right]^2 = h^2 f(x)^2 + O(h^3)$$

- De modo que:

$$\mathbb{V}[\hat{f}_h(x)] = \frac{1}{nh}\int_{-1}^{1} f(x)K^2(u)du + \frac{1}{n}f(x)^2 + O\left(\frac{h}{n}\right)$$

Erro quadrático médio do estimador de *kernel*

- Combinando os pontos anteriores, temos que:

$$\text{EQM}[\hat{f}_h(x)] = \mathbb{E}[(\hat{f}_h(x) - f(x))^2] =$$

$$h^4 f''(x)^2 \left(\int_{-1}^{1} u^2 K(u) du \right)^2 + \frac{1}{nh} \int_{-1}^{1} f(x) K^2(u) du + \frac{1}{n} f(x)^2 + o(1)$$

- Portanto, se $n \to \infty$ com $h \to 0$ e $nh \to \infty$.

$$\text{EQM}[\hat{f}_h(x)] \to 0\,,$$

e o estimador de kernel é *consistente*.

- Além disso, da expressão acima escolha h ótima, que balanceia *trade-off* viés-variância, é dada por:

$$h \propto n^{-1/5}$$

Estimação de densidades multidimensionais

- O estimador anteriormente introduzido é facilmente extensível para a estimação de uma densidade em $\mathbb{R}^d$. Seja $\boldsymbol{f}$ uma densidade em $\mathbb{R}^d$, e $\{\boldsymbol{X}_i\}_{i=1}^n$ uma amostra aleatória desta distribuição. Podemos estimar $\boldsymbol{f}(\boldsymbol{x})$ como:

$$\hat{\boldsymbol{f}}_{\boldsymbol{h}}(x) = \frac{1}{nh_1 \ldots h_d} \sum_{i=1}^{n} \boldsymbol{K}_{\boldsymbol{h}}\left(\boldsymbol{X}_i - \boldsymbol{x}\right)$$

onde $\boldsymbol{K}_{\boldsymbol{h}}(\boldsymbol{u}) = \prod_{j=1}^{d} K(u_j/h_j)$, com K um *kernel* univariado.

- Nesse caso, viés é proporcional da ordem de $d \max_j h_j^2$ e variância a $\frac{1}{nh_1 \ldots h_d}$. Portanto, se:

$$\max_j h_j \to 0, \quad nh_1 \ldots h_d \to \infty\,,$$

estimador é consistente.

Estimador de Nadaraya-Watson para a esperança condicional

Bibliografia I

Achdou, Yves et al. (2022). "Income and wealth distribution in macroeconomics: A continuous-time approach". Em: *The review of economic studies* 89.1, pp. 45–86.

Benhabib, Jess e Alberto Bisin (2018). "Skewed wealth distributions: Theory and empirics". Em: *Journal of Economic Literature* 56.4, pp. 1261–1291.