USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Unha introdución á Regresión Modal

Ramón Daniel Regueiro Espiño

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Unha introdución á Regresión Modal

Ramón Daniel Regueiro Espiño

Xullo 2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Unha introdución á Regresión Modal
Breve descrición do contido
Nun contexto clásico, a función de regresión defínese como o valor esperado da variable resposta condicionada ao valor da variable explicativa. Asemade, o tratamento inferencial dos modelos de regresión adóitase facer baixo unha perspectiva paramétrica, xeralmente lineal. Non obstante, a media condicionada pode non resultar sempre axeitada para resumir a relación entre as variables, por exemplo cando a distribución da variable resposta condicionada á explicativa é asimétrica ou multimodal. Porén, nestes casos o estudo da moda ou modas condicionais pode supoñer unha mellor alternativa para modelar a relación entre as variables. Aínda que no escenario da regresión modal tanto o enfoque paramétrico como o non paramétrico son posibles, o modelado non paramétrico aparece de xeito máis natural, dado que a estimación da regresión modal se pode abordar mediante a maximización dun estimador flexible da densidade condicional, como pode ser o estimador tipo núcleo. Neste TFG presentarase unha introdución aos modelos de regresión modal non paramétricos, para o que resulta necesaria unha revisión previa da estimación non paramétrica da densidade.

Índice xeral

Resumo

A regresión modal xorde como alternativa á regresión en media para tratar de modelar relacións entre variables nos casos nos que esta poida non ser adecuada, dado que a media condicional non é sempre unha medida resumo que representa fielmente a relación entre as mesmas. Este traballo ten como obxectivo presentar a regresión modal non paramétrica.

Comezamos mostrando varias simulacións e un caso real onde a regresión en media paramétrica non é adecuada, para abordar así a estimación da densidade non paramétrica. Co fin de contextualizar a metodoloxía da regresión modal, farase unha análise da estimación tipo núcleo da densidade e da densidade condicional, destacando a forma de seleccionar unha fiestra de suavizado adecuada. Para isto, introduciranse varios criterios de erro e diferentes métodos de selección.

Para concluír, desenvolverase a estimación modal non paramétrica. Para introducir a regresión modal, farase unha breve alusión ao modelo linear unimodal. Posteriormente, ilustrarase o principal algoritmo para a construción do modelo non paramétrico multimodal, o *mean-shift*. A continuación, describiranse varias medidas de erro, as cales deben definirse para conxuntos no canto de puntos, coa idea de poder facer unha selección adecuada das fiestras. Mostraranse dous métodos de selección do ancho de banda cuxa idea xeral coincide co da estimación tipo núcleo. Finalmente, ilustrarase a regresión modal non paramétrica sobre os exemplos, simulados e real, onde a regresión en media paramétrica non era adecuada.

Abstract

Modal regression appears as an alternative to mean regression in order to try to model relationships between variables in cases where mean regression may not be adequate, since the conditional mean is not always a summary measure that faithfully represents the relationship between them. The aim of this work is to introduce nonparametric modal

regression.

We start showing several simulations and a real case where the mean parametric regression is not adequate, to address the estimation of non-parametric density. In order to contextualize the modal regression methodology, an analysis of the kernel density estimation and kernel conditional density estimation will be made, highlighting how to select an adequate bandwidth. For this, various error criteria and different selection methods will be introduced.

To conclude, non-parametric regression estimation will be developed. To introduce modal regression, a brief recall of the unimodal linear model will be made. Then, the main algorithm for the construction of the multimodal non-parametric model, the mean-shift algorithm, will be illustrated. Next, several error measures will be described, which must be defined for sets instead of points, with the idea of being able to make a proper bandwith selection. Two bandwidth selection methods will be shown whose general idea matches that of the kernel estimation approach. Finally, the non-parametric modal regression will be illustrated on the examples, simulated and real, where the mean parametric regression was not adequate.

Prefacio

A Estatística é unha disciplina que ten como un dos seus obxectivos a descrición de relacións entre varias variables. A día de hoxe, o seu emprego é unha ferramenta fundamental de campos tan diversos como a física, a economía ou a informática.

Así, dado un conxunto de observacións de varias variables, é importante sermos capaces de describir o comportamento dunha delas, a variable resposta, en relación ás outras, as variables explicativas. Isto dá lugar á creación de modelos de regresión, os cales se comezan a explicar na materia de 3º do Grao de Matemáticas *Inferencia Estatística*, profundizando neles na materia de 4º do Grao de Matemáticas *Modelos de Regresión e Análise Multivariante*.

Os primeiros modelos de regresión que se desenvolveron, e os máis habituais, son aqueles que realizan unha estimación da media da variable resposta. Estes modelos intentan predicir o valor que lle correspondería á variable resposta segundo os valores concretos das variables explicativas empregando a media. Ademais, se a función ten forma linear, diremos que é un modelo linear. Este caso foi o primeiro en estudarse e por iso recibe o nome de clásico. Por outro lado, se a función non ten unha expresión paramétrica sinxela e facilmente interpretable, para facer a estimación consideraremos modelos non paramétricos.

Porén, para realizar inferencia sobre os modelos estudados, estes requiren de supoñer a veracidade de distintas hipóteses como a normalidade dos erros, é dicir, que os erros seguen unha distribución normal. Este feito dá como resultado que, en casos onde non se verifiquen as hipóteses necesarias, estes non sexan idóneos para captar as relacións existentes entre as distintas variables. Debido a esta eiva, é necesario desenvolver alternativas que nos permitan facer estimacións apropiadas. Un exemplo é a regresión modal non paramétrica.

A regresión modal non paramétrica baséase na procura da moda global, se é unimodal, ou das modas locais, se é multimodal, da función de densidade condicional. Para buscar estes máximos, é necesario estimar esta función mediante un método de regresión en media non paramétrica, concretamente, por medio de estimadores tipo núcleo. A forma de construír este modelo permítenos relaxar as hipóteses, facéndoo moi versátil. Por un lado, ao ser regresión non paramétrica, poden relaxarse aquelas referentes á forma da regresión. Por

outro lado, ao ser modal, e dado que a moda non ten por que ser unha característica única dunha distribución, entón permite non só ter funcións senón tamén funcións multivaluadas.

Ademais da súa gran versatilidade, con aplicacións tan variadas como a predición do deterioro cognitivo provocado polo Alzheimer, a análise do consumo eléctrico ou os patróns dos incendios forestais, este tipo de regresión ten outras virtudes. Entre elas, cabe resaltar a súa validez para construír modelos onde as observacións estean divididas en grupos, pero sen saber a que grupo pertence cada observación. Este tipo de regresión tamén é útil se os erros do modelo presentan unha distribución altamente asimétrica e que a transformación dos datos non reporta resultados adecuados.

Unha introdución histórica xunto coa descrición do modelo de regresión en media paramétrico preséntase no Capítulo 1. Esta introdución acompáñase das hipóteses requiridas para que se poida realizar inferencia sobre el. A continuación, móstranse varios exemplos, tres simulados e un real, onde non se verifican as hipóteses requiridas e o modelo construído non é adecuado. Así, motívase a necesidade de alternativas, como a regresión modal.

Baseándose principalmente en [Wand e Jones, 1995], no Capítulo 2 defínese a estimación tipo núcleo da densidade e da densidade condicional, que será un elemento fundamental na exposición da regresión modal. Ademais, móstranse varios criterios de erro para a estimación en media non paramétrica co fin de introducir varios métodos de selección da fiestra de suavizado para este tipo de estimación.

No Capítulo 3, comézase facendo unha breve alusión á regresión unimodal paramétrica. De seguido, introdúcese o algoritmo *mean-shift* e os criterios de erro definidos en [Chen et al., 2016] que se poden aplicar para escoller as fiestras adecuadas na estimación multimodal. Para isto, introdúcense dous métodos diferentes de selección de fiestra, recollidos en [Zhou e Huang, 2019], sendo un deles de validación cruzada e outro baseado na remostraxe. Na seguinte sección analízanse os resultados dun estudo de simulación, realizado polo autor, de cincocentas simulacións dos tres modelos no programa `R`. Nel recóllense os erros puntuais cometidos ao realizar unha estimación modal non paramétrica en tres puntos, empregando como criterio de selección de validación cruzada, obténdose así que este é adecuado para dous dos tres casos. Ademais, móstrase a estimación modal para a primeira simulación de cada modelo. Finalmente, no último apartado faise unha estimación modal do caso real empregando ambos os criterios e compáranse as bondades e defectos de cada un.

Capítulo 1

Introdución: contexto da regresión modal

Neste capítulo preténdese introducir o modelo de regresión en media como método estatístico. Mostrarase a súa posible clasificación, segundo se é linear ou non, e describirase algunha das propiedades que verifica baixo certas hipóteses concretas. Finalmente, ilustraranse con exemplos, simulados e reais, posibles casos onde o modelo de regresión en media obtido non verifica as hipóteses requiridas para aplicar os resultados clásicos da Inferencia Estatística.

Para isto, dividimos o capítulo en varias seccións. Na primeira faremos unha breve contextualización histórica e definiremos diferentes modelos de regresión en media. Na segunda sección mostraremos tres exemplos simulados e estimaremos un modelo de regresión paramétrico asociado. Ademais, comprobaremos que este modelo estimado non é adecuado. Finalmente, na seguinte sección, realizaremos o mesmo procedemento para datos reais mediante un exemplo obtido de [Einbeck e Tutz, 2006].

1.1. Introdución aos modelos de regresión en media

Os modelos de regresión pretenden establecer relacións entre unha ou varias variables explicativas, que denotaremos por X, e unha variable resposta, que denotaremos por Y. Para isto, a análise da regresión *estima* o valor que toma a variable resposta segundo os posibles valores das variables explicativas.

O termo regresión, acuñado por Francis Galton (1822-1911), fai referencia ao campo da bioloxía, concretamente ao feito de que a altura dos descendentes de individuos notablemente altos tende a regresar aos valores promedio da poboación. O seu nacemento asóciase á aparición da teoría de mínimos cadrados en Francia durante a Revolución Francesa. Es-

ta recoñéceselle a Adrien-Marie Legendre (1752-1833), publicándose a súa demostración formal por primeira vez na obra *Théorie analytique* de Laplace, tal e como se recolle en [Boyer, 1987]. Aínda así, é de xustiza mencionar que Gauss xa inventara e xustificara o método de mínimos cadrados con anterioridade.

A regresión en media trata de predicir o valor da variable resposta, asumindo que este é o valor da media condicionada, para cada valor do conxunto de variables explicativas, onde cada valor concreto é da forma $x \in D$, sendo D o soporte de X. Isto pode expresarse matematicamente como $m(x) = \mathbb{E}(Y|_{X=x}) \quad \forall x \in D$. Así, podemos identificar o modelo como $Y = m(X) + \varepsilon$, onde ε é o termo do erro. Cabe destacar que, para poder aplicar os resultados de inferencia clásicos, a esperanza condicional do erro ten que ser cero, $\mathbb{E}(\varepsilon|_{X=x}) = 0$, e os erros deben verificar as hipóteses de homocedasticidade, normalidade e independencia. Ademais, se o modelo construído verifica a hipótese de linearidade, diremos que é un modelo de regresión linear. Noutro caso, falaremos dun modelo de regresión non linear.

O modelo de regresión en media linear clásico, un caso concreto de modelo paramétrico, é da forma $Y_i = \beta_0 + \beta_1 X_{1\,i} + ... + \beta_{p-1} X_{p-1\,i} + \varepsilon_i$, onde $i \in \{1, ..., n\}$ indica o individuo concreto. O modelo de regresión en media non linear paramétrico baséase nas mesmas hipóteses, salvo a linearidade da función $m(x)$.

Procedemos agora a describir algunha das propiedades do estimador por mínimos cadrados do modelo de regresión en media linear. Unha demostración e análise máis detallada destas propiedades poden verse en [Cristóbal, 1992]. Así, baixo estas hipóteses, temos que o estimador dos parámetros da pendente coincide co de máxima verosimilitude[1], aquel que fai máis "verosímil" as observacións realizadas. Isto débese a que as hipóteses de independencia e normalidade dos erros implican que os valores que maximizan o logaritmo da verosimilitude, coincidentes cos que maximizan a verosimilitude, son os mesmos que minimizan a suma por mínimos cadrados.

Ademais, no modelo de regresión en media linear, e de verificárense as hipóteses mencionadas anteriormente, concretamente a relativa á normalidade dos erros, temos que os estimadores son insesgados. Un estimador dise insesgado cando a súa esperanza matemática coincide co valor do parámetro a estimar. Neste caso, xa que, se $\hat{\beta}$ é un estimador para o vector de parámetros β, tense:

$$\mathbb{E}(\hat{\beta}) = \beta \implies \text{Nesgo} = \mathbb{E}(\hat{\beta}) - \beta = 0.$$

Por outro lado, se engadimos a hipótese de non existir colinearidade entre as variables explicativas, temos que o estimador por mínimos cadrados é consistente, é dicir, que o seu

[1] Este concepto estúdase en profundidade na materia obrigatoria de *Inferencia Estatística* de 3º curso.

nesgo tende a cero se o tamaño da mostra tende a infinito.

Se denotamos por X a matriz de deseño do modelo, é dicir, aquela que verifica a ecuación $Y = X\beta + \varepsilon$, e σ^2 a varianza dos erros, tense que a matriz de varianzas-covarianzas do estimador é $\mathrm{Cov}(\hat{\beta}, \hat{\beta}) = \sigma^2(X'X)^{-1}$, onde $\mathrm{Cov}(Y, Y) = \sigma^2 I_n$, sendo I_n a matriz identidade $n \times n$. Deste xeito, o teorema de Gauss-Markov[2] afirma que, baixo estas hipóteses, para un modelo linear, o estimador obtido por mínimos cadrados ten, e é o único que ten, a varianza mínima para estimadores lineares sen nesgo.

Teorema 1.1 (de Gauss-Markov)**.** *Verifíquense as seguintes hipóteses:*
a) Os valores da variable resposta están xerados polo modelo linear $Y = X\beta + \varepsilon$.
b) Os erros son incorrelados.
c) Os erros son homocedásticos.
d) Os erros son independentes das variables explicativas.
e) As observacións non se obtiveron con erros de medida.
f) Consideramos estimadores centrados lineares.
g) Consideramos como estimador óptimo aquel que é centrado e ten mínima varianza.
Entón, o estimador obtido polo método de mínimos cadrados é óptimo.

Demostración. Supoñamos coñecido que os estimadores obtidos polo método de mínimos cadrados son centrados e lineares, vexamos que son óptimos. É dicir, probemos que teñen varianza mínima dentro do conxunto de estimadores lineares para unha observación arbitraria.

Denotando por C a matriz $(X'X)^{-1}X'$, temos que $\hat{\beta} = CY$. Consideremos outro estimador linear centrado arbitrario, $b = BY$, e comparemos as varianzas de ambos. Temos que $\mathbb{E}(b) = \mathbb{E}(BY) = \mathbb{E}(BX\beta + B\varepsilon) = BX\beta$, xa que $\mathbb{E}(\varepsilon) = 0$.

Sexa $A = B - C$, como consideramos un estimador centrado, $BX = Id_{p-1}$, tense que

$$Id_{p-1} = BX = (A + C)X = AX + CX = AX + (X'X)^{-1}X'X = AX + Id_{p-1}.$$

Logo, $AX = 0$.

Por outro lado, tense que

$$\begin{aligned} \mathrm{Var}(b) &= \mathbb{E}((b-\beta)(b-\beta)') = \mathbb{E}((BX\beta + B\varepsilon - \beta)') = \mathbb{E}(B\varepsilon \cdot \varepsilon' B') \\ &= \sigma^2 BB' = \sigma^2(A + C)(A + C)' = \sigma^2(AA' + CA' + AC' + CC'). \end{aligned} \tag{1.1}$$

Ademais, verifícase

$$CC' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$$

[2]Este resultado forma parte do contido da materia optativa de 4º curso *Modelos de Regresión e Análise Multivariante*.

e

$$CA' = (X'X)^{-1}X'A' = (X'X)^{-1}(AX)' = 0 = AC'.$$

Aplicando estas dúas igualdades na ecuación (1.1), temos que

$$\mathrm{Var}(b) = \sigma^2(X'X)^{-1} + \sigma^2 AA' = \mathrm{Var}(\hat{\beta}) + \sigma^2 AA'.$$

Como os elementos da diagonal de AA' son positivos, ao ser a suma dos cadrados dos elementos de cada fila, verifícase que a varianza do estimador b considerado é estritamente maior que a do estimador de mínimos cadrados. Entón, o estimador obtido polo método de mínimos cadrados é óptimo. □

Fronte aos modelos paramétricos, que requiren de hipóteses relativamente fortes, encóntrase o modelo de regresión non paramétrico. Neste caso, as únicas hipóteses que se teñen que cumprir son a suavidade da función m, que se traduce na súa continuidade e na súa diferenciabilidade, e que a media condicional do erro sexa nula,

$$Y = m(X) + \varepsilon, \quad \mathbb{E}(\varepsilon|_{X=x}) = 0.$$

Nótese que este relaxamento de hipóteses lévanos a poder tratar un maior número de situacións, pero perdemos a posibilidade de empregar métodos tan eficientes como os da regresión paramétrica.

1.2. Algúns exemplos ilustrativos

Procedemos a mostrar tres exemplos de simulacións, recollidas en [Zhou e Huang, 2019], onde un modelo de regresión en media non ten por que ser o adecuado para mostrar as relacións entre as variables explicativas e a variable resposta.

Modelo C1:

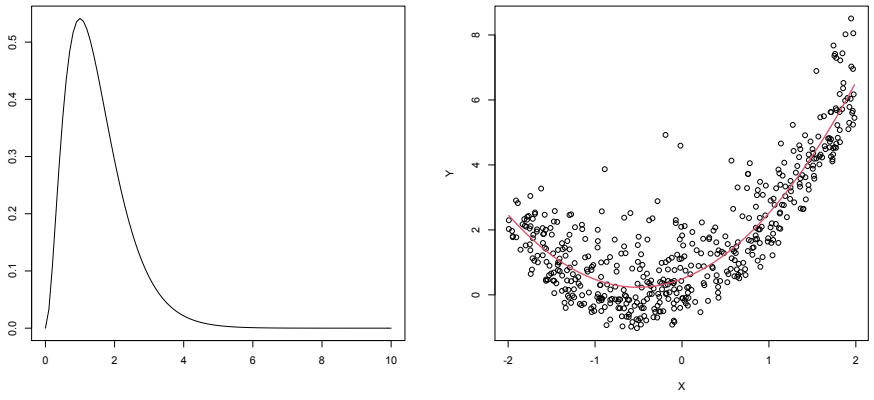
Este caso correspóndese co modelo:

$$Y = X + X^2 - 1 + \varepsilon, \quad \text{con } \varepsilon \sim \Gamma(\alpha, \beta), \tag{1.2}$$

onde o erro non é simétrico. Simuláronse, co programa `R`, cincocentos datos do intervalo $[-2, 2]$ seguindo unha distribución $X \sim U(-2, 2)$ e tomando como parámetros $\alpha = 3$ e $\beta = 2$. Represéntanse tanto a densidade dos erros como o conxunto de datos simulados na Figura 1.1.

	Estimación	Erro estándar	Valor t	$\Pr(>\|t\|)$
Intercepto	0,4935	0,0594	8,31	$9 \cdot 10^{-16}$
x	1,0205	0,0353	28,88	$< 2 \cdot 10^{-16}$
x^2	1,0032	0,0343	29,26	$< 2 \cdot 10^{-16}$

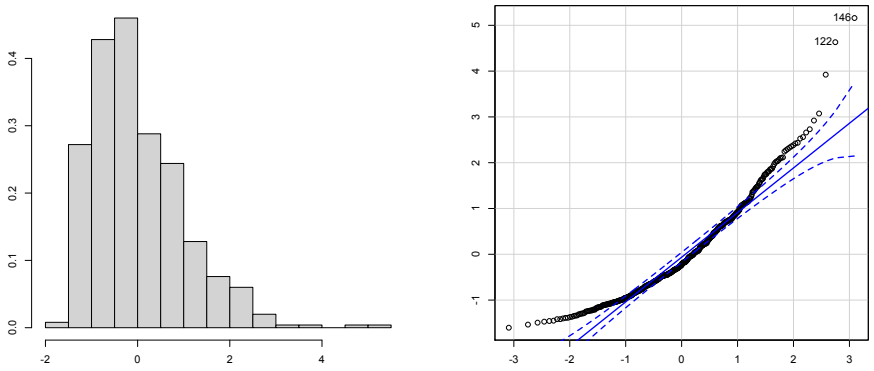
Cadro 1.1: Estimacións dos coeficientes, erro estándar, estatístico de contraste para H_0: $\beta = 0$ e p-valor asociado ao contraste do modelo polinómico de grao 2 de regresión en media para a simulación recollida na ecuación (1.2).



(a) (b)

Figura 1.1: Gráfica da función de densidade dos erros do modelo recollido na ecuación (1.2) e nube de puntos dos datos da simulación e modelo paramétrico axustado asociado do Cadro 1.1.

Visualmente, observando a Figura 1.1, parece que o modelo axustado, de regresión en media polinómico de grao 2, encaixa axeitadamente nos nosos datos. Ademais, a proporción da varianza explicada do modelo, o $R^2 - axustado$, é 0,788, o cal xeralmente sería bastante elevado. Non obstante, ao ser un modelo simulado, e tomar como modelo o baseado na nosa simulación, podemos considerar que non é o adecuado. Para comprobar isto, realizamos un test de normalidade aos residuos obtidos.



(a) (b)

Figura 1.2: Histograma dos residuos do modelo recollido no Cadro 1.1 e normalidade dos residuos do modelo recollido no Cadro 1.1.

Como se pode ver na Figura 1.2, os residuos do modelo non verifican a hipótese de normalidade, é dicir, non teñen unha distribución normal. Cabe destacar que no histograma se intúe a forma asimétrica positiva da distribución Gamma, lembremos que a súa función de densidade asociada é $f(x) = \frac{\lambda(\lambda x)^{\alpha-1}e-\lambda x}{\Gamma(\alpha)}$ para $x > 0$. Isto corrobórase ao aplicarlle un test Shapiro-Wilk[3] aos residuos do modelo axustado. O p-valor obtido é de $3{,}593 \cdot 10^{-15}$, o cal implica que os residuos do modelo non seguen unha distribución normal a calquera nivel de significación dos empregados habitualmente.

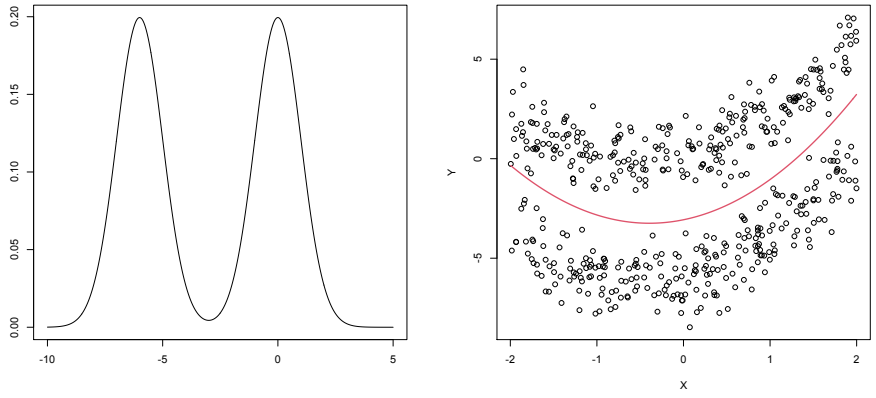
Relacionado co anterior, o p-valor de todos os coeficientes é o suficientemente pequeno para o contraste de significación, da orde de 10^{-16} ou inferior, como para poder afirmar que a un nivel de significación dos considerados habitualmente, todos os coeficientes serían significativos. Isto lévanos á conclusión de que si hai unha relación entre as variables, pero o modelo $\hat{y} = 0{,}4935 + 1{,}0205\, x + 1{,}0032\, x^2$, recollido no Cadro 1.1, non ten por que ser o adecuado. Cabe destacar o elevado valor do intercepto en comparación co verdadeiro, o cal tiña valor -1.

Modelo C2:

O modelo empregado neste caso é:

$$Y = X + X^2 + \varepsilon \quad \text{con} \;\; \varepsilon \sim 0{,}5 \cdot N(0,1) + 0{,}5 \cdot N(-6,1). \tag{1.3}$$

[3]Este test estúdase na materia obrigatoria de 3º curso de *Inferencia Estatística*.



(a) (b)

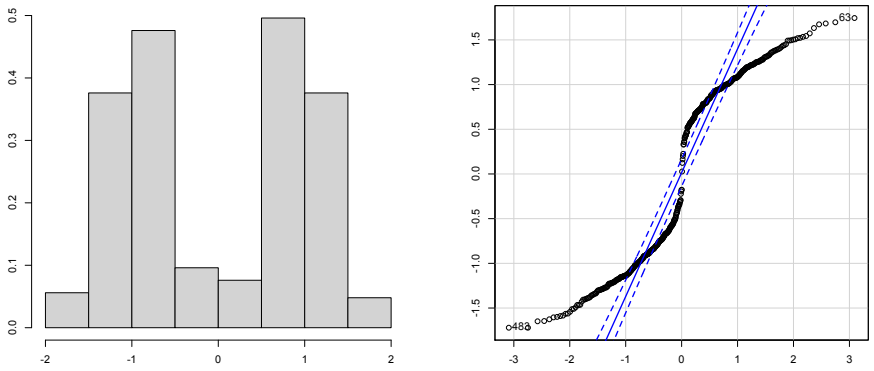
Figura 1.3: Gráfica da función de densidade dos erros do modelo recollido na ecuación (1.3) e nube de puntos dos datos da simulación e modelo paramétrico axustado asociado do Cadro 1.2.

A diferenza da simulación anterior, o erro si é simétrico. Porén, é multimodal. A simulación realizouse nas mesmas condicións que no modelo anterior, cincocentos datos obtidos dunha distribución $X \sim U(-2, 2)$. Ao igual que no caso anterior, axustouse un modelo polinómico de orde 2. Os resultados do axuste poden verse no Cadro 1.2.

	Estimación	Erro estándar	Valor t	$\Pr(>\|t\|)$
Intercepto	$-3{,}0680$	0,2129	$-14{,}41$	$< 2 \cdot 10^{-16}$
x	0,8911	0,1254	7,11	$4{,}13 \cdot 10^{-12}$
x^2	1,1305	0,1221	9,26	$< 2 \cdot 10^{-16}$

Cadro 1.2: Estimacións dos coeficientes, erro estándar, estatístico de contraste para H_0: $\beta = 0$ e p-valor asociado ao contraste do modelo de regresión en media para a simulación recollida na ecuación (1.3).

Observando a Figura 1.3, podemos supoñer que o modelo de regresión en media obtido, $\hat{y} = -3{,}0680 + 0{,}8911\, x + 1{,}1305\, x^2$, non se adecúa axeitadamente aos nosos datos. Isto podémolo corroborar neste caso co pequeno valor do $R^2 - axustado$, 0,2136. Cabe destacar que, visualmente, pola forma na que se distribúen os datos, pode intuírse que unha única curva non se amoldará axeitadamente a estes.



(a) (b)

Figura 1.4: Histograma dos residuos do modelo recollido no Cadro 1.2 e normalidade dos residuos do modelo recollido no Cadro 1.2.

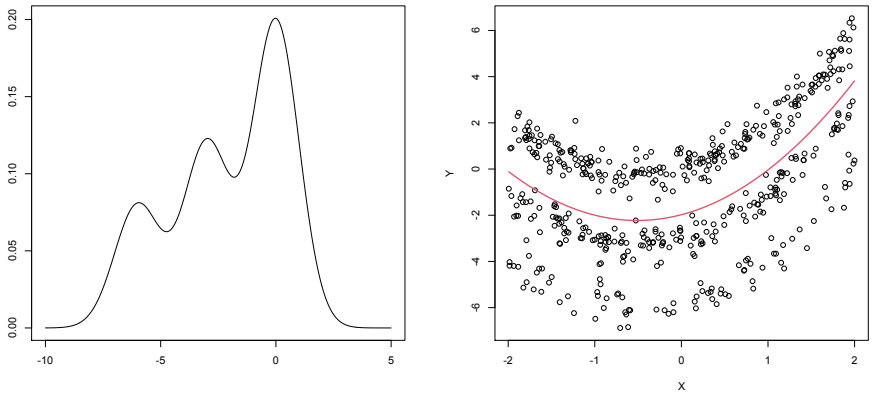
Na Figura 1.4 podemos ver que os residuos do modelo non verifican a hipótese de normalidade. Nótese que no histograma dos residuos se poden intuír as dúas modas, así como a distribución formada pola mestura de normais recollidas na ecuación (1.3).

Ao igual que no exemplo anterior, o p-valor de todos os coeficientes, recollido no Cadro 1.2, permítenos afirmar que, cun nivel de significación habitual, todos os coeficientes son significativos. A pesar disto, o feito de que non é unha distribución normal corrobórase ao realizar un test Shapiro-Wilk sobre os residuos estandarizados do modelo asociado, obtendo un p-valor inferior a $2{,}2 \cdot 10^{-16}$. En consecuencia, a distribución dos residuos do modelo non é normal e non podemos aplicar os resultados habituais de inferencia sobre o modelo.

Modelo C4:

Este modelo é unha extensión do caso anterior á situación onde o erro está xerado coa mestura de tres normais e con pesos diferentes. Así, o modelo empregado é:

$$Y = X + X^2 + \varepsilon \quad \text{con} \quad \varepsilon \sim 0{,}5 \cdot (0, 0{,}5^2) + 0{,}3 \cdot N(-3, 0{,}5^2) + 0{,}2 \cdot N(-6, 0{,}5^2). \tag{1.4}$$



(a) (b)

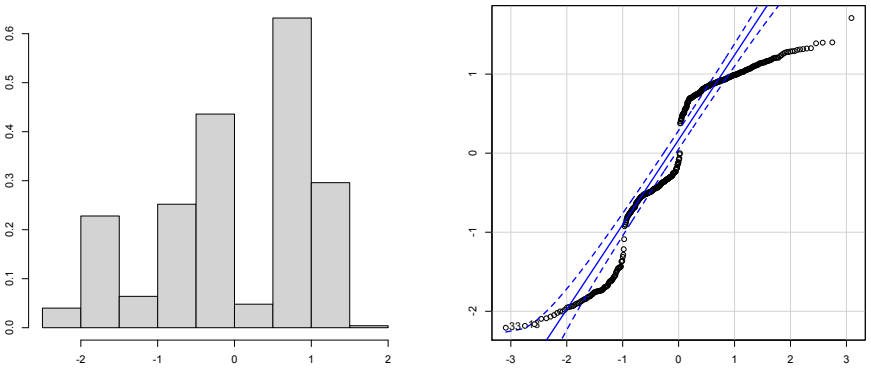
Figura 1.5: Gráfica da función de densidade dos erros do modelo recollido na ecuación (1.4) e nube de puntos dos datos da simulación e modelo paramétrico axustado asociado do Cadro 1.3.

Unha vez feita a simulación, ao igual que nos outros dous modelos, de cincocentos datos dunha distribución $X \sim U(-2, 2)$, obtemos os resultados mostrados no Cadro 1.3 ao axustar un modelo polinómico de orde 2.

	Estimación	Erro estándar	Valor t	$\Pr(>\|t\|)$
Intercepto	$-1{,}9767$	0,1512	$-13{,}07$	$< 2 \cdot 10^{-16}$
x	0,9793	0,0860	11,38	$< 2 \cdot 10^{-16}$
x^2	0,9594	0,0832	11,54	$< 2 \cdot 10^{-16}$

Cadro 1.3: Estimacións dos coeficientes, erro estándar, estatístico de contraste para H_0: $\beta = 0$ e p-valor asociado ao contraste do modelo de regresión en media para a simulación recollida na ecuación (1.4).

Ao seguir a mesma estrutura que o modelo C2, vemos que a análise que podemos facer do modelo é semellante. A estimación da regresión en media, recollida no Cadro 1.3, mediante o modelo $\hat{y} = 1{,}9767 + 0{,}9793\, x + 0{,}9594\, x^2$ non é adecuada, como se mostra visualmente na Figura 1.5. Isto vese confirmado polo baixo valor do $R^2 - axustado$, 0,3415. En consecuencia, podemos concluír que o modelo non explica a meirande parte das observacións simuladas.



(a) (b)

Figura 1.6: Histograma dos residuos do modelo recollido no Cadro 1.3 e normalidade dos residuos do modelo recollido no Cadro 1.3.

Ademais de non explicar a maior parte das observacións, e ao igual que sucedía nos dous modelos considerados anteriormente, os residuos non teñen unha distribución normal, como se pode comprobar na Figura 1.6. Ao observar o histograma dos residuos, nótase que este ten certo parecido coa mestura de normais correspondente á ecuación (1.4) así como a súa representación gráfica, recollida na Figura 1.5. Neste caso, e a diferenza do modelo C2, o feito de que na mestura das normais non se lle outorgue o mesmo peso a cada unha delas tradúcese nunha significativa diferenza de alturas das modas no histograma.

Para corroborar o intuído de forma visual, podemos realizar un test Shapiro-Wilk sobre os residuos do modelo. Neste caso, obtemos un p-valor inferior a $2{,}2 \cdot 10^{-16}$, polo que se corrobora que os residuos do modelo non verifican a hipótese de normalidade a ningún nivel de significación dos habituais. En consecuencia, ao igual que no modelo anterior, non podemos considerar válidos os resultados habituais de inferencia sobre este modelo.

1.3. Ilustración con datos reais

Procedemos a mostrar, de forma similar ao realizado cos modelos das simulacións, un exemplo con datos reais onde a regresión en media non é unha opción adecuada. Para isto, empregamos os datos de velocidade e fluxo de vehículos dun carril dunha autoestrada californiana, mostrados en [Einbeck e Tutz, 2006].

Os diagramas de velocidade e fluxo son empregados habitualmente na enxeñaría de tráfico. Neste caso concreto, mediuse a velocidade, en millas por hora, e o fluxo, en vehículos por carril e hora, en intervalos de trinta segundos. Así, a cuestión de interese é ser capaz de

explicar o diagrama de velocidade e fluxo destas observacións. Na Figura 1.7 pode intuírse que, cando o fluxo é baixo, non hai signo de asociación. Porén, un aumento do fluxo parece implicar unha diminución da velocidade.

Neste caso, como non coñecemos unha posible forma do modelo, e co fin de facer unha exploración maior sobre as observacións, comparamos distintos modelos polinómicos de grao 1, 2 e 3, onde a variable resposta é a velocidade e a variable explicativa é o fluxo de tráfico. Comparando estes modelos, mediante o Criterio de Información de Akaike (AIC), obtemos que o modelo de regresión linear simple é o que ten un menor AIC, entón é o máis adecuado dos tres empregados. A continuación, procedemos a ilustrar este modelo, cuxos valores se recollen no Cadro 1.4.

	Estimación	Erro estándar	Valor t	$\Pr(>\|t\|)$
Intercepto	57,3565	0,8061	71,15	$< 2 \cdot 10^{-16}$
x	$-0{,}0015$	0,0005	$-2{,}88$	0,0040

Cadro 1.4: Valores do modelo de regresión en media para o modelo asociado á velocidade e fluxo.

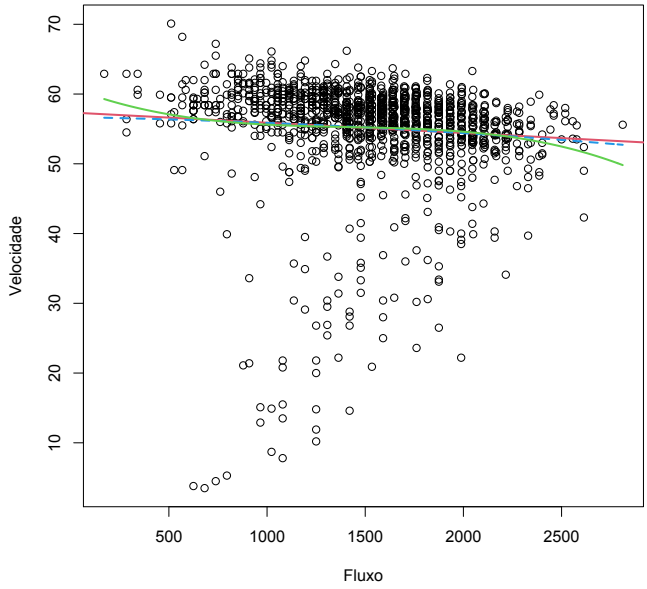


Figura 1.7: Nube de puntos dos datos da velocidade e fluxo cos posibles modelos polinómicos de grao 1 en vermello, de grao 2 en azul, e de grao 3 en verde.

Neste caso, tal e como se recolle no Cadro 1.4, o modelo paramétrico de regresión en media correspóndese con $\hat{y} = 57{,}3565 - 0{,}0015\,x$. Todos os estimadores son significativos e o $R^2 - axustado$ é 0,005526. Debido a isto, podemos afirmar que este modelo explica menos do un por cento da varianza. O valor baixo do $R^2 - axustado$ podía ser intuído visualmente, observando a Figura 1.7. Ademais, se realizamos un test Shapiro-Wilk sobre os residuos estandarizados, obtemos un p-valor inferior a $2{,}2 \cdot 10^{-16}$. Isto implica que os residuos non seguen unha distribución normal e os resultados habituais de inferencia non poden ser aplicados, xa que non se verifica a hipótese de normalidade dos residuos do modelo.

1.4. Conclusión e motivación do TFG

Deste capítulo, podemos concluír que a regresión en media non é sempre unha opción adecuada. Isto pode deberse a varios motivos, entre eles, como se mostra no modelo simulado C1, porque o erro do modelo sexa asimétrico, ou porque o erro non sexa normal, como pode ser o modelo simulado C2 ou o modelo C4. Ademais, esta falta de validez non sucede exclusivamente no plano teórico, senón que tamén se pode encontrar en datos recollidos na realidade, o cal mostramos cos datos de velocidade e fluxo de vehículos nunha autopista californiana na Sección 1.3.

Este feito lévanos a valorar a posibilidade de construír outros modelos, onde a medida empregada non sexa a media. Isto motiva a aparición, entre outras posibilidades, da regresión modal, que emprega a moda no canto da media para facer estimacións.

Capítulo 2

Revisión da estimación tipo núcleo

Unha vez vistos varios exemplos onde a regresión en media non ten por que ser unha opción axeitada, cabe preguntarse se hai outros métodos que poidamos empregar para facer estimacións. En concreto, faremos uso da estimación tipo núcleo da densidade (condicional) para introducir a regresión modal non paramétrica no seguinte capítulo. No que segue, presentaremos algúns conceptos básicos da estimación tipo núcleo da densidade e da densidade condicional que faciliten a comprensión do terceiro capítulo.

2.1. Estimación tipo núcleo da densidade

Os contidos desta sección están baseados en [Wand e Jones, 1995]. Comezamos considerando unha variable aleatoria X e unha mostra aleatoria simple $X_1, \dots, X_n$. Daquela, podemos definir o estimador tipo núcleo da densidade como unha función da forma

$$\hat{f}_n(x,h) = \frac{\sum_{i=1}^{n} K((x-X_i)/h)}{nh}, \tag{2.1}$$

onde $h \in \mathbb{R}$ é o ancho de banda e K é unha función, denominada función núcleo, que verifica $\int K(x)dx = 1$. Nótese que a construción do estimador da forma expresada na ecuación (2.1) permite estimar densidades que non pertencen a unha familia paramétrica. Ademais, o valor de h, o cal determina a escala do núcleo, decidirá a suavidade do estimador.

Aínda que a expresión máis habitual sexa a referida na ecuación (2.1), definindo a función de núcleo reescalada, $K_h(x) = \frac{K(x/h)}{h}$, podemos expresar o estimador tipo núcleo da densidade da seguinte maneira:

$$\hat{f}_n(x,h) = \frac{\sum_{i=1}^{n} K_h(x-X_i)}{n}.$$

Cabe destacar que, xeralmente, a función núcleo é unha función densidade de probabilidade e é simétrica con respecto ao cero. Neste caso, se K é a densidade dunha variable

aleatoria Z, a función K_h é a densidade da variable $h\,Z$. Como exemplos de funcións núcleo habituais podemos mencionar a Gaussiana, $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, a Triangular, $K(x) = (1-|x|)\mathbb{I}(|x|<1)$, ou a Epanechnikov, $K(x) = \frac{3}{4}(1-x^2)\mathbb{I}(|x|<1)$. Ademais, tense que o estimador resultante herdará as propiedades da función núcleo de densidade empregada na súa obtención.

No caso da estimación da densidade tipo núcleo, os criterios de erro empregados para analizar o comportamento do estimador poden medir o erro nun único punto, os criterios de erro puntuais, ou medir o erro cometido na estimación ao longo de toda a función $\hat{f}_n$, os criterios de erro globais. Un exemplo de criterio de erro puntual para un punto x é o erro cadrático medio, ECM, que se define como

$$\mathrm{ECM}(\hat{f}_n(x,h)) = \mathbb{E}(\hat{f}_n(x,h) - f(x)^2), \tag{2.2}$$

e un exemplo de criterio de erro global é o erro cadrático integrado, ECI, definido como

$$\mathrm{ECI}(\hat{f}_n(\cdot,h)) = \int (\hat{f}_n(x,h) - f(x))^2 dx. \tag{2.3}$$

A partir do erro cadrático integrado[1], o cal coincide con $\|\hat{f}_n - f\|_2^2$, onde $\|\cdot\|_2$ é a norma asociado ao espazo $\mathcal{L}^2$, podemos definir o erro cadrático medio integrado como

$$\mathrm{ECMI}(\hat{f}_n(\cdot,h)) = \mathbb{E}(\mathrm{ECI}(\hat{f}_n(\cdot,h))). \tag{2.4}$$

Ademais, ao combinarmos as ecuacións (2.2) e (2.4), verifícase a seguinte relación entre o erro cadrático medio e o erro cadrático medio integrado:

$$\begin{aligned}\mathrm{ECMI}(\hat{f}_n(\cdot,h)) &= \mathbb{E}\left(\int (\hat{f}_n(x,h) - f(x))^2 dx\right) \\ &= \int \mathbb{E}(\hat{f}_n(x,h) - f(x))^2 dx = \int \mathrm{ECM}(\hat{f}_n(x,h))dx.\end{aligned} \tag{2.5}$$

Fixando un x, obtemos que o valor esperado do estimador tipo núcleo da densidade é da forma

$$\mathbb{E}(\hat{f}_n(x)) = \int K(u) f(x-hu)du = (K_h * f)(x)^2. \tag{2.6}$$

Repárese que na ecuación (2.6) se introduce a convolución, denotada por $*$. Dadas dúas funcións $f, g : \mathbb{R} \to \mathbb{R}$, definimos a súa convolución como

$$(f * g)(x) = \int_{\mathbb{R}} f(y)g(x-y)\,dy.$$

[1] Estas nocións estúdanse de forma introdutoria na materia obrigatoria de *Series de Fourier e Introdución ás Ecuacións en Derivadas Parciais* de 3º curso e de forma máis extensa na materia optativa de *Análise Funcional en Espazos de Hilbert* de 4º curso.

[2] O concepto de convolución foi definido e estudado na materia de *Análise Funcional en Espazos de Hilbert*.

Nótese que $(f * g)(x) = (g * f)(x)$, ao aplicar na definición de convolución o cambio de variable $z = x - y$. Ademais, de ser derivable f, e existiren as funcións $f * g$ e $f' * g$, temos que $f * g$ é derivable e $(f * g)' = f' * g$. Así, no noso caso concreto, temos que $(K_h * f)' = K_h' * f$. Esta propiedade, a cal é aplicable a derivadas de orde arbitraria, recolle a idea intuitiva da regularidade que poden acadar as estimacións tipo núcleo da densidade.

Co fin de facilitar a comprensión, procederemos a introducir varias notacións. Para unha función cadrado-integrable, g, denotamos $R(g) = \int g(u)^2\,du$. Para unha función K, denotamos $\mu_2(K) = \int u^2 K(u)\,du$.

Empregando o Teorema de Taylor e supoñendo que a función f que pretendemos estimar é o suficientemente regular, dúas veces diferenciable, obtemos a seguinte igualdade:

$$f(x - hu) = f(x) - huf'(x) + \frac{1}{2}(hu)^2 f''(x) + o(h^2). \tag{2.7}$$

De verificarse que $\mu_2(K) < \infty$, e se K é unha función de densidade simétrica con respecto ao cero, podemos aplicar a igualdade recollida na ecuación (2.7) na integral da ecuación (2.6). Así, baixo estas hipóteses, como $\int uK(u)\,du = 0$, obtemos que se verifica a igualdade

$$\mathbb{E}(\hat{f}_n(x)) = f(x) + \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2).$$

Logo, tense que:

$$h \to 0 \implies \mathbb{E}(\hat{f}_n(x)) - f(x) = \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2). \tag{2.8}$$

No caso da varianza, temos que se se verifica $nh \to \infty$, entón

$$\begin{aligned}
\mathrm{Var}(\hat{f}_n(x)) &= \frac{1}{n}\mathrm{Var}(K_h(x - X_1)) \\
&= \frac{1}{n}\left[\int K_h^2(x-y)f(y)\,dy - \left(\int K_h(x-y)f(y)\,dy\right)^2\right] \\
&= \frac{1}{nh}\int K^2(u)f(x-hu)\,du - \frac{1}{n}\left(\int K(u)f(x-hu)\,du\right)^2 \\
&= \frac{1}{nh}\int K^2(u)f(x+o(1))\,du - \frac{1}{n}(f(x)+o(1))^2 = \frac{1}{nh}R(K)\,f(x) + o((nh)^{-1}).
\end{aligned} \tag{2.9}$$

Se nos fixamos nas ecuacións (2.8) e (2.9), podemos intuír que o nesgo aumentará canto máis grande sexa h e a variabilidade aumentará canto máis pequeno sexa h. Este feito implica que a elección dun h adecuado sexa crucial para reducir o posible erro. Así, para certo x e h fixados, temos que o erro cadrático medio se corresponde con

$$\begin{aligned}
\mathrm{ECM}(\hat{f}_n(x,h)) &= \mathrm{Var}(\hat{f}_n(x,h)) + (\mathbb{E}(\hat{f}_n(x,h)) - f(x))^2 \\
&= \frac{1}{nh}R(K)\,f(x) + \left(\frac{1}{2}h^2\mu_2(K)f''(x)\right)^2 + o((nh)^{-1} + h^4).
\end{aligned} \tag{2.10}$$

Desta maneira, aplicando a relación entre o erro cadrático medio e o erro cadrático medio integrado, recollida na ecuación (2.5), podemos obter o erro cadrático medio integrado ao integrar na ecuación (2.10),

$$\mathrm{ECMI}(\hat{f}_n(\cdot,h)) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') + o((nh)^{-1} + h^4).$$

Así, e se obviamos na ecuación anterior a parte relativa á orde, podemos definir o erro cadrático medio integrado asintótico como

$$\mathrm{ECMIA}(\hat{f}_n(\cdot,h)) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f''). \tag{2.11}$$

A partir destes resultados, podemos definir o ECMIA como unha función dependente de h, ECMIA $= \psi(h)$ e, ao resolver a ecuación $\psi'(h) = 0$, obtemos unha fiestra óptima para o erro cadrático medio integrado asintótico. Ademais, podemos garantir a existencia da solución ao ter que

$$\lim_{h\to 0^+} \psi(h) = \lim_{h\to 0^+} \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') = \lim_{h\to 0^+} \frac{1}{nh}R(K) = +\infty$$

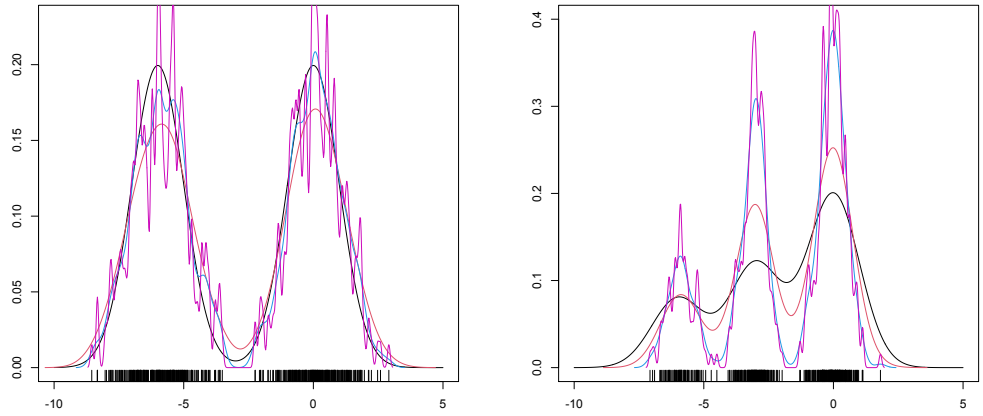
e

$$\lim_{h\to \infty} \psi(h) = \lim_{h\to \infty} \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') = \lim_{h\to \infty} \frac{1}{4}h^4\mu_2(K)^2R(f'') = +\infty.$$

Esta fiestra é da seguinte forma:

$$h_{\mathrm{ECMIA}} = \left(\frac{R(K)}{\mu_2(K)^2R(f'')n}\right)^{1/5}. \tag{2.12}$$

Procedamos a ilustrar a importancia do ancho de banda escollido baseándonos nas simulacións recollidas na Sección 1.3. Para isto, vexamos varias estimacións tipo núcleo da densidade dos erros dos modelos C2 e C4, variando a fiestra empregada.



(a) Densidade dos erros do modelo asociado á ecuación (1.3)

(b) Densidade dos erros do modelo asociado á ecuación (1.4)

Figura 2.1: Densidade real dos erros, en negro, e estimacións tipo núcleo con $h = 0{,}05$ en maxenta, $h = 0{,}2$ en azul, e $h = 0{,}6$ en vermello.

Como podemos ver na Figura 2.1, o ancho de banda escollido é crucial para obter unha boa aproximación, o cal se pode observar de forma especialmente clara na gráfica dos erros do modelo C4. A estimación realizada con ancho de banda $h = 0{,}2$, en azul, aproxímase bastante máis á curva real que a estimación realizada con ancho de banda $h = 0{,}6$, en vermello. Así e todo, temos que a relativa ao ancho de banda $h = 0{,}05$, en maxenta, é moito máis variable, axustándose de forma máis irregular ao modelo ideal. Por outro lado, na gráfica correspondente á simulación C2, pódese intuír o maior parecido da función real por parte da estimación con ancho de banda $h = 0{,}6$, en vermello. Non obstante, tamén é moito menos suave e ten unha maior variabilidade que a estimación con ancho de banda $h = 0{,}2$, en azul. Neste caso, a estimación realizada con ancho de banda $h = 0{,}05$, en maxenta, ao igual que no modelo C4, ten unha variabilidade moito maior.

Unha posible explicación da relación entre o tamaño da fiestra empregada coa suavidade da estimación e co seu axuste á curva real pode ser a relación do parámetro h e as expresións do nesgo e da varianza . O nesgo aumenta cando aumenta o tamaño da fiestra. Debido a isto, valores de h máis pequenos dan estimacións máis centradas. Pero, como unha diminución no valor de h implica un aumento da varianza, ao pretender ter estimacións máis centradas, obtemos que estas teñen unha variabilidade maior.

Por outro lado, podemos preguntarnos se é posible aplicar esta estimación da densidade para dimensións maiores, mediante vectores aleatorios. Se X é un vector aleatorio

d-dimensional e $X_1, \dots, X_n$ unha mostra aleatoria simple, de tal forma que $D \subseteq \mathbb{R}^d$ é un vector aleatorio de dimensión d, podemos definir o estimador tipo núcleo como

$$\hat{f}_{n,K,H}(x) = \frac{1}{n|H|^{1/2}} \sum_{i=1}^{n} K_d(H^{-1/2}(x - X_i)), \tag{2.13}$$

onde $H \in M_{d\times d}(\mathbb{R})$ é unha matriz de fiestras, $H^{-1/2}$ a inversa da matriz asociada á descomposición de Cholesky de H e K unha función núcleo. Neste caso, K pode obterse como unha densidade d-dimensional ou como produto de densidades unidimensionais en cada compoñente. Nótese que podemos realizar a descomposición de Cholesky de H por ser esta matriz hermitiana e definida positiva.

Cabe resaltar a gran similitude desta definición coa do caso unidimensional, recollida na ecuación (2.1), xa que basta substituír h por $|H|^{1/2}$ e $(x - X)$ por $(x - X_i)$. Ao igual que no caso 1-dimensional, adoita considerarse a función K_d como unha función densidade d-dimensional, ou ben obterse como produto de densidades unidimensionais en cada compoñente. De todas as maneiras, o feito de que a matriz de fiestras H teña $\frac{d(d+1)}{2}$ elementos distintos dificulta enormemente o cálculo da matriz de fiestras óptima conforme engadimos dimensións, á vez que se precisan máis datos para obter unha precisión semellante á do caso unidimensional, fenómeno que se coñece como a maldición da dimensionalidade.

2.2. Estimación tipo núcleo da densidade condicional

Se temos dúas variables aleatorias X e Y, podemos definir a densidade marxinal da variable X a partir da densidade conxunta de (X, Y) como a función

$$f^M(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy \quad \text{con } x \in D.$$

Ademais, se S é o soporte da variable Y, podemos definir a densidade condicional da variable Y condicionada a $x \in D$ como

$$f(y|x) = \frac{f(x, y)}{f(y)}, \quad \text{con } y \in S. \tag{2.14}$$

Estas definicións poden xeneralizarse ao caso multivariante. Así, de termos un vector aleatorio $X = (X_1, \dots, X_d)$ de variables aleatorias, podemos definir, dado $x \in (X_1, \dots, X_m)$ con $m < d$, a función densidade marxinal como

$$f^M(x_1, \dots, x_m) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \dots, x_m, x_{m+1}, \dots, x_d) \prod_{i=m+1}^{d} dx_i$$

E unha vez definida a función densidade marxinal, a partir da ecuación (2.14), podemos definir a función densidade condicional de varias variables como

$$f(x_1, \dots x_m | x_{m+1}, \dots x_d) = \frac{f(x_1, \dots, x_d)}{f^M(x_{m+1}, \dots, x_d)}. \tag{2.15}$$

A ecuación (2.14) garda certo parecido coa definición de probabilidade condicionada, pois non deixa de ser a súa xeneralización para funcións de densidade. Neste caso, temos unha igualdade análoga á proporcionada polo teorema de Bayes[3]:

$$f(y|x) = \frac{f(x|y) f^M(y)}{f^M(x)}.$$

Vexamos agora como poderiamos construír estimadores tipo núcleo para a densidade condicional, supoñendo que queremos coñecer a densidade da variable aleatoria Y, definida en $\mathbb{R}$ condicionada a $X = x$, onde X é unha variable aleatoria definida en $\mathbb{R}^d$. Para isto, basearémonos nas propostas recollidas en [Hyndman et al, 1996].

Sexa $f(x, y)$ a densidade conxunta das nosas variables e $f^M(x)$ a densidade marxinal de X, se substituímos na ecuación (2.14), a densidade condicional que pretendemos estimar verifica a igualdade

$$f(y|x) = \frac{f(x, y)}{f^M(x)}.$$

Co fin de facilitar a comprensión, consideremos o caso univariante, aínda que se pode xeneralizar ao caso multivariante, intercambiando h^{-1} por $H^{-1/2}$. Se temos unha mostra aleatoria simple $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, cuxas observacións supoñemos independentes, definimos a media condicional como $m(x) = \mathbb{E}(Y|X = x)$, aplicando a igualdade anterior, temos que un posible estimador é

$$\hat{f}_n(y|x) = \frac{\hat{f}_n(x, y)}{\hat{f}_n^M(x)}. \tag{2.16}$$

Neste caso, podemos aplicar a estimación tipo núcleo definida na ecuación (2.13) para obter o denominador da fracción do segundo membro da ecuación (2.16).

$$\hat{f}_n^M(x) = \frac{1}{n h_x} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_x}\right).$$

Igualmente, podemos considerar como estimador núcleo multivariante o núcleo produto, da seguinte maneira:

$$\hat{f}_n(x, y) = \frac{1}{n h_x h_y} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_x}\right) K\left(\frac{y - Y_j}{h_y}\right),$$

[3]Este teorema forma parte dos contidos da materia de formación básica de 1º curso *Elementos de Probabilidade e Estatística*.

onde h_x e h_y serían as fiestras correspondentes para as variables X e Y, respectivamente, é dicir, controlan a suavidade da estimación.

Se substituímos na ecuación (2.16), obtemos que o estimador que construímos é

$$\hat{f}_n(y|x) = \frac{1}{h_y} \sum_{j=1}^{n} w_j(x) K\left(\frac{y - Y_j}{h_y}\right), \tag{2.17}$$

onde

$$w_j(x) = \frac{K\left(\frac{x-X_j}{h_x}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_x}\right)}, \quad j = 1, \dots, n.$$

Vexamos agora as propiedades asintóticas deste estimador. Para isto, supoñeremos que $h_x \to 0$, $h_y \to 0$ e $n \to \infty$. Ademais, consideremos que as funcións $f(y|x), f^M(x)$ sexan funcións $\mathcal{L}^2$ e o suficientemente regulares. Se a función $m(x)$ tamén é o suficientemente regular, tal e como se recolle en [Hyndman et al, 1996], podemos expresar o nesgo mediante a igualdade

$$\begin{aligned} \mathbb{E}(\hat{f}_n(y|x)) - f(y|x) = \frac{h_x^2 \sigma_K^2}{2} \left(2 \frac{(f^M)'(x)}{f^M(x)} \frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2} + \frac{h_y^2}{h_x^2} \frac{\partial^2 f(y|x)}{\partial y^2} \right) \\ + O(h_x^4) + O(h_y^4) + O(h_x^2 h_y^2) + O\left(\frac{1}{n h_x}\right), \end{aligned} \tag{2.18}$$

e a varianza mediante a igualdade

$$\mathrm{Var}(\hat{f}_n(y|x)) = \frac{R(K) f(y|x)}{n h_x h_y f^M(x)} \left[R(K) - h_y f(y|x)\right] + O\left(\frac{1}{n}\right) + O\left(\frac{h_y}{h_x n}\right) + O\left(\frac{h_x}{h_y n}\right). \tag{2.19}$$

Logo, se sumamos o cadrado da ecuación (2.18) e a ecuación (2.19), obtemos unha nova medida do erro, o erro cadrático medio asintótico. Este defínese da seguinte maneira:

$$\begin{aligned} \mathrm{ECMA}(\hat{f}_n(y|x)) = \frac{h_x^4 \sigma_K^4}{4} \left(2 \frac{(f^M)'(x)}{f^M(x)} \frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2} + \frac{h_y^2}{h_x^2} \frac{\partial^2 f(y|x)}{\partial y^2} \right)^2 \\ + \frac{R(K) f(y|x)}{n h_x h_y f^M(x)} \left[R(K) - b f(y|x)\right] + O\left(\frac{1}{n}\right) + O\left(\frac{h_y}{h_x n}\right) \\ + O\left(\frac{h_x}{h_y n}\right) + O(h_x^6) + O(h_y^6) + O(h_x^2 h_y^4) + O(h_x^4 h_y^2) + O\left(\frac{1}{n h_x}\right), \end{aligned}$$

o cal é consistente se ás hipóteses mencionadas anteriormente se lles engade que, cando $n \to \infty$, se verifique $n h_x h_y \to \infty$. Podemos mencionar que, ao igual que na estimación tipo núcleo da densidade, canto máis grande sexa a fiestra seleccionada, o nesgo é máis grande e a varianza máis pequena.

2.3. Selección da fiestra de suavizado

Na Sección 2.1 mostramos diferentes criterios de erro para a estimación tipo núcleo. Ademais, ilustramos a importancia de escoller un ancho de banda óptimo e chegamos a un ancho de banda óptimo, de forma teórica, para o ECMIA:

$$h_{\text{ECMIA}} = \left(\frac{R(K)}{\mu_2(K)^2 R(f'')n} \right)^{1/5}.$$

Porén, na práctica non podemos calculalo, xa que precisariamos coñecer a segunda derivada da propia función que pretendemos estimar. Polo tanto, é necesario empregar outros métodos que, dependendo do criterio de erro empregado, nos darán un bo ancho de banda co cal facer estimación tipo núcleo. Procedemos a introducir varios destes métodos.

Métodos de tipo Plug-In:

Na expresión da fiestra óptima para o ECMIA o que se descoñece é $R(f'')$. Entón, coñecendo este termo, ou unha aproximación del, poderemos obter o ECMIA, ou unha aproximación deste. Estes métodos baséanse en asumir que e $R(f'')$ é dunha forma concreta, para obter o h óptimo nese caso. Deste tipo, introducimos a regra do polgar e a regra de Sheather e Jones, recollidas en [Wand e Jones, 1995].

A regra do polgar, ou de Silverman, baséase en substituír a integral descoñecida pola integral correspondente dunha densidade $N(0, \sigma)$. Así, temos que

$$R(f'') = \frac{3}{8\sqrt{\pi}} \sigma^{-5},$$

e o ancho de banda correspóndese con:

$$h_{\text{ECMIA}} = \left(\frac{8\sqrt{\pi} R(K)}{3\mu_2(K)^2 n} \right)^{1/5} \hat{\sigma},$$

onde $\hat{\sigma}$ é unha estimación de σ. Destacamos que este método tende a sobresuavizar a función que pretendemos estimar.

A regra de Sheather e Jones é un método iterativo que consiste en estimar de forma non paramétrica $R(f'')$. Se denotamos por $\Psi_r = \mathbb{E}(f^{(r)}(X))$, integrando por partes obtemos que $R(f'') = \Psi_4$. De supoñermos que a función que pretendemos estimar é normal con desviación típica σ, podemos asumir que

$$\Psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1} (r/2)! \sqrt{\pi}}.$$

Mediante un proceso iterativo *backward*, estimando da seguinte maneira:

$$\hat{f}_n^r(X_i) = \frac{1}{n} \sum_{j=1}^{n} L_h^{(r)}(X_i - X_j),$$

onde h é unha fiestra que pode ser estimada ao coñecer un posible estimador de $\hat{\Psi}_{r+2}$ e L é unha función núcleo. Cabe mencionar que, pese á suposición da normalidade de f, necesaria para estimar $\hat{\Psi}_{r_0}$, e da regularidade de f, na práctica este método dá resultados xeralmente satisfactorios.

Aproximación de erros:

Este tipo de métodos fundaméntanse en facer aproximacións para seleccionar fiestras óptimas, no canto de basearse en posibles estimacións das expresións de fiestras óptimas. Destacamos o método de validación cruzada.

O método de validación cruzada consiste en aproximar as medidas do erro a partir dunha mostra, calculando os erros cometidos para cada dato ao facer a estimación sobre a mostra considerada obviando ese dato concreto. Así, a partir do erro integrado, expresado na ecuación (2.3), e para un h fixado, obtemos a seguinte igualdade:

$$\begin{aligned} \mathrm{ECI}(h) &= \int \hat{f}_n^2(x)\,dx - 2\int \hat{f}_n(x)\,f(x)\,dx + \int f^2(x)\,dx \\ &= R(\hat{f}_n) - 2\int \hat{f}_n(x)\,f(x)\,dx + R(f). \end{aligned} \tag{2.20}$$

Nótese que na ecuación (2.20), $R(f)$ é independente do h seleccionado. Logo, ao intentar obter a fiestra óptima, podemos omitilo. Desta maneira, a fiestra óptima é:

$$h_{\mathrm{ECI}} = \arg\,\mín_h \left(R(\hat{f}_n) - 2\int \hat{f}_n(x)\,f(x)\,dx \right). \tag{2.21}$$

Nótese que temos as seguintes igualdades:

$$R(\hat{f}_n) = \frac{1}{n^2 h}\sum_{i,j} K * K\left(\frac{X_i - X_j}{h}\right) \tag{2.22}$$

e

$$\int \hat{f}_n(x)\,f(x)\,dx = \mathbb{E}(\hat{f}_n(x)). \tag{2.23}$$

Ao non dispoñermos doutra mostra $\tilde{X}_1, \ldots, \tilde{X}_m$, independente de $X_1, \ldots, X_n$, que nos permita aproximar a cantidade da ecuación (2.23) por

$$\frac{1}{m}\sum_{i=1}^{m} \hat{f}_m(\tilde{X}_i),$$

podemos estimar o valor de $f(X_i)$ empregando a mostra sen o dato i-ésimo. Así, se $\hat{f}_n^{-i}$ denota o estimador núcleo obtido a partir da mostra salvo o dato i-ésimo, tense que podemos realizar a estimación

$$\hat{\mathbb{E}}(\hat{f}_n(x)) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_n^{-i}(X_i).$$

Desta maneira, obtemos a función

$$CV(h) = \frac{1}{n^2 h} \sum_{i,j} K * K \left(\frac{X_i - X_j}{h} \right) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_n^{-i}(X_i). \tag{2.24}$$

Entón, a fiestra óptima para este método é o h que minimiza a función CV. Cabe mencionar que as problemáticas das fiestras obtidas por este método son unha alta variabilidade e unha tendencia a infrasuavizar a función a estimar.

Finalmente, cómpre lembrar que tanto os criterios de erro como os selectores da fiestra mostrados non son os únicos existentes. Por exemplo, outro posible criterio de erro é o erro absoluto medio, que se define como $\mathbb{E}(|\hat{f}(x,h) - f(x)|)$, e outro método de selección de fiestra é o *Bootstrap*, baseado en técnicas de remostraxe. Así e todo, os descritos nesta sección son varios dos principais e máis empregados.

Capítulo 3

Regresión modal non paramétrica

Como vimos no primeiro capítulo, a regresión en media non é sempre unha boa opción para o modelado da relación entre dúas variables, xa que require de hipóteses que non sempre se verifican. Porén, existen métodos alternativos ao da regresión en media, por exemplo, a regresión modal. Neste capítulo, introduciremos este concepto e profundizaremos no algoritmo *mean-shift*, un método de estimación baseado na regresión en moda.

3.1. Introdución á regresión modal

A moda dunha distribución é o valor ou valores nos cales a función densidade de probabilidade acada un máximo local. Desta maneira, supoñamos que queremos estudar a relación entre unha variable resposta Y e unha variable explicativa X mediante regresión modal. A diferenza doutras medidas de tendencia central, como poden ser a media ou a mediana, pode suceder que a moda non tome un único valor. A causa disto, é necesario diferenciar entre buscar unha única moda condicional global, o que recibe o nome de regresión unimodal, ou buscar as modas locais condicionais, o que se coñece como regresión multimodal.

Se $f(x, y)$ é a densidade conxunta das nosas variables e esta é o suficientemente regular, dúas veces continuamente diferenciable, podemos definir as modas condicionais locais nun punto x como

$$M(x) := \left\{ y;\ \frac{\partial}{\partial y} f(y|x) = 0,\ \frac{\partial^2}{\partial y^2} f(y|x) < 0 \right\}. \tag{3.1}$$

Notemos que, a diferenza dos métodos de regresión considerados anteriormente, neste caso $M : D \to \mathcal{P}(Y)$, onde D é o soporte de X e $\mathcal{P}(Y)$ denota partes de Y. Así, o rango da función son subconxuntos de Y, polo que podemos ter, para certo x dado, máis dun elemento de Y contido na súa imaxe. O feito de considerar que Y pode ser unimodal ou

multimodal xogará un papel crucial na estimación que realicemos.

Por outro lado, aínda que na definición dada na ecuación (3.1) empreguemos a densidade condicional de Y para o x dado, podemos ter unha expresión equivalente empregando unicamente a densidade conxunta:

$$M(x) = \left\{ y; \frac{\partial}{\partial y} f(x,y) = 0, \frac{\partial^2}{\partial y^2} f(x,y) < 0 \right\}. \tag{3.2}$$

Cabe destacar que, tal e como se recolle en [Chen, 2018], as estimacións unimodais e multimodais teñen vantaxes e desvantaxes diferentes. Por un lado, o estimador construído no caso unimodal adoita ter unha expresión máis simple e facilmente interpretable. Por outro lado, a regresión multimodal detecta mellor as relacións entre a variable explicativa e a variable resposta. Isto nótase especialmente cando hai varias modas, como poden ser os modelos C2 e C4 do capítulo inicial, ou cando a relación entre a variable explicativa e a resposta ten unha expresión complicada. Desta maneira, na estimación multimodal, ao considerar modas locais, as rexións de predición tenden a ser considerablemente menores, a cambio de ter funcións máis dificilmente interpretables e de carácter máis complexo.

Unha vez introducidos estes conceptos, procedamos a construír o modelo linear da regresión modal. Este modelo é un caso concreto de regresión unimodal, é dicir, asumimos que só existe unha única moda. Tanto o modelo como as súas propiedades veñen recollidas en [Lee, 1989].

Supoñamos que o soporte de Y sexa real, $S \subseteq \mathbb{R}$, e que a nosa variable explicativa X ten soporte compacto $D \subset \mathbb{R}$. Deste xeito, e analogamente ao correspondente da regresión en media, o noso modelo pode construírse como

$$\text{Moda}(Y|X = x) = \beta_0 + \beta x, \tag{3.3}$$

onde $\beta_0 \in \mathbb{R}$ e $\beta \in \mathbb{R}$ son os parámetros a estimar e $\text{Moda}(Y|X = x)$ denota a moda global de Y para un certo x dado.

Sen entrar en detalle, pódese mencionar que este modelo se pode xeneralizar para máis dunha variable explicativa con soporte compacto $D' \subset \mathbb{R}^d$. Para isto, considérase o modelo

$$\text{Moda}(Y|X = x) = \beta_0 + \beta' x, \quad \text{con } x \in D',$$

onde X denota o conxunto de variables explicativas.

Con todo, e ao igual que para a regresión en media, o modelo paramétrico é excesivamente ríxido e non sempre é unha opción axeitada. Por exemplo, pódese intuír que no caso do modelo C2 simulado, recollido na ecuación (1.3), non será adecuado. Aínda que existan estimadores non paramétricos para a regresión unimodal, os cales nos permitirían relaxar

algunha hipótese, continuariamos asumindo a existencia dunha única moda. Co fin de non asumir esta hipótese, centrarémonos no problema da regresión multimodal.

Para intentar dar unha solución axeitada a este problema, podemos definir un modelo de regresión modal non paramétrico considerando os estimadores tipo núcleo. Así, substituíndo a densidade conxunta mediante un estimador tipo núcleo na ecuación (3.2), obtemos o seguinte modelo:

$$\widehat{M}_n(x) = \left\{ y;\ \frac{\partial}{\partial y}\hat{f}_n(x,y) = 0,\ \frac{\partial^2}{\partial y^2}\hat{f}_n(x,y) < 0 \right\}. \tag{3.4}$$

Definido desta maneira, pode darse o caso, mencionado anteriormente, de que $\widehat{M}_n(x)$ conteña máis dun elemento da variable resposta. Os estimadores que teñen esta propiedade coñécense como estimadores de avaliación múltiple. Ademais, como indicamos no capítulo anterior, as diferentes $\hat{f}_n$ son funcións regulares. Deste xeito, o estimador recollido na ecuación (3.4) é de avaliación múltiple e regular. Estas propiedades resultan en que o modelo de regresión modal non paramétrica sexa máis versátil que os considerados anteriormente.

Se a variable explicativa ten soporte en $D \subset \mathbb{R}$, un posible estimador é o obtido a partir da aplicación do estimador unidimensional produto de estimadores tipo núcleo. Así, neste caso concreto,

$$\hat{f}_n(x,y) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_x}\right) K_2\left(\frac{y-Y_i}{h_y}\right),$$

onde K_1 e K_2 son funcións núcleo radialmente simétricas e $h_x, h_y > 0$ son as fiestras consideradas, respectivamente.

3.2. Estimación mediante o algoritmo *mean-shift*

Pese a que a regresión multimodal teña propiedades boas, como a súa versatilidade, a obtención do estimador $\widehat{M}_n(x)$ pode non ser sinxelo, xa que require do cálculo de máximos locais dunha estimación mediante unha función núcleo multivariante. Isto dificúltase, principalmente, porque a ecuación a resolver non ten solución explícita. Debido a isto desenvolvéronse, e desenvólvense, algoritmos asociados a métodos numéricos co fin de aproximar as estimacións. Un exemplo é o algoritmo *mean-shift*, que procedemos a mostrar a continuación.

Supoñamos que X e Y teñen soporte real e $f(x,y)$ é a súa función densidade conxunta. Supoñamos, ademais, que temos unha mostra $(X_1,Y_1),\dots,(X_n,Y_n)$ de vectores aleatorios independentes e identicamente distribuídos. Consideraremos funcións núcleo radialmente simétricas K_2, é dicir,

$$K_2(x) = c_k K(\|x\|^2), \quad \forall x \in D,$$

con c_k constante positiva. Denominaremos á función $K : [0, \infty) \to \mathbb{R}$ perfil de K_2.

Un posible exemplo ilustrativo de función núcleo radialmente simétrica é a función Gaussiana, neste caso teriamos $K_2(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. Logo, a constante positiva sería $c_k = (2\pi)^{-1/2}$ e o perfil de K_2 sería $K(x) = \exp(-x/2)$. Outro exemplo de función núcleo radialmente simétrica é a función Epanechnikov, $K_2(x) = \frac{3}{4}(1 - x^2)\mathbb{I}(|x| < 1) = \frac{3}{4}(1 - x^2)\mathbb{I}(x^2 < 1)$.

A partir da ecuación (2.14), tal e como vén recollido en [Einbeck e Tutz, 2006], podemos estimar a densidade condicional no caso univariante como

$$\hat{f}_n(y|x) = \frac{\hat{f}_n(x, y)}{\hat{f}_n(x)} = \frac{\sum_{i=1}^n K_1\left(\frac{x-X_i}{h_x}\right) K_2\left(\frac{y-Y_i}{h_y}\right)}{h_y \sum_{i=1}^n K_1\left(\frac{x-X_i}{h_x}\right)}, \tag{3.5}$$

onde K_1 e K_2 son os núcleos asociados aos estimadores de X e Y, respectivamente.

Da ecuación (3.5), dedúcese a expresión

$$\hat{f}_n(y|x) = \frac{c_k}{h_y} \sum_{i=1}^n p_i(x) K\left(\left(\frac{y - Y_i}{h_y}\right)^2\right),$$

onde p_i é unha función de peso non dependente da variable resposta.

Ademais, considerando a primeira das condicións recollidas na ecuación (3.1), temos que

$$\frac{\partial \hat{f}_n(y|x)}{\partial y} = \frac{2c_k}{h_y^3} \sum_{i=1}^n p_i(x) K'\left(\left(\frac{y - Y_i}{h_y}\right)^2\right)(y - Y_i) = 0.$$

Logo, temos que a ecuación anterior é equivalente á seguinte, a cal ten como primeiro membro o estimador modal buscado,

$$y_m(x) = \frac{\sum_{i=1}^n p_i(x) K'\left(\left(\frac{y_m - Y_i}{h_y}\right)^2\right) Y_i}{\sum_{i=1}^n p_i(x) K'\left(\left(\frac{y_m - Y_i}{h_y}\right)^2\right)}. \tag{3.6}$$

Se denotamos $y_m \equiv y_m(x)$, $g = -K'$ e $G(y) = c_g g(y^2)$, con c_g constante positiva, podemos expresar a ecuación (3.6), con $\omega(y_m) = y_m$ como

$$\omega(y) = \frac{\sum_{j=1}^n K\left(\frac{x-X_j}{h_x}\right) G\left(\frac{y-Y_j}{h_y}\right) Y_j}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_x}\right) G\left(\frac{y-Y_j}{h_y}\right)}. \tag{3.7}$$

A función ω así definida é unha ponderación das observacións onde os pesos varían segundo a proximidade ao punto avaliado. Porén, a ecuación $\omega(y_m) = y_m$ non pode resolverse analiticamente. Por mor disto, aplícase un proceso iterativo calculando medias locais mediante o algoritmo *mean-shift*.

Para poder aplicar o algoritmo, definimos a función *mean-shift* como

$$\omega(y) - y.$$

Neste caso, un máximo local da función densidade condicionada é un cero desta función. Logo, basta aplicar un proceso iterativo[1] da forma $y_{l+1} = \omega(y_l),\ l \in \mathbb{N}$.

Cabe resaltar o feito de que, xeralmente, a función densidade f non é cóncava. A causa disto, non podemos asegurar a converxencia ao máximo global deste método gradiente. Isto débese a que se f é cóncava, entón $-f$ é convexa e o algoritmo *mean-shift* é equivalente a aplicar un método gradiente de descenso a $-f$. Pero ao non poder asegurarmos que $-f$ é convexa, non se teñen por que verificar as condicións dos resultados de converxencia para o método gradiente de descenso[2]. Desta maneira, temos que este método ten carácter local. Cada moda local ten unha rexión de atracción, é dicir, unha veciñanza na cal, se o punto inicial considerado pertence a ela, o método converxerá a esa moda local.

Para o caso multimodal, ao pretender obter máis dunha moda nun x concreto, e como cada punto inicial do algoritmo converxe a unha única posible estimación modal, temos que considerar diferentes $y_0(x)$. Consideraremos $y_0^{(1)}(x), \dots, y_0^{(p)}(x)$ puntos de inicio, verificando que $y_0^{(1)}(x) < \dots < y_0^{(p)}(x)$. Así, para cada $t \in \{1, \dots, p\}$, faremos iteracións $y_{l+1}^{(t)}(x) = \omega(y_l^{(t)}(x))$, con $l = 1, 2, \dots$ ata que se verifique o criterio de converxencia considerado. Deste xeito, obteremos $\hat{y}_m^{(1)}, \dots, \hat{y}_m^{(p)}$ estimacións e poderemos estimar a función de regresión multimodal para o punto x como

$$\widehat{M_n}(x) = \{\hat{y}_m^{(1)}(x), \dots, \hat{y}_m^{(p)}(x)\}.$$

Claramente, ten que verificarse que p sexa o suficientemente grande para que teñamos alomenos un punto inicial no cal haxa converxencia a cada moda local. Notemos, ademais, que pode darse que para varios puntos iniciais distintos converxamos á mesma moda local, obviando así outras posibles estimacións, dependendo de en que rexións de atracción de cada moda local se encontre cada punto inicial.

O caso multimodal máis simple é no cal só consideramos dúas modas locais. Neste caso, basta considerar un punto inicial como o valor superior da distribución da nosa mostra e outro como o valor inferior, xa que converxerán á moda local de maior valor e á moda local de menor valor, respectivamente. Emporiso, de considerarmos máis de dúas modas locais, non temos unha forma refinada de obtelas máis aló de considerar un gran número

[1] Este é un algoritmo de iteración funcional para puntos fixos, estudado na materia obrigatoria de *Cálculo Numérico nunha Variable* de 2º curso.

[2] Os resultados de converxencia do método gradiente de descenso son contidos da materia obrigatoria de 3º curso *Métodos Numéricos en Optimización e Ecuacións Diferenciais.*

de puntos iniciais[3], tal e como se recolle en [Einbeck e Tutz, 2006].

Vexamos agora as propiedades xeométricas da regresión modal, para o que nos basearemos en [Chen et al., 2016]. Definimos a colección de variedades modais como a unión de todos os $M(x)$, é dicir,

$$\mathcal{S} = \{(x, y);\ x \in D,\, y \in M(x)\}.$$

Ao termos que o soporte da variable explicativa está contido en $\mathbb{R}$, polo teorema da función implícita, temos asegurado que a dimensión de $\mathcal{S}$ é 1. Supoñeremos que $\mathcal{S}$ pode expresarse como a unión de l variedades onde cada unha delas é unha variedade conexa admitindo unha parametrización do seguinte xeito:

$$S_j = \{(x, m_j(x)) :\ x \in A_j\},$$

onde m_j é unha función e A_j é un aberto. Así, temos que $\{A_1, \dots, A_l\}$ é unha cobertura por abertos do compacto D. Deste xeito, denotaremos $m_j(x) = \emptyset$ se $x \notin A_j\ \forall\, j \in \{1, \dots, l\}$. Logo, podemos expresar a función M como $M(x) = \{m_1(x), \dots, m_l(x)\}\ \forall\, x \in D$. Tal e como describimos a función M, cabe resaltar que é unha multifunción, téndose que o número de modas locais non ten por que ser constante. De feito, pode aumentar ou diminuír o número de modas locais cando varía x, sen ter que darse en puntos onde coincidan dúas variedades modais. Por exemplo, pode aparecer unha moda nova nun x concreto sen que haxa unha moda próxima a esta en ningún $z \in D$, con $z < x$.

Supoñamos que os m_j son funcións diferenciables. Denotemos, por simplicidade, $f_y(x, y) = \frac{\partial f}{\partial y}(x, y)$. Se consideramos que f é dúas veces diferenciable e que a colección de variedades modais, $\mathcal{S}$, se pode factorizar, entón tense que, para cada $x \in A_j$,

$$\nabla m_j(x) = -\frac{f_{yx}(x, m_j(x))}{f_{yy}(x, m_j(x))},$$

onde $f_{yx}(x, y) = \nabla_x \frac{\partial}{\partial y} f(x, y)$ é o gradiente sobre x de $f_y(x, y)$. Esta propiedade, de ser certa, daríanos unha relación entre a suavidade da función modal e a suavidade da función conxunta. Vexamos que se verifica.

Como $x \in A_j$, temos que $f_y(x, m_j(x)) = 0$ por definición de moda local. Entón, tomando o gradiente sobre x, tense que

$$0 = \nabla_x f_y(x, m_j(x)) = f_{yx}(x, m_j(x)) + f_{yy}(x, m_j(x)) \nabla m_j(x).$$

Despexando na igualdade anterior, e dado que $f_{yy}(x, m_j(x)) \neq 0$ por definición de moda local, temos que $\nabla m_j(x) = -\frac{f_{yx}(x, m_j(x))}{f_{yy}(x, m_j(x))}$.

[3] A idea intuitiva pode considerarse análoga ao caso de aplicar o método de dicotomía, estudado na materia de 2º curso *Cálculo Numérico nunha Variable*, onde haxa máis dun cero.

Antes de enunciar algún resultado sobre a suavidade de M, definamos a distancia Hausdorff. O feito de non empregar a distancia usual vén motivado porque, para cada x, consideramos o conxunto de modas locais en x, é dicir, traballamos con conxuntos no canto de puntos.

Dados dous conxuntos non baleiros $A, B \subset \mathbb{R}^d$, a distancia Hausdorff defínese como

$$\text{Haus}(A,B) = \text{máx}\left\{\sup_{x\in A} d(x,B), \sup_{x\in B} d(x,A)\right\},$$

onde $d(x,A) = \text{ínf}_{z\in A}||x-z||$ é a distancia[4] do punto x ao conxunto A. Empregamos o supremo das distancias e non o máximo porque cabe a posibilidade de que os conxuntos teñan infinitos puntos. Por outro lado, ao considerar soamente dous conxuntos, temos unicamente dous supremos e podemos afirmar a existencia do máximo de ambos. A noción desta distancia é que é pequena cando cada elemento de cada un dos dous conxuntos é próximo a, polo menos, un elemento do outro conxunto. Desta maneira, evitamos que o erro teña en conta a distancia entre as modas locais.

Teorema 3.1 (Suavidade da colección de variedades modais, de [Chen et al., 2016])**.** *Supoñamos que f é dúas veces diferenciable, $\mathcal{S}$ pode factorizarse, todas as derivadas parciais de f están limitadas e existe $\lambda_2 > 0$ tal que $f_{yy}(x,y) < -\lambda_2$ para todo $y \in M(x)$, $x \in D$. Entón, tense que*

$$\lim_{|\varepsilon|\to 0} \frac{Haus(M(x), M(x+\varepsilon))}{|\varepsilon|} \leq \max_{j=1,\dots,K} \|\nabla m_j(x)\| \leq \frac{C}{\lambda_2} < \infty,$$

onde C é unha cota do conxunto de derivadas parciais de f.

Demostración. Por ser f dúas veces diferenciable, e como $f_{yy}(x, m_j(x)) < -\lambda_2 < 0$, ao termos a igualdade $\nabla m_j(x) = -\frac{f_{yx}(x,m_j(x))}{f_{yy}(x,m_j(x))}$, $\forall j \in \{1,\dots,K\}$, $\nabla m_j(x)$ está ben definida para todo x. Ademais, ao terse $||f_{yx}(x, m_j(x))|| \leq C$, verifícase a segunda desigualdade. Para obter a primeira desigualdade, basta aplicar a definición de distancia Hausdorff. □

Deste xeito, podemos estimar $\mathcal{S}$, como $\hat{\mathcal{S}} = \hat{S}_1 \cup \dots \cup \hat{S}_{\hat{l}}$, sendo $\hat{l}$ a estimación do número de variedades e sendo cada $\hat{S}_j$ unha variedade conexa. Así, podemos considerar $\hat{m}_j(x)$, $\forall\, j \in \{1,\dots,\hat{l}\}$ e $\widehat{M}_n(x) = \{\hat{m}_1(x), \dots, \hat{m}_{\hat{l}}(x)\}$.

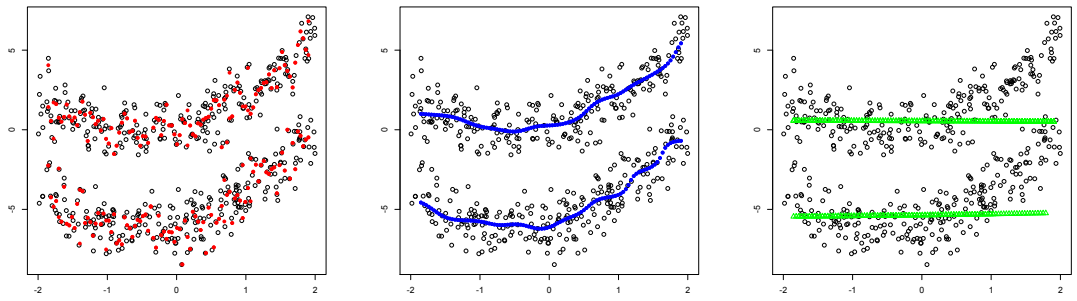
[4] A distancia dun punto a un conxunto foi introducida na materia de formación básica *Topoloxía dos espazos euclidianos* de 1º curso, aínda que a distancia entre conxuntos definida nesa materia fose outra distinta da Hausdorff.

3.3. Efecto dos parámetros de suavizado

Unha vez construído un método para obter estimacións, o algoritmo *mean-shift*, podemos ver na ecuación (3.7) que este depende de dúas fiestras h_x e h_y. No caso da regresión en media non paramétrica, a elección de fiestras óptimas era crucial para poder obter estimacións axeitadas. Ilustremos, mediante os exemplos da Sección 1.3, como varían as posibles estimacións para a regresión modal non paramétrica segundo se modifiquen os parámetros, h_x e h_y.

Nótese que o parámetro h_x é aplicado na estimación núcleo da variable explicativa, polo que afecta á suavidade das modas dun xeito similar á forma que afectaba o ancho de banda na estimación en media da densidade non paramétrica. Analogamente, o parámetro h_y controla a suavidade da estimación núcleo da variable Y. Debido a isto, este último parámetro afectará o número de variedades modais estimadas consideradas na nosa estimación. Este efecto pode verse claramente se empregamos unha función núcleo Epanechnikov, onde para cada y non ten en conta o valor da variable resposta nas observacións cuxo valor diste máis de h_y de y.

Co fin de mostrar a importancia da escolla dunha fiestra adecuada, vexamos varias estimacións feitas a partir dun mesmo conxunto de cincocentas observacións do modelo C2, recollido na ecuación (1.3). Para isto, variaremos unicamente h_x.

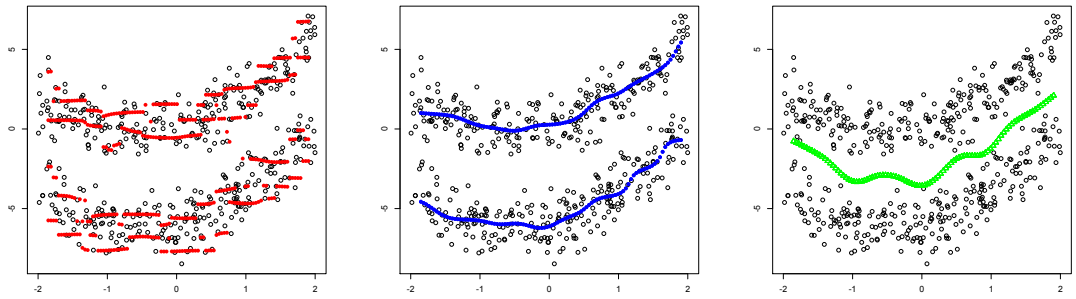


(a) (b) (c)

Figura 3.1: Simulacións e estimacións multimodais do modelo C2 con anchos de banda $h_y = 1$ e $h_x = 0{,}01$ en (a), $h_x = 0{,}2$ en (b) e $h_x = 3$ en (c).

Como se observa na Figura 3.1, ao manter fixo h_y o número de modas en cada x non varía. Pero vemos claramente o efecto de variar h_x, similar ao ancho de banda na estimación en media non paramétrica. Se este é excesivamente pequeno, obtemos unha variabilidade e unha irregularidade maiores. Por outro lado, se h_x é excesivamente grande, como na

terceira gráfica, temos que as dúas ramas tenden a ser paralelas ao eixo horizontal.



(a) (b) (c)

Figura 3.2: Simulacións e estimacións multimodais do modelo C2 con anchos de banda $h_x = 0{,}2$ e $h_y = 0{,}2$ en (a), $h_y = 1$ en (b) e $h_y = 5$ en (c).

Na Figura 3.2 vese o efecto de mudar o parámetro h_y e a súa relación co número de modas locais. Se este é moi pequeno, como no primeiro caso, chegaremos á conclusión errónea de que o número de modas locais é superior ao real. Por outro lado, se escollemos un h_y excesivamente grande estimaremos que hai unha única moda, obtendo un modelo que non se adecúa ás nosas observacións. Cabe destacar que, visualmente, o terceiro caso garda certo parecido coa estimación mediante regresión en media paramétrica, recollida na Figura 1.3. Isto débese ao feito de que, ao fixar unha función núcleo K_2, un maior h_y reduce o efecto da distancia entre os posibles valores da variable resposta, parecéndose a estimación á obtida mediante regresión en media.

3.4. Medidas de erro

Unha vez vista a importancia da selección adecuada dos parámetros h_x e h_y, cabe preguntarse se existe algún método para facer unha selección óptima destes. Para isto, é necesario establecer criterios de erro que nos permitan avaliar a bondade dunha estimación concreta. A diferenza doutros modelos de regresión considerados anteriormente, como o modelo de regresión en media, ao estar ante unha multifunción no canto dunha función univaluada, traballamos con distancias entre conxuntos e non entre puntos. Isto lévanos obrigatoriamente a descartar os criterios de erro descritos no Capítulo 2 e definir outros novos.

Cabe destacar que as nomenclaturas para as medidas de erro empregadas non teñen por que estar unificadas en toda a literatura. Un exemplo disto podemos velo na definición

do erro cadrático medio integrado, concepto que explicaremos posteriormente. Mentres en [Chen, 2018] se considera como ECMI a integral do erro puntual, en [Zhou e Huang, 2019] emprégase unha ponderación no interior da integral. No exposto a continuación, basearémonos no recollido en [Chen et al., 2016].

Unha vez definida a distancia Hausdorff e vistas varias das propiedades xeométricas, supoñendo as condicións descritas para que se verifiquen, vexamos o caso puntual. Así, definimos o erro puntual cometido pola estimación $\hat{M}$ no punto x como

$$\Delta(x) = \text{Haus}\left(M(x), \widehat{M}(x)\right). \tag{3.8}$$

Procedamos agora a profundizar na análise asintótica do erro. Supoñamos que temos unha estimación $\widehat{M}$ da colección de variedades modais M. No dominio considerado das variables explicativa e resposta, denotaremos por $\text{BC}^k(C)$ o conxunto de funcións k veces continuamente diferenciables e cuxas derivadas parciais están limitadas en valor absoluto pola constante real $C \in \mathbb{R}^+$.

Supoñamos que a densidade conxunta verifica as seguintes propiedades:

(A1): $f \in \text{BC}^4(C_f)$ para certa $C_f > 0$.

(A2): $\mathcal{S}$ pode ser factorizado como $\mathcal{S} = S_1 \cup ... \cup S_l$, onde $\forall\, j \in \{1, ..., l\}$, S_j é unha curva conexa que admite unha parametrización da forma $\{(x, m_j(x)) : x \in A_j\}$ con $\{A_1, ..., A_{K_{S_j}}\}$ cobertura por abertos de D, o cal supoñíamos compacto.

(A3): $\exists \lambda_2 > 0$ tal que $\forall\, (x, y) \in D \times S$ verificando $f_y(x, y) = 0$ se ten $|f_{yy}(x, y)| > \lambda_2$.

Nótese que estas propiedades implican que se cumpran as hipóteses do teorema da suavidade da colección de variedades modais. Concretamente, a condición (A1) permítenos afirmar que o nesgo da segunda derivada está limitado. Ademais, (A2) asegúranos que a colección formada polas modas locais pode representarse como unión finita de variedades. Por outro lado, (A3) permítenos afirmar que non existen casos onde as variedades converxan ou se separen.

Antes de introducir outras propiedades que debe verificar a propiedade núcleo co fin de probar outros resultados, introduzamos os conceptos de ε-cobertura dun espazo métrico e de clase de Vapnik-Chervonenkis.

Consideremos un espazo métrico (T, g), o número da ε-cobertura é o número mínimo de bolas esféricas de radio ε necesarias para cubrilo, podendo ter varias delas intersección non baleira.

Cabe mencionar que esta definición tamén é válida para espazos semimétricos, é dicir, espazos onde g verifica as condicións de ser distancia salvo a relativa á desigualdade triangular, a cal non se ten por que dar sempre.

Sexa S un conxunto, $\mathcal{K}$ unha colección de subconxuntos de S e F un subconxunto finito de S. Entón, $\mathcal{K}$ pulveriza a F se para cada subconxunto A de F se ten un subconxunto de $\mathcal{K}$, C, tal que $A = C \cap F$. Se existe un k, de tal xeito que $\mathcal{K}$ pulveriza alomenos un conxunto de cardinal k, e non existe $k' > k$ verificando esta propiedade, dise que $\mathcal{K}$ é da clase de Vapnik-Chervonenkis, ou VC. Ademais, dise que a súa dimensión Vapnik-Chervonenkis é $S(\mathcal{K}) = k$.

Procedamos agora a enunciar as propiedades que debe verificar a función núcleo para que se poidan afirmar os seguintes resultados.

(K1): $K \in \mathrm{BC}^2(C_K)$ para certo $C_K > 0$ e verifique

$$\int_{-\infty}^{\infty} (K^{(\alpha)})^2(u)\, du < \infty \quad \text{e} \quad \int_{-\infty}^{\infty} u^2 K^{(\alpha)}(u)\, du < \infty \quad \alpha \in \{0, 1, 2\},$$

onde $K^{(\alpha)}$ indica a derivada de orde α de K.

(K2): A colección $\mathcal{K} = \{v \mapsto K^{(\alpha)}\left(\frac{u-v}{h}\right) : u \in \mathbb{R},\, h > 0, \alpha = 0, 1, 2\}$ é de clase VC. Entón verifica que existen $A, v > 0$ tales que $\forall\, \varepsilon \in (0, 1)$, se ten

$$\sup_Q \mathcal{M}(\mathcal{K}, L_2(Q), C_K \varepsilon) \leq \left(\frac{A}{\varepsilon}\right)^v,$$

onde $\mathcal{M}(T, g, \varepsilon)$ é o número da ε-cobertura para o espazo semiparamétrico (T, g) e Q é unha medida de probabilidade.

A consecuencia de ser de clase VC recollida na condición (K2) pode obterse como resultado do exposto en [Haussler, 1995].

Vexamos agora un resultado que nos dá unha taxa do erro puntual, recollido en [Chen et al., 2016], e que non demostraremos. Para isto, empregamos as seguintes notacións:

$$\begin{aligned}
\|\hat{f}_n - f\|_\infty^{(0)} &= \sup_{x,y} \|\hat{f}_n(x, y) - f(x, y)\|, \\
\|\hat{f}_n - f\|_\infty^{(1)} &= \sup_{x,y} \|\hat{f}_{y,n}(x, y) - f_y(x, y)\|, \\
\|\hat{f}_n - f\|_\infty^{(2)} &= \sup_{x,y} \|\hat{f}_{yy,n}(x, y) - f_{yy}(x, y)\|, \\
\|\hat{f}_n - f\|_{\infty,2}^{(*)} &= \text{máx}\{\|\hat{f}_n - f\|_\infty^{(0)}, \|\hat{f}_n - f\|_\infty^{(1)}, \|\hat{f}_n - f\|_\infty^{(2)}\}.
\end{aligned}$$

Consideraremos un proceso estocástico, unha colección de variables aleatorias ordenadas $\{X_i;\ i \in T\}$, e denotaremos por $O_{\mathbb{P}}(1)$ cando a secuencia de variables aleatorias estea limitada en probabilidade. Analogamente, $o_{\mathbb{P}}(1)$ indicará converxencia en probabilidade a cero, é dicir, $\lim_{i \to \infty} P(|X_i| > \varepsilon) = 0$ para calquera $\varepsilon > 0$ arbitrario. Ademais,

tense que $O_{\mathbb{P}}(R_i)$ quere dicir que $X_i = Y_i R_i$ verifica que a secuencia $\{Y_i;\ t \in T\}$ está limitada en probabilidade e $o_{\mathbb{P}}(R_i)$ indica que $\{Y_i;\ i \in T\}$ converxe en probabilidade a cero. Tanto estas definicións, como unha análise das súas propiedades, veñen recollidas en [Van der Vaart, 1998].

Para enunciar e demostrar os seguintes resultados supoñeremos que podemos escoller unha fiestra de tal xeito que $h := h_x = h_y$. Esta suposición, tal e como se afirma en [Zhou e Huang, 2019], resulta en estimacións modais peores. Seguindo o traballo de [Chen et al., 2016], vexamos como podemos asegurar a existencia de certas cotas do erro teóricas.

Teorema 3.2 (Taxa do erro puntual, de [Chen et al., 2016])**.** *De verificarse as condicións previas, (A1), (A2), (A3), (K1) e (K2), se definimos o proceso estocástico*

$$A_n(x) := \begin{cases} \frac{1}{\Delta_n(x)} |\Delta_n(x) - \text{máx}_{z \in M(x)} \{|f_{yy}^{-1}(x,z)|, \hat{f}_{y,n}(x,z)|\}|, & \text{se } \Delta_n(x) > 0, \\ 0, & \text{se } \Delta_n(x) = 0. \end{cases}$$

Daquela, se $\|\hat{f}_n - f\|_{\infty,2}^{(*)}$ *é o suficientemente pequeno, tense que* $\sup_{x \in D} A_n(x) = O_{\mathbb{P}}(\|\hat{f}_n - f\|_{\infty}^{(2)})$*. Ademais, fixado* $x \in D$*, se* $\frac{nh^6}{\log n} \to \infty$ *e* $h \to 0$*,* $\Delta_n(x) = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^4}}\right)$*.*

A partir da expresión do erro puntual recollida na ecuación (3.8), se o soporte da variable explicativa é D, podemos considerar o erro cadrático integrado como

$$ECI = \int_D \Delta^2(x)\, dx.$$

Logo, definindo o erro uniforme, Δ_S, como $\Delta_S = \sup_{x \in D} \Delta(x)$ tense o seguinte resultado.

Teorema 3.3 (Taxa do erro uniforme, de [Chen et al., 2016])**.** *Se se ten que* $\frac{nh^6}{\log n} \to \infty$ *e* $h \to 0$*, baixo as condicións (A1), (A2), (A3), (K1) e (K2), entón*

$$\Delta_S = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^4}}\right).$$

Para a súa demostración, empregaremos o seguinte resultado, que non demostraremos, recollido en [Giné e Guillou, 2002].

Teorema 3.4. *Asumindo (K1), que a secuencia de bandas regulares* $\{h_n;\ n \in \mathbb{N}\}$ *verifica* $h_n \to 0$*,* $\frac{nh_n}{|\log h_n|} \to \infty$*,* $\frac{|\log h_n|}{\log\log n} \to \infty$*,* $h_n \leq c h_{2n}$ *para algún* $c > 0$*, e que* f *é unha función densidade limitada en* $\mathbb{R}^d$*, tense que*

$$\limsup_{n \to \infty} \sqrt{\frac{nh_n}{\log h_n^{-1}}} \|f - \hat{f}_n\|_{\infty} = C,$$

onde $C^2 \leq r^2 c \|f\|_{\infty} \|K\|_2^2$ *para unha constante* r *dependente da característica VC do espazo de funcións considerado.*

Demostración. (do teorema da taxa do erro uniforme, de [Chen et al., 2015]) Supoñamos que se verifica o teorema da taxa do erro puntual. Sexa n o número de observacións consideradas, $x \in D$ un punto fixado e y_j unha moda local estimada como $\hat{y}_j$. Entón,

$$\begin{aligned}\Delta_n(x) &= \max_j\{|f_{yy}^{-1}(x,y_j)||\hat{f}_{y,n}(x,y_j)|\} + o_{\mathbb{P}}(1)\\ &= \max_j\{|f_{yy}^{-1}(x,y_j)||(\hat{f}_{y,n}(x,y_j) - \mathbb{E}(\hat{f}_{y,n}(x,y_j))| + B(x,y_j))\} + o_{\mathbb{P}}(1),\end{aligned}$$

onde $B(x,y_j) = |\mathbb{E}(\hat{f}_{y,n}(x,y_j)) - f_y(x,y_j)| = O(h^2)$ denota o nesgo e $O_{\mathbb{P}}(\|\hat{f}_n - f\|^*_{\infty,2}\Delta_n(x))$ implica $o_{\mathbb{P}}(1)$.

Se $f_{yy}(x,y_j)^{-1}$ está limitado, entón $|f_{yy}(x,y_j)^{-1}|B(x,y_j) = O(h^2)$ e tense a igualdade

$$\Delta_n(x) = \max_j\{|f_{yy}^{-1}(x,y_j)||\hat{f}_{y,n}(x,y_j) - \mathbb{E}(\hat{f}_{y,n}(x,y_j))\} + O(h^2) + o_{\mathbb{P}}(1),$$

onde $O(h^2)$ indica o nesgo e é independente do punto x. Así, se consideramos o supremo sobre $x \in D$, e denotando

$$Z = \sup_{x\in D}\max_j\{|f_{yy}(x,y_j)^{-1}||\hat{f}_{y,n}(x,y_j) - \mathbb{E}(\hat{f}_{y,n}(x,y_j))|\},$$

tense que

$$\Delta_n = Z + O(h^2) + o_{\mathbb{P}}(1). \tag{3.9}$$

Designemos o espazo de funcións

$$\mathcal{F}_0 = \left\{(u,v) \mapsto g_{x,y}(u,v) = f_{yy}^{-1}(x,y)\cdot K\left(\frac{\|x-u\|}{h}\right)K^{(1)}\left(\frac{y-v}{h}\right),\, y \in M(x), x \in \mathbb{R}\right\},$$

e sexa o proceso empírico

$$\mathbb{G}_n(g) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n g(Z_i) - \mathbb{E}(g(Z_i))\right), \quad g \in \mathcal{F}_0.$$

Entón,

$$Z = \sup_{x\in D}\max_j\{|f_{yy}(x,y_j)^{-1}||\hat{f}_{y,n}(x,y_j) - \mathbb{E}(\hat{f}_{y,n}(x,y_j))|\} = \frac{1}{h^4\sqrt{n}}\sup_{g\in\mathcal{F}_0}|\mathbb{G}_n(g)|.$$

Como se ten (A1), (K1) e (K2), $\mathcal{F}_0$ é de clase VC con constante $\frac{C_K^2}{\lambda_2}$.

En consecuencia, polo teorema previo,

$$Z = \sup_{x\in D}\max_j\{|f_{yy}(x,y_j)^{-1}||\hat{f}_{y,n}(x,y_j) - \mathbb{E}(\hat{f}_{y,n}(x,y_j))|\} = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^4}}\right).$$

Aplicando isto á ecuación (3.9), tense

$$\Delta_n = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^4}}\right),$$

tal e como queriamos demostrar. □

De xeito análogo ao caso da estimación non paramétrica da densidade, podemos definir o erro cadrático medio integrado como

$$\mathrm{ECMI}(\hat{M}) = \mathbb{E}\left(\int_{x\in D} \Delta^2(x)dx\right).$$

Ademais, verifícase o seguinte resultado.

Teorema 3.5 (Taxa do ECMI, de [Chen et al., 2016])**.** *Nas condicións (A1), (A2), (A3), (K1), e (K2), e se se verifica* $\frac{nh^6}{\log n} \to \infty$ *e* $h \to 0$*, tense que*

$$ECMI(\hat{M}) = O(h^4) + O\left(\frac{1}{nh^4}\right).$$

Demostración. Polo teorema da taxa do erro puntual, tense que

$$\mathbb{E}(\Delta_n^2(x)) = O(h^4) + O\left(\frac{1}{nh^4}\right) = \mathrm{Nesgo}^2(x) + \mathrm{Var}(x).$$

Pódese demostrar que, ao igual que na regresión en media non paramétrica, o nesgo e a varianza teñen a mesma taxa de converxencia. □

Enunciemos agora, sen demostración, un corolario destes resultados para a estimación da densidade conxunta suavizada, $\tilde{f}(x,y) = \mathbb{E}(\hat{f}_n(x,y))$. Neste caso, consideraremos $\widetilde{M}(x) = \mathbb{E}(\widehat{M}_n(x))$, $\tilde{\Delta}_n(x) = \mathrm{Haus}(\widetilde{M}(x), \widehat{M}_n(x))$, $\tilde{\Delta}_n = \sup_{x\in D} \tilde{\Delta}_n(x)$ e $\widetilde{\mathrm{ECMI}}(\widehat{M}_n) = \mathbb{E}\left(\int_{x\in D} \tilde{\Delta}_n^2(x)dx\right)$.

Corolario 3.6 (Taxas de erro para modas condicionais suavizadas, de [Chen et al., 2016])**.** *Nas condicións do teorema da taxa do ECMI, tense que:*

$$\sqrt{nh^4}\sup_{x\in D}\left|\tilde{\Delta}_n(x) - \max_{z\in\widetilde{M}(x)}\{\tilde{f}_{yy}^{-1}(x,z)\hat{f}_{y,n}(x,z)\}\right| = O_{\mathbb{P}}(\varepsilon_{n,2}),$$

$$\tilde{\Delta}_n(x) = O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^4}}\right),$$

$$\tilde{\Delta}_n = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^4}}\right),$$

$$\widetilde{ECMI}(\widehat{M}_n) = O\left(\frac{1}{nh^4}\right),$$

onde

$$\varepsilon_{n,2} = \sup_{x,y}|\hat{f}_{yy,n}(x,y) - \tilde{f}_{yy}(x,y)|.$$

Deste xeito, se consideramos as estimacións suavizadas obtemos taxas de converxencia mellores.

3.5. Selección de fiestras de suavizado

Unha vez definidas as medidas de erro, vexamos como podemos facer unha selección adecuada de $h = (h_x, h_y)$ para estas medidas de erro. En primeiro lugar, mostraremos un método de validación cruzada, recollido en [Zhou e Huang, 2019]. Notemos que o esquema básico do método é igual ao análogo empregado para a estimación da densidade condicional descrito no Capítulo 2.

Supoñamos que temos $(X_1, Y_1), \dots, (X_n, Y_n)$ observacións independentes e identicamente distribuídas. Daquela, escollemos a fiestra que minimiza a función

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} d^2(\hat{M}_{-i}(X_i), Y_i) N_{-i}^2(X_i) p(X_i), \tag{3.10}$$

onde $\hat{M}_{-i}(X_i)$ é a estimación obtida ao aplicar o algoritmo *mean-shift* sobre a mostra sen o valor i-ésimo con parámetro de suavizado h, $N_{-i}(X_i)$ é o número de elementos de $\hat{M}_{-i}(X_i)$ e p é unha función de peso non dependente da variable resposta.

Logo, o valor de $h = (h_x, h_y)$ que minimice a función recollida na ecuación (3.10) daranos unhas posibles fiestras adecuadas.

Este método, tal e como se mostra en [Zhou e Huang, 2019], de momento, non ten unha xustificación teórica que avale a súa consistencia. De feito, no modelo simulado, recollido na ecuación (1.2), non semella consistente para o ECI. Así, habería aberta unha liña de investigación consistente en probar a consistencia, ou inconsistencia, de CV ou atopar un criterio de validación cruzada mellor.

Ante o feito de que este método poida non ser adecuado en certas ocasións, procedemos a mostrar outro posible selector de ancho de banda empregando un método *Bootstrap*. A idea xeral deste selector é aproximar un criterio de erro global mediante remostraxe, o que lle dá nome ao método, e buscar a fiestra que minimiza esta aproximación. De todos os xeitos, como se menciona en [Alonso-Pena, 2020], este método é computacionalmente lento.

Para a súa explicación, consideraremos un criterio de erro baseado no erro puntual e similar ao erro cadrático integrado. Este erro, que denominaremos erro cadrático integrado ponderado, defínese como

$$\mathrm{ECI}_p(h) = \int_D \Delta^2(x) f(x) p(x)\, dx, \tag{3.11}$$

onde h é unha fiestra concreta e p unha función de peso. A diferenza do ECI, introducimos na definición deste criterio de erro a función de densidade co fin de darlle unha importancia maior ao erro puntual cometido onde a función de densidade sexa maior.

Supoñamos que temos $(X_1, Y_1), \ldots, (X_n, Y_n)$ observacións independentes e identicamente distribuídas. Sexa h o ancho de banda considerado, entón, podemos obter unha estimación $\widehat{M}(x)$ de $M(x)$ mediante *mean-shift* considerando como fiestra h. Deste xeito, definimos a función

$$\mathrm{A}(h; X, Y, \widehat{M}, M) = \frac{1}{n} \sum_{i=1}^{n} \Delta(X_i)^2 p(X_i). \tag{3.12}$$

Pero, ao igual que na estimación non paramétrica da regresión en media, descoñecemos os $M(X_i)$. Para solventar isto, consideramos θ mostras obtidas mediante remostraxe, $\{(X, Y^{(i)});\ i = 1, \ldots, \theta\}$. Cabe mencionar que, sen ser a única posibilidade, a forma empregada é unha estimación en media paramétrica da densidade condicional. Así, unha posible estimación de A é

$$\hat{A}(h) = \frac{1}{\theta} \sum_{i=1}^{\theta} A(h; X, Y^{(i)}, \widehat{M}^{(i)}, \widehat{M}^{*}),$$

onde $\widehat{M}^{(i)}$ denota a estimación modal da i-ésima mostra obtida mediante remostraxe e $\widehat{M}^{*}$ denota a estimación modal obtida ao considerar as modas da estimación da densidade condicional.

3.6. Estudo de simulación

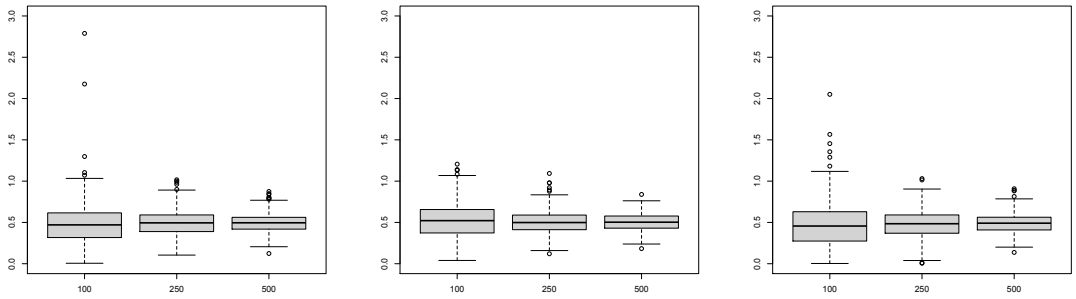
Para ilustrar a idea do erro puntual e como varía segundo o tamaño da mostra, vexamos o erro local cometido na estimación para $x \in \{0,1, 5, -1,5\}$, empregando o algoritmo *mean-shift*.

Para isto, en cada un dos modelos descritos no primeiro capítulo, C1, C2 e C4, realizáronse cincocentas simulacións con $n = 100$, $n = 250$ e $n = 500$. Vexamos agora a evolución do erro puntual segundo aumenta o tamaño da mostra para cada modelo. Co fin de escoller anchos de banda adecuados, empregouse en cada simulación en `R` a función `moderegbw` do paquete `lpme`, de [Zhou e Huang, 2019]. Esta función selecciona a fiestra óptima a partir dunha grella delas empregando o método de validación cruzada visto na sección anterior, e recollido na ecuación (3.10).

Erro puntual para o modelo C1:

Neste caso, na Figura 3.3, podemos observar que para o modelo C1, recollido na ecuación (1.2), existen similitudes entre os erros puntuais para $x = -1,5$, $x = 0$ e $x = 1,5$. Comparando os diagramas de caixa para os distintos x, parece observarse que unha maior cantidade de observacións non se traduce necesariamente nun erro puntual menor, pois, para os tres puntos, a mediana dos erros non se aproxima a cero cando aumenta n. Non

obstante, tense que os erros puntuais se concentran nun intervalo menor. Isto resulta en que o posible erro máximo cometido nunha estimación sexa cada vez menor. Este feito nótase especialmente no caso de $x = -1{,}5$, onde, para $n = 100$, chegamos a ter estimacións onde o erro puntual é superior a 2. Non obstante, para $n = 500$, e $x = -1{,}5$, o erro puntual non chega a ser superior a 1, o cal indica unha redución notable do erro puntual máximo cando aumenta n.



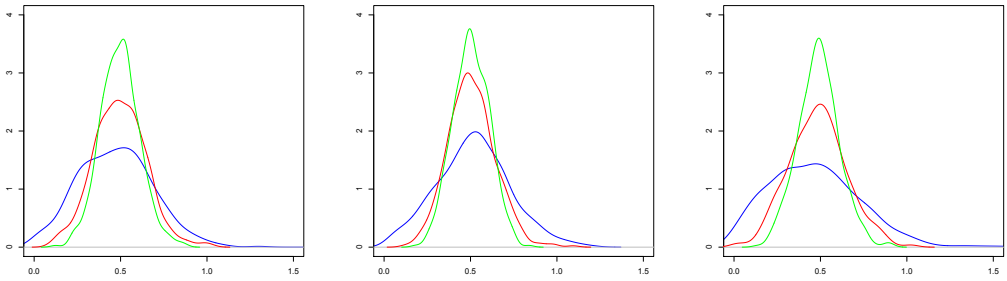
(a) (b) (c)

Figura 3.3: Diagramas de caixa dos erros puntuais cometidos aplicando o algoritmo *mean-shift* para 500 simulacións do modelo C1 en $x = -1{,}5$, $x = 0$ e $x = 1{,}5$ respectivamente.

Unha posible explicación deste fenómeno é que estamos ante un caso unimodal e o h_y empregado é excesivamente grande. Deste xeito, como explicabamos ao falarmos da importancia da selección dos parámetros, isto lévanos a que a estimación modal se aproxima a unha estimación en media. Neste caso, e aínda que a moda do erro sexa cero, este non está centrado. Isto pode verse xa que a media da distribución $\Gamma(3, 2)$ é $\frac{3}{2} = 1{,}5$, e como o erro considerado na ecuación (1.2) é $\Gamma(3, 2) - 1$, a súa media é un medio. Por outro lado, a moda da distribución $\Gamma(3, 2)$ é un, e tense que a moda do erro do modelo é cero. En consecuencia, asume que o modelo a estimar é

$$Y = 0{,}5 + X + X^2 + \varepsilon. \tag{3.13}$$

Cabe resaltar que, de ser certa a explicación deste feito, teriamos un exemplo onde o criterio de validación cruzada non é axeitado para seleccionar unha fiestra adecuada neste caso. Pois, como mencionamos ao introducilo, ao non termos unha xustificación teórica que o apoie, non podemos afirmar que funcione correctamente en todos os casos.



(a) Estimación tipo núcleo da densidade dos erros puntuais para $x = -1{,}5$

(b) Estimación tipo núcleo da densidade dos erros puntuais para $x = 0$

(c) Estimación tipo núcleo da densidade dos erros puntuais para $x = 1{,}5$

Figura 3.4: Estimación tipo núcleo para 500 simulacións do modelo C1 do erro puntual. En azul, para $n = 100$, en vermello para $n = 250$ e en verde para $n = 500$.

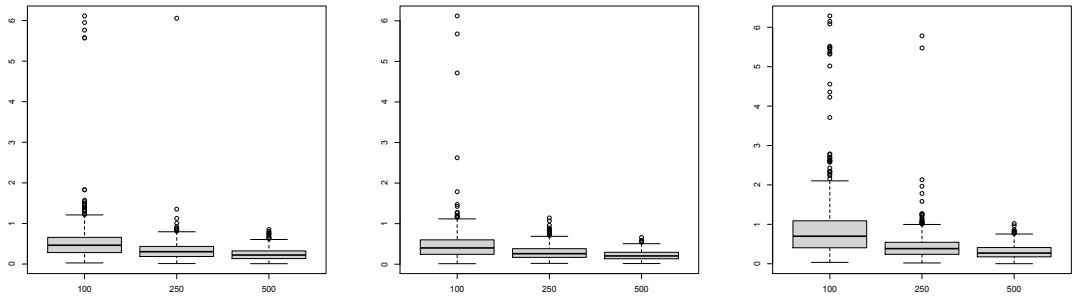
n	$x = -1{,}5$	$x = 0$	$x = 1{,}5$
100	0,48	0,52	0,47
250	0,50	0,51	0,48
500	0,50	0,50	0,49

Cadro 3.1: Medias dos erros puntuais para cada punto e tamaño mostral.

Aplicando a función `density` de `R`, a cal nos proporciona unha estimación tipo núcleo da densidade dos erros, obtemos para os tres puntos as gráficas recollidas na Figura 3.4. Podemos observar que a nosa intuición de que había unha concentración maior dos erros puntuais nunha mesma franxa cando aumenta n, pero sen ter unha mellora dos erros puntuais cometidos, parece ser a correcta. De feito, observando o Cadro 3.1, temos que nos casos de $x = 1{,}5$ e $x = -1{,}5$ incluso se produce un pequeno, pero non significante, aumento das medias.

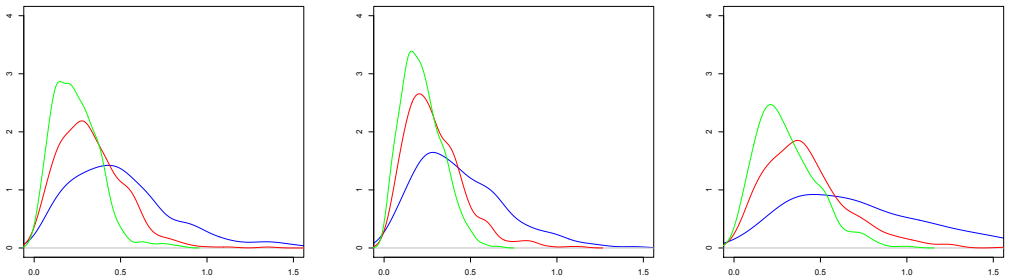
Erro puntual para o modelo C2:

No modelo C2, recollido na ecuación (1.3), temos que para os tres valores de x considerados hai unha clara redución do erro puntual cometido conforme aumentamos o valor de n. Isto obsérvase na Figura 3.5 e para os tres casos, ao ser menor a mediana conforme aumenta o tamaño da mostra. Ademais, o aumento da mostra tamén parece reducir o número de casos, puntuais, onde o erro pode ser considerado atipicamente grande.



(a) (b) (c)

Figura 3.5: Diagramas de caixa dos erros puntuais cometidos aplicando o algoritmo *mean-shift* para 500 simulacións do modelo C2 en $x = -1{,}5$, $x = 0$ e $x = 1{,}5$ respectivamente.



(a) Estimación tipo núcleo da densidade dos erros puntuais para $x = -1{,}5$

(b) Estimación tipo núcleo da densidade dos erros puntuais para $x = 0$

(c) Estimación tipo núcleo da densidade dos erros puntuais para $x = 1{,}5$

Figura 3.6: Estimación tipo núcleo para 500 simulacións do modelo C2 do erro puntual. En azul para $n = 100$, en vermello para $n = 250$ e en verde para $n = 500$.

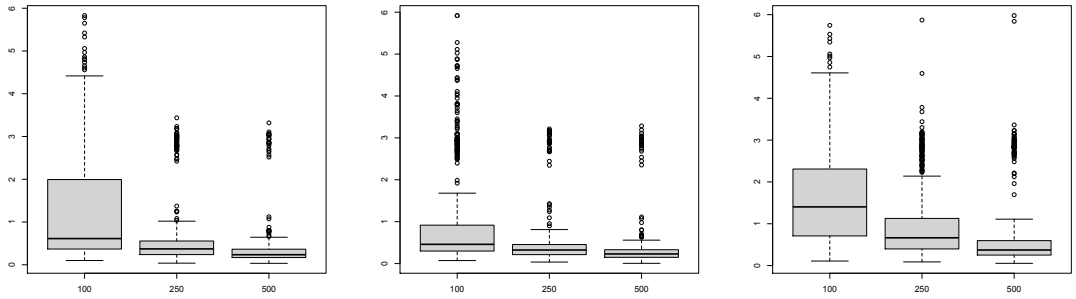
n	$x = -1{,}5$	$x = 0$	$x = 1{,}5$
100	0,56	0,48	0,92
250	0,34	0,29	0,45
500	0,24	0,22	0,31

Cadro 3.2: Medias dos erros puntuais para cada punto e tamaño mostral.

Tanto a Figura 3.6 como o Cadro 3.2 confirman o observado no diagrama de caixas. Tense que o aumento do tamaño mostral implica un descenso significativo da media dos erros puntuais en cada un dos tres puntos considerados.

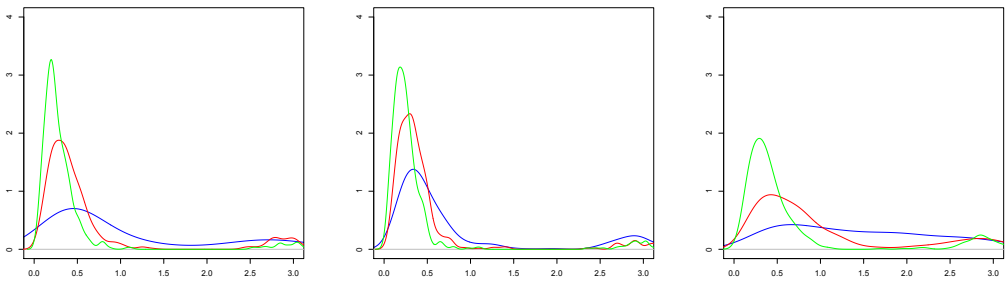
Erro puntual para o modelo C4:

Do mesmo xeito que no modelo C2, observando os diagramas de caixas dos erros chegamos á conclusión de que hai unha diminución dos erros ao aumentar o tamaño da mostra. Isto corrobórase ao observar o Cadro 3.3. Esta similitude pode deberse a que os dous modelos son mesturas de normais, a diferenza do modelo C1, no cal se empregou unha distribución Gamma. Con todo, neste caso si que continúa habendo estimacións cuxo erro puntual se afasta de forma atípica cando aumenta o tamaño mostral, de forma máis destacada no punto $x = 1{,}5$.



(a) (b) (c)

Figura 3.7: Diagramas de caixa dos erros puntuais cometidos aplicando o algoritmo *mean-shift* para 500 simulacións do modelo C4 en $x = -1{,}5$, $x = 0$ e $x = 1{,}5$ respectivamente.



(a) Estimación tipo núcleo da densidade dos erros puntuais para $x = -1{,}5$

(b) Estimación tipo núcleo da densidade dos erros puntuais para $x = 0$

(c) Estimación tipo núcleo da densidade dos erros puntuais para $x = 1{,}5$

Figura 3.8: Estimación tipo núcleo para 500 simulacións do modelo C4 do erro puntual. En azul para $n = 100$, en vermello para $n = 250$ e en verde para $n = 500$.

n	$x = -1{,}5$	$x = 0$	$x = 1{,}5$
100	1,21	1,02	1,61
250	0,63	0,50	1,00
500	0,39	0,38	0,68

Cadro 3.3: Medias dos erros puntuais para cada punto e tamaño mostral.

Vexamos agora cal sería a estimación modal para as tres simulacións realizadas, as cales compararemos coas nosas observacións e coas modas teóricas. Para isto, empregaremos as fiestras obtidas no caso puntual e aplicarémolas ao conxunto de datos de cada unha das tres simulacións. Cabe lembrar que o selector empregado para a obtención do ancho de banda é o de validación cruzada.

Estimación modal para o modelo C1:

Como adiantabamos ao valorar o erro puntual, a estimación modal obtida con este método para este caso non é moi adecuada. Esta afirmación pode apoiarse visualmente ao observar a Figura 3.9, onde a estimación modal para o modelo se asemella moito máis á estimación en media do modelo recollido na ecuación (3.13). Noutras palabras, recolle a estimación en media do modelo (1.2) asumindo que os erros deste están centrados.

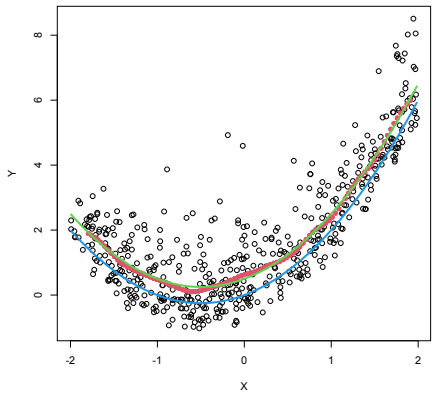


Figura 3.9: Moda teórica, en azul para a ecuación (1.2), en verde media do modelo recollido na ecuación (3.13), estimación modal, en vermello, e nube de puntos dos datos da simulación.

Estimación modal para o modelo C2:

Fronte ao modelo paramétrico da regresión en media, recollido na Figura 1.3, a estimación modal da Figura 3.10 adáptase de forma máis axeitada ás nosas observacións,

así como á moda real. Así, para un valor da variable explicativa concreto, teriamos unha mellor estimación do valor que pode tomar a variable resposta.

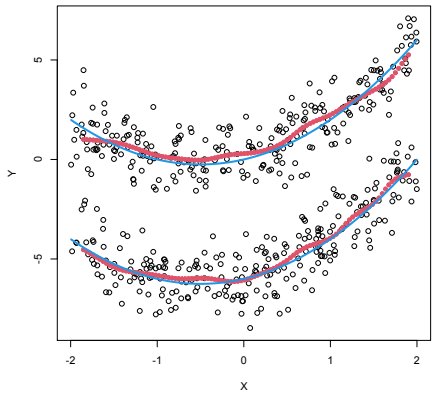


Figura 3.10: Moda teórica, en azul para a ecuación (1.3), estimación modal, en vermello, e nube de puntos dos datos da simulación.

Estimación modal para o modelo C4:

Como se observa na Figura 3.11, e ao igual que no modelo C2, neste caso temos unha estimación do modelo recollido na ecuación (1.4) moito máis adecuada que a feita no modelo linear da regresión en media asociado, recollida na Figura 1.5.

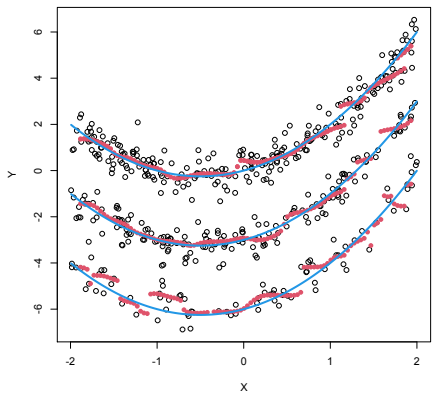


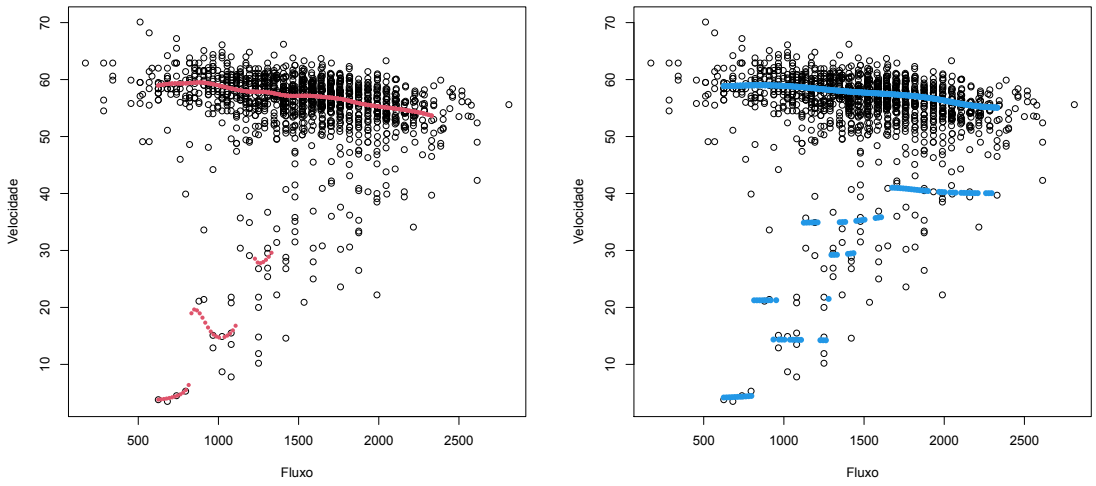
Figura 3.11: Moda teórica, en azul para a ecuación (1.4), estimación modal, en vermello, e nube de puntos dos datos da simulación.

Cabe mencionar unha problemática asociada a este tipo de estimación, e que se pode observar claramente na Figura 3.11, o feito de non distinguir nunha mesma estimación

que modas locais son maiores e cales son menores. Aínda que sexa capaz de captar as modas locais, cada estimación é capaz de detectar, pero non de distinguir, unhas certas modas. Así, non nos indica se, entre elas, estas son maiores ou menores. Por exemplo, neste caso concreto, a moda correspondente a $m(X) = X + X^2$, a superior das tres, a cal englobaría a metade das observacións aproximadamente, non se distingue das outras dúas, correspondentes ás ecuacións $m(X) = -3 + X + X^2$ e $m(X) = -6 + X + X^2$, englobando aproximadamente tres décimos e un quinto das observacións, respectivamente. Este defecto vén causado polo carácter local da estimación multimodal.

3.7. Ilustración con datos reais

Vexamos agora un exemplo de aplicación a datos reais de estimación da regresión multimodal. Para isto, empregaremos os datos de velocidade e fluxo de vehículos da Sección 1.3, onde vimos que o modelo de regresión paramétrica en media non era axeitado, ao non verificarse a normalidade dos residuos do modelo.



(a) (b)

Figura 3.12: Estimacións modais da velocidade segundo o fluxo empregando como selector de fiestra validación cruzada, en a), e o método *bootstrap* con 500 interaccións, en b).

Comparando a estimación representada na Figura 1.7 coas estimacións da Figura 3.12, chegamos á conclusión de que a simulación modal recolle de forma máis adecuada a posibilidade de que os vehículos vaian lentos cando o fluxo é baixo. Ademais, neste caso si se verifican todas as hipóteses requiridas, polo que, aínda que requira empregar un mé-

todo máis complexo e custoso computacionalmente, esta estimación é máis acaída para o conxunto de datos.

Unha vez feita a avaliación da estimación modal en comparación coa regresión en media para a velocidade e o fluxo, procedamos a valorar se hai diferenzas entre a fiestra obtida con cada un dos selectores de fiestra vistos e o custo computacional de facer a estimación modal empregando ese selector.

Estimación	Validación cuzada	*Bootstrap*
Fiestra	(69,508, 7,018)	(208,505, 1,76105)
Tempo de execución	920,17 s	73365,27 s

Cadro 3.4: Fiestra e tempo de execución para o cálculo da fiestra e a estimación modal en `R` para os datos de fluxo e velocidade segundo o selector.

Observando a Figura 3.12, vemos pequenas diferenzas nas estimacións modais obtidas. Entre elas, tal e como se recolle no Cadro 3.4, o h_x empregado no algoritmo *mean-shift* é maior no caso do selector *Bootstrap*, mentres o h_y é menor. Isto resulta en que no caso do selector de validación cruzada haxa unha maior regularidade nas estimacións e tenda a considerar un menor número de modas locais. Isto pode verse na Figura 3.12, pois en b) hai practicamente dúas modas locais para case todo valor do fluxo, mentres en a), para valores do fluxo maiores que 1500 só considera unha única moda local.

Por outro lado, o tempo de execución necesario é bastante menor empregando validación cruzada, como se indica no Cadro 3.4. Esta eiva do selector *Bootstrap* xa fora mencionada na sección referente á selección de fiestras adecuadas, e fai que poida ser preferible o selector de validación cruzada. Pero, cabe recordar que este é un selector sen sustento teórico que nos garanta a súa converxencia, polo que non podemos asegurar o seu funcionamento adecuado en todos os casos.

3.8. Conclusións

En conclusión, temos que os modelos construídos mediante a regresión en media, tanto paramétrica como non paramétrica, non son sempre os idóneos para captar relacións entre variables por diversos motivos. Entre estes, podemos mencionar o feito de que non se verifique a normalidade dos erros, ou que estes sigan unha distribución multimodal, como, por exemplo, unha formada pola mestura de normais ou calquera outro patrón que dea lugar a "distintas" nubes de puntos. Debido a isto, os modelos de regresión modal non paramétrica son capaces de detectar, e modelar axeitadamente, moitas das relacións que non se poden

obter por medio dos métodos de regresión convencionais. Esta gran versatilidade fai da regresión modal non paramétrica unha alternativa útil.

Este tipo de modelos poden construírse empregando o algoritmo *mean-shift*, un método iterativo de gradiente. Pese a isto, a necesidade de seleccionar un ancho de banda para obter resultados satisfactorios implica basearse en métodos de selección sen un sustento teórico amplo, debido a que é relativamente recente, o cal resulta en que, aínda que habitualmente o seu uso sexa o idóneo, non temos un resultado teórico que nolo asegure. Isto é a causa de que poida darse o caso de que a fiestra obtida mediante estes métodos de selección non sexa a acaída, como podía verse na Sección 3.5, no erro puntual ou na estimación modal do modelo C1.

Finalmente, o feito de que a regresión multimodal non paramétrica comezase a ser investigada recentemente resulta en que haxa moitas, e diversas, liñas de investigación abertas teoricamente e practicamente. A nivel teórico, por exemplo, non hai unha resposta contundente sobre a idoneidade da distancia Hausdorff para este contexto. A nivel práctico, podemos mencionar o desenvolvemento dun selector de fiestra que se adecúe a todas as posibilidades que se poidan dar.

Apéndice A

O algoritmo mean-shift como método de gradiente

Un método gradiente[1] é un método de carácter iterativo que se aproxima a un máximo, se é ascendente, ou mínimo, se é descendente, dunha función obxectivo. Para isto, en cada iteración obtense un punto a partir do cálculo do gradiente, ou vector de máximo crecemento, do punto anterior, o sentido variará se é ascendente ou descendente.

Procedamos agora a mostrar que a versión clásica do algoritmo *mean-shift*, para estimar as modas dunha función de densidade, é un método gradiente de ascenso aplicado á función densidade. Baseándonos en [Comaniciu e Meer, 2002], vexamos que en cada iteración a súa dirección e sentido coinciden co gradiente da función densidade estimada e é proporcional a esta.

Consideremos n observacións independentes e identicamente distribuídas, entón temos que, para unha función núcleo radialmente simétrica da forma $K_2(x) = c_k K(||x||^2)$, sendo $c_k > 0$ constante, podemos facer a estimación tipo núcleo, cun ancho de banda axeitado,

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K_2 \left(\frac{x - x_i}{h} \right).$$

Polas propiedades que caracterizan a esta función tipo núcleo tense

$$\hat{f}_n(x) = \frac{c_k}{nh} \sum_{i=1}^{n} K \left(\left\| \frac{x - x_i}{h} \right\|^2 \right).$$

En consecuencia, se a función perfil, K, é derivable en $[0, \infty)$, xa que ao provir dunha función núcleo radialmente simétrica, podemos restrinxir o seu dominio aos reais non ne-

[1] O concepto de método gradiente pertence ao contido da materia obrigatoria de 3º curso *Métodos Numéricos en Optimización e Ecuacións Diferenciais.*

gativos, podemos expresar a estimación do gradiente como

$$\hat{\nabla} f(x) \equiv \nabla \hat{f}_n(x) = \frac{2c_k}{nh^3} \sum_{i=1}^{n} (x - x_i) K' \left(\left\| \frac{x - x_i}{h} \right\|^2 \right). \tag{A.1}$$

Deste xeito, se definimos a función $g(x) = -K'(x)$, podemos definir unha función tipo núcleo radialmente simétrico do xeito $G(x) = c_g g(\|x\|^2)$, onde c_g é unha constante normalizadora. Así, substituíndo na ecuación (A.1), e supoñendo que $\sum_{i=1}^{n} g \left(\left\| \frac{x - x_i}{h} \right\|^2 \right)$ é estritamente positivo, temos que

$$\hat{\nabla} f(x) = \frac{2c_k}{nh^3} \sum_{i=1}^{n} (x_i - x) g \left(\left\| \frac{x - x_i}{h} \right\|^2 \right) = \frac{2c_k}{nh^3} \left[\sum_{i=1}^{n} g \left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} x_i g(\left\| \frac{x - x_i}{h} \right\|^2)}{\sum_{i=1}^{n} g(\left\| \frac{x - x_i}{h} \right\|^2)} - x \right].$$

Aplicando a ecuación (3.7) neste caso concreto, temos que

$$\hat{\nabla} f(x) = \frac{2c_k}{nh^3} \left[\sum_{i=1}^{n} g \left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \right] [\omega(x) - x].$$

Para concluír, se denotamos por $\hat{f}_g$ a estimación tipo núcleo da densidade f mediante a función tipo núcleo G, temos que

$$\hat{\nabla} f(x) = \hat{f}_g(x) \frac{2c_k}{h^2 c_g} (\omega(x) - x),$$

e queda probado que é un método gradiente.

Bibliografía

[Alonso-Pena, 2020] Alonso-Pena, M. (2020). An introduction to nonparametric multimodal regression. *Boletín de Estadística e Investigación Operativa 36*, 5-23.

[Boyer, 1987] Boyer, C. (1987). *Historia de la Matemática*. Alianza Editorial.

[Chen et al., 2016] Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., e Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics, 44*, 489–514.

[Chen et al., 2015] Chen, Y.-C., Genovese, C. R., Tibshirani, R. J. e Wasserman, L. (2015). Supplement to "Nonparametric modal regression". DOI:10.1214/15-AOS1373SUPP.

[Chen, 2018] Chen, Y-C. (2018). Modal regression using kernel density estimation: A review. *WIREs Computational Statistics, 10*, e1431.

[Comaniciu e Meer, 2002] Comaniciu, D. e P. Meer (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 603–619.

[Cristóbal, 1992] Cristóbal, J. (1992). *Inferencia Estadística*. Servicio de Publicaciones Universidad de Zaragoza.

[Einbeck e Tutz, 2006] Einbeck, J. e Tutz, G. (2006). Modelling beyond regression functions: an application of multimodal regression to speed-flow data. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 55*, 461-475.

[Giné e Guillou, 2002] Giné, E. e Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics, 38*, 907–921.

[Haussler, 1995] Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean nn-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A, 69*, 217–223.

[Hyndman et al, 1996] Hyndman, R., Bashtannyk, D. e Grunwald, G. (1996). Estimating and Visualizing Conditional Densities. *Journal of Computational and Graphical Statistics, 5*, 315-336.

[Lee, 1989] Lee, M. (1989). Mode Regression. *Journal of Econometrics 42*, 337-349.

[R] R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing.*

[Van der Vaart, 1998] Van der Vaart, A. (1998). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics).* Cambridge University Press.

[Wand e Jones, 1995] Wand, M. e Jones, M. (1995). *Kernel Smoothing.* Chapman & Hall.

[Zhou e Huang, 2019] Zhou, H. e Huang, X. (2019). Bandwidth selection for nonparametric modal regression. *Communications in Statistics - Simulation and Computation, 48*, 968-984.