USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Predición electoral mediante promedios de enquisas

Javier Couselo Silveira

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Predición electoral mediante promedios de enquisas

Javier Couselo Silveira

Xullo, 2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Predición electoral mediante promedios de enquisas
Breve descrición do contido
Nos últimos anos os intentos por predicir os resultados electorais incrementáronse e cada vez hai máis empresas se dedican á predición electoral. Isto explica a gran cantidade de enquisas que hai actualmente, que sen embargo difiren nos seus resultados, xa que empregan mostras distintos, e axustan os datos a través de modelos diferentes, non necesariamente idénticos. Cabe destacar que, a pesar da súa discrepancia, a información que conteñen é moi relevante, xa que están baseadas en distintas mostras. Agregar a información que conteñen dunha forma consistente, permitiría incorporar toda esa información mostral nun único índice, que debería ser máis preciso, que as estimacións individuais. O obxectivo deste traballo é estudar como se pode predicir un resultado electoral a partir desta variedade de enquisas. Estudarase como facer unha boa predición mediante unha media local ponderada na que se lle asignará a cada unha das enquisas distintos pesos dependendo de diversos factores que serán estudados neste traballo tales como o tamaño da enquisa, a data na que se realizaron cada unha delas, a fiabilidade da empresa que fai a enquisa. Revisaranse os métodos empregados coñecidas e aplicaranse a un conxunto de datos reais.

Índice xeral

Resumo

Ante un escenario electoral xorde o interese de anticipar quen gañará e como será o reparto de votos resultante. Nas semanas e meses previos véñense facendo sondaxes por parte de diferentes casas de enquisas. Neste contexto, aparece a idea de agregar enquisas, isto é, considerar un conxunto amplo de enquisas feitas ata ese momento e usalas para facer un promedio que prediga o resultado o día das eleccións. Este será un promedio local que priorice as enquisas máis próximas ó día do que se quere facer a predición sobre aquelas que estean máis afastadas deste día.

Para este promedio local empregaremos un modelo estadístico de regresión non paramétrica que recibe o nome de regresión lineal local. Así, durante os tres primeiros capítulos farase un desenvolvemento teórico detallado deste modelo cubrindo os seus aspectos fundamentais, como poden ser o cálculo dos estimadores ou a elección do parámetro ventana que será o que determine como de local será o modelo. Finalmente, aplicarase todo este desenvolvemento teórico ó caso real das eleccións presidenciais estadounidenses do 3 de novembro do 2020.

Abstract

In the presence of an electoral scenario, an interest arises in anticipating who will win and how the resulting vote distribution will be. In the previous weeks and months, polls have been carried out by different survey companies. In this context, the idea of aggregating polls appears, that is, considering a broad set of polls completed up to that point and using them to obtain an average that predicts the result on election day. This will be a local average that prioritizes the polls closest to the day for which you want to make the prediction over those that are further away from this day.

For this local average, we will use a nonparametric regression statistical model that is called local linear regression. Thus, during the first three chapters a detailed theoretical development of this model will be made, covering its fundamental aspects, such as the calculation of the estimators or the choice of the window parameter that will determine how local the model will be. Finally, all of this theoretical development will be applied to the real case of the US presidential elections of the third of November 2020.

Introdución

Historicamente, o ser humano sempre intentou anticiparse ós acontecementos e predicilos para poder garantirse o éxito. No contexto social, unha das principais formas de conseguilo é coñecer a opinión que ten a poboación sobre certos ámbitos, xa que permite conxecturar en que dirección vai avanzar a sociedade. Un propósito que non fixo máis ca aumentar debido ás cada vez maiores ansias de conseguir o éxito e saber en que sentido fixar decisións estratéxicas.

Anticipar e coñecer o que o conxunto da poboación pensa é o que veñen a ofrecer as enquisas, unha ferramenta clave e eficaz para pescudar a opinión de case calquera tipo de poboación sobre practicamente calquera ámbito da realidade social, económica, política,... Partindo dunha mostra considerablemente menor ca poboación da que se quere saber a opinión, as sondaxes obteñen grandes éxitos na predición de sucesos e na anticipación a eles.

Este crecente afán de adiantarse ós acontecementos deriva nun gran aumento da demanda destes servizos e nun constante perfeccionamento, convertendo as enquisas nunha ferramenta cada vez máis fidedigna. Isto explica o seu auxe no ámbito político e tamén, no empresarial, onde conseguen chegar mellor ó cliente mellorando o servizo e a atención que recibe o consumidor, desembocando nunha maior produtividade, eficacia e competitividade.

O uso das enquisas esténdese a un amplo número de ámbitos, destacando nitidamente no eido da política, onde se explotan ata ser consideradas habitualmente como un índice real da aceptación que van tendo os partidos e as figuras políticas. Por exemplo, fóra de épocas electorais para saber como son valoradas as estratexias de diferentes formacións políticas. Porén, é evidente que o momento no que máis se emprega este recurso é nos períodos electorais, que é precisamente o contexto no que se desenvolve este traballo.

Motivados por coñecer de antemán os resultados electorais, xorde a idea de axustar un modelo que sexa capaz de dar ca intención real de voto empregando as sondaxes publicadas.

As diferentes enquisas difiren nos seus resultados, xa que son axustadas de maneiras diferentes e conteñen mostras distintas. Precisamente, este último motivo fai que a información de cada unha de estas enquisas sexa relevante, xa que agregando as diferentes mostras conseguimos empregar o coñecemento dispoñible de xeito eficiente, corrixindo así os posibles nesgos que cada enquisa de xeito individual poida ter. Así, perfeccionamos a estimación mediante un promedio que involucre a varias sondaxes, pois tense visto que os promedios de varias enquisas se achegan máis á realidade de voto que cada unha delas individualmente, ó estar baseados en información provinte de varias mostras. Isto analízase en Graefe (2015), onde se conclúe que a agregación de enquisas consegue pronósticos electorais máis precisos ca outros métodos de predición electoral.

O noso obxectivo é estimar a partir das enquisas dispoñibles, a porcentaxe de votos para certo candidato; concretamente, orientaremos este traballo de cara a estimar a porcentaxe de votos para Biden nas eleccións presidenciais estadounidenses do 2020. A razón desta escolla é que foron os comicios que se celebraron nos primeiros meses de elaboración deste traballo. Non obstante, poderíase aplicar a mesma metodoloxía en calquera outro proceso electoral con máis forzas políticas.

Con este fin, plantexarei un modelo de regresión lineal local despois de ter vistas as limitacións que nos levan a apartarnos do modelo lineal simple e farei un análise profundo dos estimadores lineais locais e da relación que gardan ca regresión lineal simple. Así mesmo, abordarei o tema da elección do parámetro ventana, que controlará cantas observacións interveñen nos promedios locais e presentarei unha forma de escribir o modelo lineal local dun xeito computacionalmente máis eficiente, reducindo os tempos de execución.

Finalmente facendo uso do software estadístico R (R Core Team, 2021), concretamente a versión 4.0.4 subida o día 15 de febreiro do 2021, aplicarei todo o desenvolvemento teórico ás eleccións de EEUU, en concreto a varios estados que eran decisivos ó non ter clara a intención de voto en ditos comicios. Así, partindo dunha ampla cantidade de sondaxes levadas a cabo en distintas datas preelectorais, estimaremos a intención de voto o día das eleccións.

En canto á estrutura do traballo, no Capítulo 1 revisaremos as nocións máis importantes da regresión lineal simple, especialmente a expresión do estimador, así como as súas limitacións. Estas últimas levarannos a plantexarnos no Capítulo 2 un modelo lineal local, do que faremos un estudo exhaustivo (parámetro ventana, función kernel, axustes, predición,...). No Capítulo 3 afondaremos sobre o transcendental papel do parámetro ventana na regresión lineal local analizando con detalle a súa influencia sobre o nesgo e a varianza. Por último, no Capítulo 4 aplicaremos o modelo lineal local desenvolvido ó longo do traballo ó caso real das eleccións presidenciais de EEUU do 2020 empregando R .

Capítulo 1

Regresión Lineal Simple

A **análise de regresión** é unha ferramenta estadística que permite estudar a dependencia entre distintas variables, é dicir como inflúen unhas variables que reciben o nome de variables explicativas noutra variable que chamaremos variable resposta. O propósito de axustar un modelo de regresión pode ser coñecer a dependencia ou influencia que teñen as variables explicativas sobre a variable resposta ou ben, obter predicións da variable resposta para novos valores das variables explicativas.

A regresión estadística ten dúas vertentes, a **regresión paramétrica**, que engloba os modelos nos que a función de regresión toma unha forma determinada e a **regresión non paramétrica**, onde pola contra, a función de regresión do modelo non ten unha forma predefinida, senón que se constrúe a partir dos datos partindo soamente de hipóteses de suavidade (no sentido de continuidade e diferenciabilidade).

Ademais, interesarémosnos en concreto pola **regresión univariante**, é dicir queremos explicar unha variable resposta Y a partir dunha única variable explicativa X. Empezarei presentando o modelo de regresión lineal simple (que é un modelo paramétrico) para posteriormente comparalo ca regresión lineal local (que se enmarca na estadística non paramétrica) e facer un estudo profundo deste segundo modelo.

No contexto da regresión univariante, pretendemos axustar:

$$Y = m(x) + \varepsilon,$$

onde m é unha función real de variable real descoñecida que recibe o nome de **función de regresión**, que é a que queremos estimar, e onde denotamos o **erro** por ε, que vén sendo a variabilidade da variable resposta Y que o modelo non é capaz de explicar a partir da variable explicativa X. Asumiremos que este erro ten media cero.

A función de regresión m non é máis có valor esperado da variable Y para certo valor da covariable $X = x$; é dicir, a esperanza de Y condicionada a $X = x$:

$$m(x) \;=\; \mathbb{E}\left[Y|X = x\right].$$

Polo tanto, dado un conxunto de n observacións das variables explicativa e resposta, $\{(x_i, Y_i)\}_{i=1}^n$, o **modelo de regresión** que queremos axustar é o seguinte:

$$Y_i \;=\; m(x_i) \;+\; \varepsilon_i, \qquad\qquad i = 1, \dots, n,$$

onde x_i e Y_i son as i-ésimas observacións da variables explicativa e resposta respectivamente e ε_i denota o erro asociado á observación i-ésima.

No caso que nos ocupa, o de estudar a pocentaxe de voto para Biden en certos estados nas eleccións de EEUU, para cada observación $i \in \{1, \dots, n\}$, o valor da variable explicativa x_i fará referencia á data na que se publicou a enquisa, en concreto será, como explicarei no último capítulo, o número de días que pasaron dende o día no que se fixo a primeira enquisa dese estado, considerando a este como o día 1. Por outro lado, o valor da variable resposta Y_i será a estimación da porcentaxe de votos que dita enquisa lle outorga a Biden.

Empezamos plantexando un modelo de regresión lineal simple.

1.1. Formulación da regresión lineal simple

Comezámosnos interesando por un modelo de regresión lineal simple, xa que ó ter unha estrutura determinada, neste caso unha recta, resúltanos atractivo para manternos nunha complexidade o máis baixa posible e cunha interpretación dos resultados moi clara.

Aínda que só trataremos algúns contidos básicos da regresión lineal simple, para afondar neste tema pódese consultar o Capítulo 5 de Peña (2002).

Na regresión lineal simple, asúmese que a función de regresión toma a forma:

$$m(x) \;=\; \beta_0 + \beta_1 \cdot x, \tag{1.1}$$

polo tanto, o modelo de regresión lineal simple consiste en axustar do seguinte xeito:

$$Y_i \;=\; \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \qquad\qquad i \in \{1, \dots, n\},$$

onde $\beta_0, \beta_1 \in \mathbb{R}$ son os parámetros descoñecidos a estimar, que reciben o nome de **intercepto** e **pendente** respectivamente. Como xa indiquei anteriormente, ε_i representa o erro asociado á observación i-ésima.

Deste xeito, é evidente que para poder aplicar un modelo lineal simple se ten que cumprir a hipótese de linealidade, é dicir, o modelo debe ter a forma especificada en (1.1). Non obstante, en moitos casos isto non é así, e polo tanto teremos que recorrer a outro tipo de modelos. Nese suposto, pódese considerar un modelo lineal local, xa que ofrece unha maior flexibilidade e vaise adaptando localmente ós datos, como veremos no Capítulo 2.

En todo caso, o axuste lineal local, que é a nosa finalidade, pretenderá mellorar o axuste lineal simple polo que é conveniente revisar os conceptos básicos desta metodoloxía, especialmente no tocante á estimación dos parámetros por mínimos cadrados.

1.2. Estimación dos parámetros da regresión lineal simple

Para estimar os parámetros intercepto e pendente, habitualmente emprégase o **método de mínimos cadrados**, que consiste en minimizar a suma de residuos ó cadrado (RSS, segundo as súas siglas en inglés: *Residual Sum Squares*), que veñen sendo os cadrados das diferenzas entre os valores da resposta e as predicións:

$$\text{RSS} \;=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \;=\; \sum_{i=1}^{n} (Y_i - a - b \cdot x_i)^2,$$

sendo $\hat{Y}_i = a + b \cdot x_i$. Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ son os que resolven o problema de minimización:

$$\min_{a,b \in \mathbb{R}} \text{RSS} \;=\; \min_{a,b \in \mathbb{R}} \sum_{i=1}^{n} e_i^2 \;=\; \min_{a,b \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - a - b \cdot x_i)^2. \tag{1.2}$$

Entón, a nosa finalidade é minimizar a función obxectivo:

$$\begin{aligned} \phi :\; \mathbb{R}^2 &\longrightarrow \mathbb{R}, \\ (a,b) &\rightsquigarrow \phi(a,b) \;:=\; \sum_{i=1}^{n} (Y_i - a - b \cdot x_i)^2 . \end{aligned} \tag{1.3}$$

Os estimadores da regresión lineal simple serán precisamente os mínimos da función obxectivo. No seguinte teorema expoño a súa expresión analítica.

Teorema 1.1. *Os estimadores da regresión lineal simple $\hat{\beta}_0$ e $\hat{\beta}_1$ teñen as seguintes expresións que se corresponden co mínimo absoluto da función obxectivo ϕ dada en (1.3):*

$$\hat{\beta}_0 \;=\; \overline{Y} - \frac{S_{xY}}{S_x^2} \cdot \overline{x}, \qquad\qquad \hat{\beta}_1 \;=\; \frac{S_{xY}}{S_x^2},$$

sendo $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as medias mostrais das variables explicativa e resposta respectivamente, $S_{xY} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})\left(Y_i - \overline{Y}\right)$ a covarianza entre as variables x e Y e $S_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$ a varianza da variable explicativa x.

Proba. Como os parámetros $\hat{\beta}_0$ e $\hat{\beta}_1$ son os que minimizan a función obxectivo ϕ dada en (1.3), calculamos as derivadas parciais da mesma:

$$\begin{aligned}\frac{\partial \phi}{\partial a}(a,b) &= -2 \cdot \sum_{i=1}^{n}(Y_i - a - b \cdot x_i),\\ \frac{\partial \phi}{\partial b}(a,b) &= -2 \cdot \sum_{i=1}^{n}(Y_i - a - b \cdot x_i) \cdot x_i.\end{aligned}$$

Para obter os estimadores de mínimos cadrados debemos resolver o sistema resultante de igualar ambas ecuacións a cero, coñecidas como **ecuacións normais da regresión**:

$$\left\{\begin{array}{rcl} \displaystyle\sum_{i=1}^{n} Y_i &=& n \cdot a + b \cdot \displaystyle\sum_{i=1}^{n} x_i,\\ \displaystyle\sum_{i=1}^{n} x_i \cdot Y_i &=& a \cdot \displaystyle\sum_{i=1}^{n} x_i + b \cdot \displaystyle\sum_{i=1}^{n} x_i^2. \end{array}\right.$$

Se dividimos ambas ecuacións entre o número de datos n, poderemos reescribir algúns termos en función das medias mostrais, $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, chegando así ó sistema:

$$\left\{\begin{array}{rcl} \overline{Y} &=& a + b \cdot \overline{x},\\ \dfrac{\sum_{i=1}^{n} x_i \cdot Y_i}{n} &=& a \cdot \overline{x} + b \cdot \dfrac{\sum_{i=1}^{n} x_i^2}{n}. \end{array}\right. \tag{1.4}$$

Se agora consideramos agora a diferenza da segunda ecuación menos a primeira multiplicada pola media mostral de x, $\overline{x}$, chegamos a seguinte expresión:

$$\frac{\sum_{i=1}^{n} x_i \cdot Y_i}{n} - \overline{Y} \cdot \overline{x} = b \cdot \left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 \right). \tag{1.5}$$

O termo da esquerda da igualdade correspóndese ca covarianza entre x e Y, S_{xY}:

$$\begin{aligned} S_{xY} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i \cdot Y_i - x_i \cdot \overline{Y} - \overline{x} \cdot Y_i + \overline{x} \cdot \overline{Y})\\ &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i \cdot Y_i\right) - \overline{x} \cdot \overline{Y} - \overline{x} \cdot \overline{Y} + \overline{x} \cdot \overline{Y} = \frac{\sum_{i=1}^{n} x_i \cdot Y_i}{n} - \overline{x} \cdot \overline{Y}. \end{aligned} \tag{1.6}$$

Por outro lado, o termo que multiplica a b en (1.5) é a varianza de x, S_x^2:

$$\begin{aligned} S_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2-2\cdot x_i\cdot\overline{x}+\overline{x}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right)-2\cdot\overline{x}^2+\overline{x}^2 \\ &= \frac{\sum_{i=1}^{n}x_i^2}{n}-\overline{x}^2. \end{aligned} \tag{1.7}$$

Deste xeito, tendo en conta (1.6) e (1.7), reescribimos a expresión (1.5) como segue:

$$S_{xY} = b\cdot S_x^2,$$

e así, obtense a expresión do estimador $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{S_{xY}}{S_x^2}.$$

Se agora despexamos a na primeira ecuación de (1.4) obtemos a seguinte expresión:

$$a = \overline{Y}-b\cdot\overline{x},$$

que nos permite obter a expresión para $\hat{\beta}_0$:

$$\hat{\beta}_0 = \overline{Y}-\frac{S_{xY}}{S_x^2}\cdot\overline{x}.$$

Así temos que $\left(\hat{\beta}_0,\hat{\beta}_1\right)$ é un punto crítico da función obxectivo ϕ. Queremos comprobar que é un mínimo, para o cal recorremos ás segundas derivadas e á matriz hessiana:

$$\begin{aligned} \frac{\partial^2\phi}{\partial a^2}(a,b) &= -2\cdot\sum_{i=1}^{n}(-1) = 2\cdot n, \\ \frac{\partial^2\phi}{\partial b^2}(a,b) &= -2\cdot\sum_{i=1}^{n}-x_i^2 = 2\cdot\sum_{i=1}^{n}x_i^2, \\ \frac{\partial^2\phi}{\partial a\partial b}(a,b) &= \frac{\partial^2\phi}{\partial b\partial a}(a,b) = 2\cdot\sum_{i=1}^{n}x_i = 2\cdot n\cdot\overline{x}. \end{aligned}$$

Polo tanto a matriz hessiana é:

$$\mathcal{H} = \begin{pmatrix} \frac{\partial^2\phi}{\partial a^2}(a,b) & \frac{\partial^2\phi}{\partial a\partial b}(a,b) \\ \frac{\partial^2\phi}{\partial b\partial a}(a,b) & \frac{\partial^2\phi}{\partial b^2}(a,b) \end{pmatrix} = \begin{pmatrix} 2\cdot n & 2\cdot n\cdot\overline{x} \\ 2\cdot n\cdot\overline{x} & 2\cdot\sum_{i=1}^{n}x_i^2 \end{pmatrix}.$$

Tense que ambos os dous menores principais son estritamente positivos, xa que por unha parte, $|2n|>0$ e por outra $|\mathcal{H}|>0$ como comprobaremos nas vindeiras liñas.

Así, o determinante da matriz hessiana é:

$$|\mathcal{H}| \;=\; \left|\begin{pmatrix} 2\cdot n & 2\cdot n\cdot\overline{x} \\ 2\cdot n\cdot\overline{x} & 2\cdot\sum_{i=1}^{n} x_i^2 \end{pmatrix}\right| \;=\; 4\cdot n\cdot\sum_{i=1}^{n} x_i^2 - 4\cdot n^2\cdot\overline{x}^2 \;=\; 4n^2\cdot\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2\right).$$

Pero, o termo entre parénteses correspóndese ca varianza mostral de x, S_x^2, como vimos en (1.7), polo que o determinante da matriz hessiana é estritamente positivo:

$$|\mathcal{H}| \;=\; 4\cdot n^2\cdot S_x^2 \;>\; 0.$$

Efectivamente, a varianza mostral de x, $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$, é estritamente positiva por ser suma de termos ó cadrado (non negativos) e non todos nulos, xa que nese caso todos os valores da variable explicativa serían identicamente iguais, o cal non ocorre en ningún conxunto de datos que interese estudar.

Como os dous menores principais da matriz $\mathcal{H}$ son estritamente positivos, en virtude do criterio de Sylvester, a función obxectivo ϕ alcanza un mínimo relativo en $\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

É o único mínimo relativo e ademais cando algún dos parámetros tende a $\pm\infty$, ϕ tende a ∞; co cal, efectivamente $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ é un mínimo absoluto da función obxectivo ϕ. □

A continuación veremos que aínda que a regresión lineal simple conta con grandes vantaxes en modelos nos que procede (como por exemplo unha expresión e interpretación sinxela), hai outros moitos modelos que non é capaz de axustar.

1.3. Fortalezas e debilidades da regresión lineal simple

Como acabo de adiantar, a regresión lineal simple é incapaz de axustar un amplo número de modelos. Nesta sección empezarei presentando un exemplo no que a regresión lineal simple axusta moi ben a relación entre as variables para posteriormente expoñer outro que no que non é así, pero que pola contra, a regresión lineal local si o fai sen ningún problema.

Comezaremos traballando co conxunto de datos `trees`, do paquete `datasets`, que ven cargado por defecto en R (R Core Team, 2021) e que contén observacións sobre o diámetro (en polgadas), altura (en pés) e o volume (en pés cúbicos) do tronco de distintas cereixeiras. Propoñémosnos saber o volume do tronco a partir do seu diámetro; así que axustamos un modelo lineal simple co comando `lm` e que representamos ca axuda de `plot` e `abline`, obtendo a recta axustada da Figura 1.1.

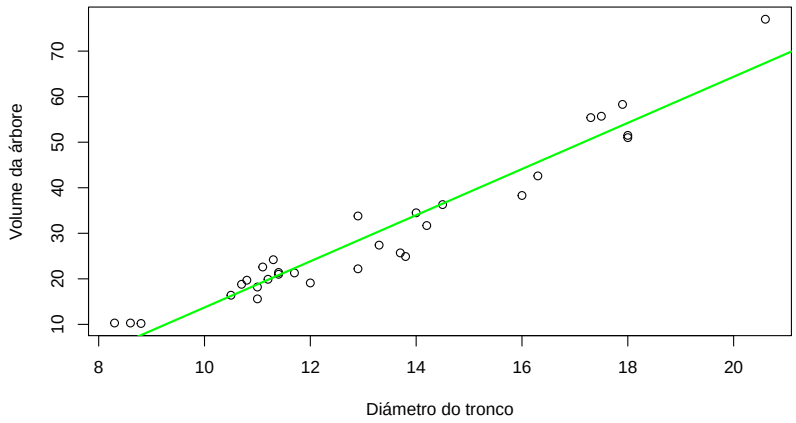


Figura 1.1: Axuste lineal dos datos de `trees` explicando `Volume` (eixo y) en función de `Diámetro` (eixo x).

Apreciamos na Figura 1.1 que a regresión lineal simple axusta perfectamente este modelo. De feito, o coeficiente de determinación, que é unha medida da bondade do axuste, vale 0,935 sobre un máximo de 1. Neste caso, quere dicir que o modelo é capaz de explicar nun 93,5 % a variabilidade da variable resposta a partir da explicativa, o cal está moi ben.

Facendo uso do comando `coefficients` de  obteño os coeficientes do modelo de regresión, que quedaría escrito como segue:

$$\texttt{Volume} \;=\; -36{,}94 \;+\; 5{,}07 \cdot \texttt{Diámetro}.$$

A interpretación do intercepto non procede, xa que se correspondería con árbores de nulo diámetro de tronco. Porén, o valor da pendente interprétase como segue: por cada unidade que "aumenta" o diámetro do tronco, o seu volume faino en 5,07 unidades, o que explica a pendente positiva que apreciamos na Figura 1.1. Esta sinxela expresión, que permite unha interpretación directa e clara, é unha das vantaxes da regresión lineal simple, que ademais non precisa dunha gran cantidade de datos para realizar un bo axuste.

Non obstante, hai modelos que a regresión lineal simple é incapaz de axustar. Ilustrareino co conxunto de datos `economics` da libraría `ggplot2` (Wickham, 2016), que contén variables de carácter económico para os meses entre xullo de 1967 e abril do 2015. Interesámosnos por explicar a evolución temporal mensual da súa variable `uempmed`, que nos dá a duración media (en semanas) do desemprego en cada mes. Para que se vexa con máis claridade só considerarei as 120 primeiras observacións. Con un modelo lineal simple, o axuste que obtemos para `uempmed` en función da data é o que represento na Figura 1.2, que como vemos é pésimo.

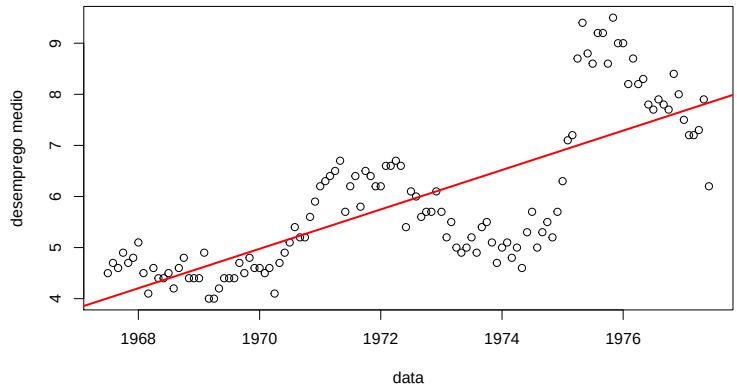


Figura 1.2: Axuste lineal simple dos datos de `economics` explicando `uempmed` (eixo y) en función de `data` (eixo x).

Neste caso o modelo non é lineal e iso explica que a regresión lineal simple non sexa capaz de axustalo, pois non é posible aproximar cunha única recta este modelo. Non obstante, acudindo á regresión lineal local obtemos un bo axuste que si se adapta ós datos, tal e como se aprecia na Figura 1.3, na que ilustro este axuste lineal local.

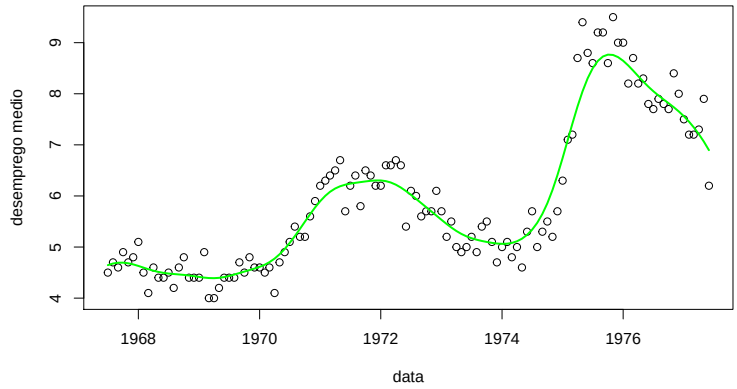


Figura 1.3: Axuste lineal local dos datos de `economics` explicando `uempmed` (eixo y) en función de `data` (eixo x).

A razón pola cal este último axuste ilustrado na Figura 1.3 si é bo vai dada no seu nome: é un axuste local. O modelo lineal local non conta cunha estrutura determinada, senón que se adapta ós datos, acomodándose en cada valor da variable explicativa á información que localmente ten, é dicir, ás observacións máis próximas, como podemos apreciar nesta figura.

En conclusión, o modelo lineal simple, cando procede, é unha moi boa elección debido á súa baixa complexidade e a súa fácil comprensión e interpretación. Non obstante, non é capaz de axustar modelos máis complexos que si se poden axustar cun modelo lineal local, polo feito de non partir dunha forma específica. Vexamos entón a regresión lineal local.

Capítulo 2

Regresión Lineal Local

Para o plantexamento deste modelo baseeime nos libros de Fan e Gibels (1996) e Wand e Jones (1995) nos que se trata este modelo con gran detalle.

Recordemos que a nosa finalidade é estimar a función de regresión m tal que:

$$Y \;=\; m(x) \;+\; \varepsilon. \tag{2.1}$$

Mentres que, como acabamos de ver, a regresión lineal consiste en atopar un par de parámetros β_0, $\beta_1 \in \mathbb{R}$ tales que $m(x) = \beta_0 + \beta_1 \cdot x$, na regresión lineal local non suporemos ningunha forma especial á función de regresión, simplemente nos limitaremos a traballar baixo a hipótese de que a función de regresión m é suave no sentido que precisaremos máis adiante e axustando localmente esta función de regresión de xeito que:

$$m(x) \;=\; \mathbb{E}\left[Y|X = x\right],$$

de onde se segue que o erro que aparece (2.1) ten esperanza cero, $\mathbb{E}\left[\varepsilon\right] = 0$.

Non obstante, na práctica o que estaremos estimando non será exactamente $m(x)$ senón un promedio local dos $m(x_i)$, pero isto será ilustrado con suficiente detalle máis adiante.

Así, unha vez **fixado** x_0, asumiremos que a función de regresión localmente é da forma:

$$m(x) \;=\; \beta_0(x_0) + \beta_1(x_0) \cdot x + e, \qquad \text{para } x \text{ nunha veciñanza de } x_0,$$

sendo e un residuo que, como veremos, non ten esperanza nula.

Para conseguir este axuste faremos un promedio local ponderado no que consideraremos só os datos máis próximos ó punto de interese x_0, restrinxindo o promedio a unha veciñanza cuxo tamaño virá determinado polo parámetro ventana h. Este promedio local será ademais ponderado, é dicir priorizaremos as observacións máis proximas ó dato x_0, facendo uso dunha función kernel que será a encargada de facer esas asignacións de pesos.

2.1. Significado dos parámetros $\beta_0(x_0)$ e $\beta_1(x_0)$

Nesta sección abordarase a interpretación dos parámetros $\beta_0(x_0)$ e $\beta_1(x_0)$, así como a súa relación ca aproximación lineal que se considera localmente con este modelo.

Se asumimos que m é suave nunha veciñanza de x_0 (que xa fixamos na páxina anterior), podemos considerar a aproximación de Taylor de grao 1 ó redor do punto x_0:

$$m(x) \simeq m(x_0) + m'(x_0) \cdot (x - x_0). \tag{2.2}$$

Agora, se desenvolvemos o produto obtemos o seguinte:

$$m(x) \simeq \underbrace{m(x_0) - m'(x_0) \cdot x_0}_{\beta_0(x_0)} + \underbrace{m'(x_0)}_{\beta_1(x_0)} \cdot x. \tag{2.3}$$

Entón pretendemos obter o noso estimador localmente como:

$$\hat{m}(x) = \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0) \cdot x. \tag{2.4}$$

Como vemos o termo constante de (2.3) correspóndese con $\beta_0(x_0)$ mentres que $\beta_1(x_0)$ ven sendo a derivada da función de regresión no punto x_0:

$$\begin{cases} \beta_0(x_0) = m(x_0) - m'(x_0) \cdot x_0, \\ \beta_1(x_0) = m'(x_0). \end{cases}$$

O valor de $\beta_0(x_0)$ correspóndese ca estimación da variable resposta para o valor da explicativa nulo, o cal non terá interpretación no noso caso, xa que para nos a variable x fará referencia á data das enquisas, e será estritamente positiva. Pola contra, o parámetro $\beta_1(x_0)$ si que ten unha interpretación, pois é a derivada da función de regresión no punto x_0, co cal permitirá saber a tendencia crecente ou decrecente da función de regresión en x_0 e cuantificar esa tendencia.

Agora ben, debemos ser conscientes de que aínda que a función de regresión m si se pode aproximar localmente como se mostra en (2.2), dende o momento no que aceptamos o estimador $\hat{m}$ dado en (2.4), estamos renunciando a ser capaces de explicar certa variabilidade, xa que pretendemos aproximar localmente m por un polinomio de grao 1.

Isto terá como consecuencia que non estaremos estimando a propia función de regresión, senón un promedio local da mesma avaliada nos valores observados da variable explicativa, $m(x_i)$, como veremos no Capítulo 3.

Reflexionando un pouco máis nesta dirección, dámosnos de conta de que o erro que estamos cometendo na aproximación de Taylor de grao 1 presentada en (2.2) é precisamente o dado polo Teorema de Taylor:

$$\frac{m''(\xi)}{2} \cdot (x - x_0)^2,$$

onde ξ é un punto que se atopa entre x e x_0. Nótese entón que estamos supoñendo que m suave significa dúas veces derivable.

Neste erro aparece a segunda derivada, que provocará que o estimador lineal local teña dificultades para modelar rexións cunha curvatura pronunciada, en particular nos máximos e mínimos. Ilustrarei este feito con detalle na Sección 3.2.

Así, teremos este erro debido á aproximación lineal que consideramos en (2.2) e tamén o erro asociado á propia observación dos datos mostrais que viñamos denotando por ε. Englobamos ambos baixo un residuo que denotaremos por e.

Polo tanto, partindo dun conxunto de observacións $\{x_i, Y_i\}_{i=1}^n$, o axuste que faremos nunha veciñanza de x_0 será:

$$Y_i \;=\; \beta_0(x_0) + \beta_1(x_0) \cdot x_i + e_i, \qquad \text{se } x_i \text{ está nunha veciñanza de } x_0, \tag{2.5}$$

onde e_i xa non é o erro de esperanza cero que denotabamos por ε_i, senón un residuo que deixa de ter esperanza nula, pois como acabo de indicar, ademais deste erro de observación ε_i tamén está involucrado o correspondente a asumir en (2.2) a aproximación lineal dunha función que non o é.

A aproximación (2.2) será mellor canto máis cerca estea x de x_0. Isto lévanos a plantexarnos o papel do parámetro ventana, que regula o tamaño da veciñanza, sendo así determinante na regresión lineal local, polo que trataremos este tema na seguinte sección.

De cara a aplicar no Capítulo 4 o modelo ó caso real das eleccións estadounidenses do 2020, recordo que para nós a variable explicativa X fará referencia á data da enquisa, é dicir, na expresión (2.5), x_i será o día no que se fai a enquisa i. Do mesmo xeito, como a variable Y se refire á porcentaxe de votos de certo candidato (en concreto tomaremos o de Biden), os termos Y_i de (2.5) denotarán a porcentaxe de votos que este candidato recibiría de acordo ca enquisa i-ésima.

2.2. Papel do parámetro ventana

Aquí entramos a tratar unha cuestión de gran importancia nun modelo lineal local, o parámetro ventana, ou parámetro ancho de banda, h. A súa función é controlar a amplitude do intervalo que estamos considerando para a regresión lineal local.

O modelo lineal local tal e como indica o seu nome baséase nun axuste local, é dicir, un axuste no que consideramos as observacións dunha veciñanza do punto para o que queremos obter a estimación. Xorden así unha cuestións de máxima importancia: como de grande queremos a veciñanza? cantos datos queremos empregar para o noso promedio local?

Pois ben, o parámetro ventana vén a precisar estas cuestións, xa que a súa función será determinar a extensión da veciñanza na que consideramos as observacións. Deste xeito, propoñémosnos estimar nunha veciñanza de x_0 do seguinte xeito:

$$\hat{m}(x) \;=\; \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0)\cdot x, \qquad x \in (x_0 - h, x_0 + h), \qquad \text{con } h > 0.$$

Entón, reescribimos o axuste (2.5) como segue:

$$Y_i \;=\; \beta_0(x_0) + \beta_1(x_0)\cdot x_i + e_i, \qquad \text{para todo } i \text{ tal que } x_i \in (x_0 - h, x_0 + h). \quad (2.6)$$

Así, neste modelo estamos a considerar só as observacións tales que $x_i \in (x_0-h, x_0+h)$, que podemos reescribir como aquelas para as que o cociente da súa distancia a x_0 entre a ventana h sexa menor cá unidade en valor absoluto:

$$x_i \in (x_0 - h, x_0 + h) \quad \Longleftrightarrow \quad |x_i - x_0| < h \quad \Longleftrightarrow \quad \frac{|x_i - x_0|}{h} < 1. \qquad (2.7)$$

Deste xeito só estaríamos considerando as observacións cuxa variable explicativa caia dentro do intervalo $(x_0 - h, x_0 + h)$. Isto implica que se unha observación ten o seu valor da explicativa fóra dese intervalo non se terá en conta, ignorarase para a predición en x_0. Tamén no caso de que o valor da explicativa estea moi cerca de caer dentro do intervalo pero non o faga.

Decatémosnos ademais de que o parámetro ventana h está intimamente relacionado ca aproximación de Taylor, pois como controla a amplitude do intervalo, se toma un valor moi grande a aproximación de Taylor de m será deficitaria, pero se toma un valor menor do debido, quédanos un modelo moi inestable, xa que a estimación en x_0 estará baseada en moi poucos datos, incrementando a variabilidade da aproximación, tal e como veremos. Isto levaranos a unha busca do parámetro parámetro ventana óptimo que consiga un equilibrio entre ambas as dúas situacións, cuestión que abordemos en detalle no seguinte capítulo.

Así, para axustar o modelo (2.6), e tendo en conta a condición (2.7) para os puntos que estamos a considerar, formalizamos o seguinte **problema de mínimos cadrados**:

$$\min_{a,b\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - a - b \cdot x_i)^2 \cdot \mathbb{I}\left[\frac{|x_i - x_0|}{h} < 1\right], \tag{2.8}$$

onde $\mathbb{I}[\cdot]$ toma valor 1 se se cumpre a condición entre corchetes e 0 noutro caso.

Non obstante, estamos considerando todos os puntos do entorno $(x_0 - h, x_0 + h)$ igual de importantes e isto é claramente mellorable, pois m para os puntos máis próximos a x_0 tenderá a parecerse máis a $m(x_0)$ cos puntos máis afastados.

Para o caso que nos ocupa, o das enquisas das eleccións estadounidenses (ás que aplicaremos todo o desenvolvemento teórico no Capítulo 4), como a variable x recolle os días nos que se fai cada enquisa, a ventana h será a que determine a distancia para coller ou non as observacións, determinando o rango de días considerados. Así, no problema de mínimos cadrados que acabamos de plantexar en (2.8), entrarían as observacións que verifican $x_i \in (x_0 - h, x_0 + h)$, e dicir as enquisas feitas os anteriores h días e os posteriores h días.

2.3. A función kernel

Queda clara a substancial importancia do parámetro ventana h, que non é outra ca controlar as observacións que caen no intervalo que consideramos para facer a regresión lineal local, restrinxíndonos a esa veciñanza. Pero ademais, non ten demasiado sentido considerar todos eses puntos por igual, senón que os puntos máis próximos terán unha maior similitude ó punto no que queremos estimar que os puntos máis afastados. Isto lévanos a asignarlles tanto máis peso canto máis cerca estean do punto x_0. Para iso, faremos uso das coñecidas como funcións kernel K, que explicarei nesta sección.

Unha **función kernel** K é unha función real de variable real non negativa e integrable, que asigna pesos ás diferentes observacións dunha mostra. Son frecuentemente empregadas en estadística non paramétrica. Nos usaremos kernels simétricos respecto á orixe.

Na práctica non son máis ca funcións de densidade simétricas unimodais no 0, pois integran a unidade:

$$\int K(u)\, du \;=\; 1 \qquad \text{(Normalización)}.$$

Ben, presentada a función kernel, prosigamos ca formulación do modelo.

Lembremos que na regresión lineal simple para obter os parámetros recorríamos ó problema de mínimos cadrados (1.2). Pois ben, agora seguimos interesados en minimizar os erros, pero a diferenza do caso lineal simple e tal e como veño explicando, nin queremos ter en conta todas as observacións, nin tampouco pretendemos ponderalas igual. Para conseguir esta ponderación que premia ós puntos máis próximos, faremos uso dunha función kernel, K, que asinará pesos ás observacións de acordo ca súa proximidade ó punto x_0.

Retomando a expresión (2.6), despexamos o residuo asociado á observación i-ésima:

$$e_i \;=\; Y_i - \beta_0(x_0) - \beta_1(x_0)\cdot x_i, \qquad\qquad x_i \in (x_0-h, x_0+h). \tag{2.9}$$

Pero como veño comentando, ademais de restrinxirnos só as observacións tales que $x_i \in (x_0-h, x_0+h)$, tamén queremos priorizar aquelas observacións cuxo valor da explicativa estea máis preto de x_0. Así, reformulamos o problema de minimización plantexado en (2.8) pero agora multiplicando o cadrado do residuo e_i pola ponderación que lle outorga a función kernel. Quédanos entón o seguinte **problema de mínimos cadrados ponderados**:

$$\min_{a,b\in\mathbb{R}} \quad \sum_{i=1}^{n} \{Y_i - a - b\cdot x_i\}^2 \cdot K_h(x_i - x_0), \tag{2.10}$$

onde a función K_h, que se soe usar para maior comodidade, defínese do seguinte xeito:

$$K_h(\cdot) \;=\; K\left(\frac{\cdot}{h}\right) \qquad\qquad h>0.$$

O uso de K_h responde a unha escritura máis limpa, pois así, o peso para a observación i ven dado por $K_h(x_i - x_0)$, que tamén se podería escribir sen usar K_h como $K\left(\frac{x_i-x_0}{h}\right)$.

A modo de observación, nótese que K_h non é unha función de densidade, pero $\frac{1}{h}K_h$ si que o é, como se pode apreciar facendo uso do cambio de variable $v=\frac{u}{h}$, $dv=\frac{du}{h}$:

$$\int \frac{1}{h}K_h(u)\; du \;=\; \int \frac{1}{h}K\left(\frac{u}{h}\right)\; du \;=\; \int K(v)\; dv = 1.$$

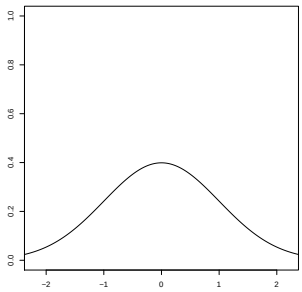
Insisto en que dentro do sumatorio (2.10), o termo que vai entre chaves elevado ó cadrado sería o residuo asociado á observación i-ésima, que sinalei en (2.9), cuxo valor non podemos saber, xa que $\beta_0(x_0)$ e $\beta_1(x_0)$ son descoñecidos. A continuación vai a ponderación mediante a función kernel, e así conseguimos que as observacións máis próximas teñan máis importancia ca aquelas que están máis afastadas.

A continuación farei un repaso das **funcións kernel máis habituais**. Indicarei as súas fórmulas e dominios xunto con unha representación gráfica.

En primeiro lugar, temos o kernel de Gauss (que ven dado pola función de densidade dunha normal estándar), o kernel de Epanechnikov (que é óptimo no sentido do erro cadrático medio no problema de estimación da densidade) e o kernel uniforme (que asina igual peso a todas as observacións do intervalo $(x_0 - h, x_0 + h)$). Nótese que empregar o kernel uniforme lévanos a resolver localmente un problema de mínimos cadrados (sen ponderar), tal e como se sinalaba en (2.8). A continuación mostro as expresións destes tres kernels xunto ca súa representación gráfica na Figura 2.1.

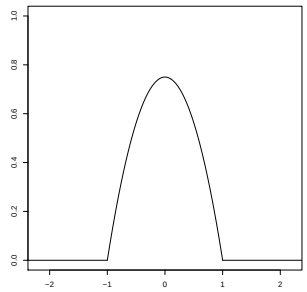
Kernel de Gauss

$$K(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}},\ u \in \mathbb{R}$$



Kernel de Epanechnikov

$$K(u) = \frac{3}{4} \cdot \left(1 - u^2\right),\ u \in [-1, 1]$$



Kernel uniforme

$$K(u) = \frac{1}{2},\ u \in [-1, 1]$$

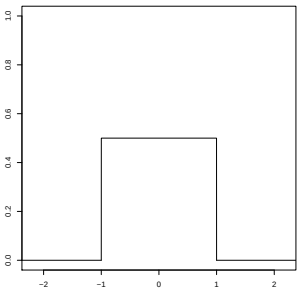
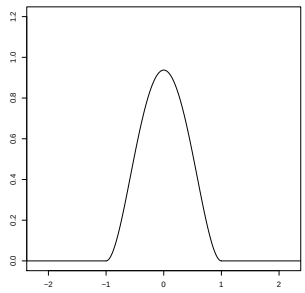


Figura 2.1: Kernels de Gauss, de Epanechnikov e uniforme.

Estes son os tres kernels da Figura 2.1 máis convencionais, aínda que particularmente o kernel uniforme non nos acaba de convencer pois ó ponderar por igual a todas as observacións do intervalo $(x_0 - h, x_0 + h)$ perde a esencia da ponderación como xa indiquei. Ademais pasar dun peso $\frac{1}{2}$ a 0 implicará unha perda de suavidade de $\hat{m}$. Finalmente, tamén convén mencionar o biweight e o triweight que presento na Figura 2.2:

Kernel biweight

$$K(u) = \frac{15}{16} \cdot \left(1 - u^2\right)^2,\ u \in [-1, 1]$$



Kernel triweight

$$K(u) = \frac{35}{32} \cdot \left(1 - u^2\right)^3,\ u \in [-1, 1]$$

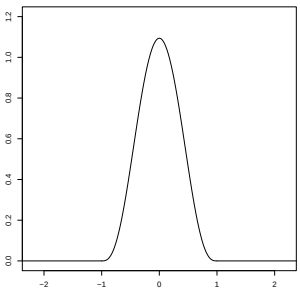


Figura 2.2: Kernels biweight e triweight.

O kernel de Gauss é o único, de entre os que expuxen, cuxo soporte é todo $\mathbb{R}$, co cal estritamente non nos estaremos restrinxindo ó intervalo $(x_0 - h, x_0 + h)$, senón que consideraremos todas as observacións. Porén, como fóra do intervalo $(x_0 - 3h, x_0 + 3h)$ este kernel asigna pesos menores á centésima, na práctica será como se os estivemos ignorando.

Non obstante, usar un kernel ou outro apenas supón cambios na estimación de m. Por iso, aínda que que as función kernel son importantes no marco da regresión lineal local, elixir unha ou outra non inflúe demasiado. Comprobaremos isto máis adiante.

Para entender mellor o comportamento da función K_h propoño fixarse na Figura 2.3, na que represento $K_h(x - x_0)$ para distintos valores de x_0 e h. Exemplifícoo tomando como kernel K o kernel de Gauss. En cor negra considero $x_0 = 0$ e $h = 1$, en cor vermella $x_0 = 1$ e $h = 0{,}5$ e en cor azul $x_0 = -1$ e $h = 0{,}25$.

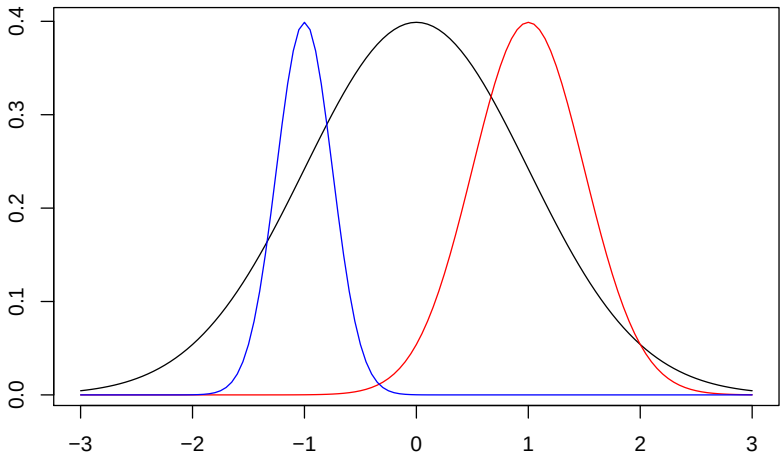


Figura 2.3: Función $K_h(x_i - x_0)$ para distintos valores de x_0 e h. En negro $x_0 = 0$ e $h = 1$, en vermello $x_0 = 1$ e $h = 0{,}5$ en azul $x_0 = -1$ e $h = 0{,}25$.

Pois ben, efectivamente, ó reducir o valor de h dende $h = 0{,}5$ na curva vermella a $h = 0{,}25$ na curva azul, o que estamos facendo é achatar o intervalo no que collemos os valores para o axuste lineal local; e ocorre o contrario cando pasamos da curva vermella con $h = 0{,}5$ á curva negra con $h = 1$. Por outra parte, o termo x_0 representa o centro de magnitude de datos e polo tanto o valor no que a ponderación é maior. A medida que nos afastamos deste valor de x_0 a ponderación vaise reducindo. Efectivamente, na curva vermella que considera $x_0 = 1$, o axuste priorizará as observacións proximas a 1, mentres que a curva azul aquelas preto de -1 e a negra as proximas a 0.

2.4. Estimador lineal local

Nesta sección amosarei a expresión do estimador da regresión lineal local.

Recordemos que tiñamos plantexado o seguinte problema de mínimos cadrados ponderados, no cal lle dabamos máis peso ós datos máis próximos:

$$\min_{a,b\in\mathbb{R}} \quad \sum_{i=1}^{n}\{Y_i - a - b\cdot x_i\}^2 \cdot K_h(x_i - x_0). \tag{2.11}$$

Co cal, unha vez estimados os valores que minimizan dita expresión, que denotaremos por $\hat{\beta}_0(x_0)$ e $\hat{\beta}_1(x_0)$, o valor do estimador lineal local para x_0 será:

$$\hat{m}(x_0) \;=\; \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0)\cdot x_0. \tag{2.12}$$

Como xa sinalei, aínda que $\hat{\beta}_0(x_0)$ non é interpretable, $\hat{\beta}_1(x_0)$ si o é, e correspóndese co crecemento/decrecemento da función de regresión. Así, se $\hat{\beta}_1(x_0)$ é positivo indicaranos que a variable resposta esta crecendo. No noso caso, o das eleccións de EEUU, significaría que a porcentaxe de votos estimados para Biden crece ó redor dese día. Analogamente, se $\hat{\beta}_1(x_0)$ toma un valor negativo pois significará que eses días a estimación da porcentaxe de votos sofre unha baixada.

Para resolver o problema de mínimos cadrados ponderados (2.11) é conveniente reescribilo en **notación matricial**, como unha forma cadrática que terá por matriz unha matriz diagonal, $W \in \mathbb{R}^{n\times n}$ que contén os pesos das observacións, estes son as ponderacións $K_h(x_i - x_0)$. Por outra banda, o erro e_i que tamén aparecía elevado ó cadrado en (2.11), reexprésase como un vector, $\boldsymbol{Y} - X\boldsymbol{\beta}$, que multiplica matricialmente á matriz W a ambos os lados e actuando trasposto cando o fai pola esquerda.

En notación matricial, o problema de minimización (2.11) quedaría como segue:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \; (\,\boldsymbol{Y} - X\boldsymbol{\beta})^T\, W\, (\,\boldsymbol{Y} - X\boldsymbol{\beta})\,, \tag{2.13}$$

onde:

$$\begin{gathered}\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n\times 2}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0(x_0) \\ \beta_1(x_0) \end{pmatrix} \in \mathbb{R}^2, \\ W = \operatorname{diag}\{K_h(x_i - x_0)\}_{i=1}^n = \begin{pmatrix} K_h(x_1 - x_0) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & K_h(x_n - x_0) \end{pmatrix} \in \mathbb{R}^{n\times n}.\end{gathered} \tag{2.14}$$

Como vemos, $\boldsymbol{Y}$ é o vector dos valores da variable resposta observados. A matriz X ten tantas filas como observacións e dúas columnas, a primeira está formada por 1s pois será multiplicada polo parámetro $\beta_0(x_0)$ que recorda ó intercepto da regresión lineal simple e a segunda columna contén os valores observados da variable explicativa, x_i.

Temos tamén o vector $\boldsymbol{\beta}$ que depende do punto x_0 no que esteamos interesados. É un vector de dúas entradas, que son precisamente as que queríamos estimar en (2.11) e polo tanto acudindo a notación matricial obteremos a estimación dos dous mediante un vector $\hat{\boldsymbol{\beta}}$.

Finalmente, a matriz cadrada diagonal W recolle os pesos que recibe cada observación. Esta matriz depende por un lado do punto x_0 e tamén depende do parámetro ventana h.

Pois ben, a función obxectivo que queremos minimizar é:

$$\begin{array}{rcl} \phi: \mathbb{R}^2 & \longrightarrow & \mathbb{R}, \\ \boldsymbol{\beta} & \rightsquigarrow & \phi(\boldsymbol{\beta}) := (\boldsymbol{Y} - X\boldsymbol{\beta})^T W (\boldsymbol{Y} - X\boldsymbol{\beta}). \end{array} \tag{2.15}$$

É importante recalcar que a pesar de que na función obxectivo ϕ interveñen matrices e vectores, non deixa de ser unha función real (precisamente responde á estrutura dunha unha forma cadrática). Os estimadores da regresión lineal local serán os que a minimicen.

Teorema 2.1. *(Expresión do estimador da regresión lineal local)*

Sexa un $x_0 \in \mathbb{R}$ un punto para o que queremos obter unha predición empregando regresión lineal local (que tamén pode ser algún dos puntos observados, $x_0 = x_i$) e sexa un h que determina a veciñanza que queremos empregar para obter dita predición.

Entón, se a matriz $X^T W X$ ten inversa, o estimador de mínimos cadrados ponderados de regresión lineal local ten por expresión:

$$\hat{\boldsymbol{\beta}} = \left(X^T W X\right)^{-1} X^T W \boldsymbol{Y},$$

que é o mínimo absoluto da función obxectivo (2.15) asociada ó problema de mínimos cadrados ponderados (2.11).

Observación 2.2. *No Teorema 2.4 analizaremos a existencia da inversa $(X^T W X)^{-1}$.*

Proba do Teorema. A demostración é elemental usando as propiedades usuais de derivación vectorial e matricial.

Primeiramente, para unha función real de variable vectorial, $\varphi : v \in \mathbb{R}^n \to \varphi(v) \in \mathbb{R}$, defínese o seu gradiente, $\nabla\varphi(v)$, ou derivada respecto a variable vectorial v como o vector formado polas derivadas usuais de ψ con respecto á cada unha das compoñentes de v:

$$\nabla\varphi(v) \;=\; \frac{\partial\varphi}{\partial v}(v) \;:=\; \left(\frac{\partial\varphi}{\partial v_1}(v), \cdots, \frac{\partial\varphi}{\partial v_j}, \cdots, \frac{\partial\varphi}{\partial v_n}(v)\right).$$

Por outro lado, para unha función vectorial de variable vectorial, $\psi : v \in \mathbb{R}^n \to (\psi_1(v), \ldots, \psi_n(v)) \in \mathbb{R}^m$, defínese a súa derivada con respecto a variable v como unha matriz, chamada xacobiana, $\mathcal{J}_\psi$, cuxas filas son os gradientes das compoñentes de ψ, funcións reais $\psi_i : v \in \mathbb{R}^n \to \psi_i(v) \in \mathbb{R}$. Así, a entrada (i, j) é a derivada de ψ_i con respecto a v_j:

$$\mathcal{J}_\psi(v) \;=\; \frac{\partial\psi}{\partial v}(v) \;:=\; \begin{pmatrix} \frac{\partial\psi_1}{\partial v_1}(v) & \cdots & \frac{\partial\psi_1}{\partial v_j}(v) & \cdots & \frac{\partial\psi_1}{\partial v_n}(v) \\ \vdots & \ddots & \vdots & & \vdots \\ \frac{\partial\psi_i}{\partial v_1}(v) & \cdots & \frac{\partial\psi_i}{\partial v_j}(v) & \cdots & \frac{\partial\psi_i}{\partial v_n}(v) \\ \vdots & & \vdots & \ddots & \vdots \\ \frac{\partial\psi_m}{\partial v_1}(v) & \cdots & \frac{\partial\psi_m}{\partial v_j}(v) & \cdots & \frac{\partial\psi_m}{\partial v_n}(v) \end{pmatrix}.$$

Proposición 2.3. *(Propiedades para a derivación vectorial e matricial)*

1. *Produto dunha matriz polo vector respecto ó que se deriva: Sexan unha matriz $A \in \mathbb{R}^{m\times n}$ e $\psi : \mathbb{R}^n \to \mathbb{R}^m$ unha función vectorial de variable vectorial dada por:*

   $$\psi(v) := Av,$$

   entón a súa derivada con respecto a v é a matriz A:

   $$\frac{\partial\psi}{\partial v}(v) \;=\; A.$$

2. *Forma cadrática: Se temos unha matriz simétrica $A \in \mathbb{R}^{n\times n}$ e unha función real de variable vectorial $\psi : \mathbb{R}^n \to \mathbb{R}$, definida como unha forma cadrática:*

   $$\psi(v) := v^T Av,$$

   entón a súa derivada con respecto a v ou gradiente ven dado por:

   $$\nabla\psi(v) \;=\; \frac{\partial\psi}{\partial v}(v) \;=\; 2v^T A.$$

3. *Regra da cadea: Se $\psi : \mathbb{R}^n \to \mathbb{R}$ é unha función real de variable vectorial que se pode expresar como a composición de dúas funcións, $f : \mathbb{R}^m \to \mathbb{R}$ e $g : \mathbb{R}^n \to \mathbb{R}^m$:*

   $$\psi(v) := (f \circ g)(v),$$

   entón a derivada de ψ con respecto a v ou gradiente pódese obter como segue:

   $$\nabla\psi(v) \;=\; \frac{\partial\psi}{\partial v}(v) \;=\; \frac{\partial f}{\partial v}(g(v))\,\frac{\partial g}{\partial v}(v).$$

A primeira propiedade dedúcese das propiedades usuais da derivada e pódese consultar na páxina 383 de Trench (2013), e a terceira na páxina 388 xunto ca súa proba. A segunda dedúcese das outras dúas.

Unha vez formalizada a derivación vectorial e matricial e dadas as propiedades de derivación, procedemos a derivar ϕ con respecto a $\boldsymbol{\beta}$ para obter os seus puntos críticos. Como ϕ é unha función real de variable vectorial, ó derivala obteremos o seu gradiente.

Podemos reescribir a nosa función obxectivo como a seguinte composición de funcións:

$$\phi(\boldsymbol{\beta}) \;=\; (f \circ g)(\boldsymbol{\beta}),$$

sendo $f : \mathbb{R}^n \to \mathbb{R}$ dada por: $f(\boldsymbol{v}) := \; \boldsymbol{v}^T W \boldsymbol{v}$ e $g : \mathbb{R}^2 \to \mathbb{R}^n$ dada por: $g(\boldsymbol{\beta}) := \; \boldsymbol{Y} - X\boldsymbol{\beta}$.

Así, pola terceira propiedade de derivación da Proposición 2.3, a da regra da cadea:

$$\nabla\phi(\boldsymbol{\beta}) \;=\; \frac{\partial \phi}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \;=\; \frac{\partial f}{\partial \boldsymbol{\beta}}\left(g(\boldsymbol{\beta})\right) \frac{\partial g}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}). \tag{2.16}$$

En primeiro lugar, como a función f é unha forma cadrática; en virtude da segunda propiedade de derivación dada na Proposición 2.3 para estas, teremos:

$$\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}) \;=\; 2\boldsymbol{v}^T W,$$

e como $g(\boldsymbol{\beta}) \;=\; \boldsymbol{Y} - X\boldsymbol{\beta}$, terase:

$$\frac{\partial f}{\partial \boldsymbol{\beta}}(g(\boldsymbol{\beta})) \;=\; 2(\,\boldsymbol{Y} - X\boldsymbol{\beta})^T W.$$

Por outro lado, como $g(\boldsymbol{\beta}) \;=\; \boldsymbol{Y} - X\boldsymbol{\beta}$, en virtude da primeira propiedade temos:

$$\frac{\partial g}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \;=\; -\,X.$$

Así, substituíndo en (2.16), obtemos a expresión do gradiente da función obxectivo tendo en conta que W é simétrica ($W = W^T$):

$$\nabla\phi(\boldsymbol{\beta}) \;=\; \frac{\partial \phi}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \;=\; -\,2(\,\boldsymbol{Y} - X\boldsymbol{\beta})^T W^T X \;=\; -\,2(\,\boldsymbol{Y} - X\boldsymbol{\beta})^T W X. \tag{2.17}$$

Así, igualando o gradiente a cero para obter os puntos críticos chegamos a:

$$\phi'\,(\boldsymbol{\beta}) \;=\; 0 \iff -\,2(\,\boldsymbol{Y} - X\boldsymbol{\beta})^T W X \;=\; 0 \iff \boldsymbol{Y}^T W X \;=\; (X\boldsymbol{\beta})^T W X \overset{\text{traspoñemos}}{\underset{\text{W simétrica}}{\iff}}$$

$$\iff X^T W\,\boldsymbol{Y} \;=\; X^T W X \boldsymbol{\beta}.$$

Polo tanto, como por hipótese $X^T W X$ é invertible, podemos despexar $\hat{\boldsymbol{\beta}}$ como:

$$\boxed{\hat{\boldsymbol{\beta}} \;=\; \left(X^T W X\right)^{-1} X^T W\,\boldsymbol{Y}.} \tag{2.18}$$

Tense así que $\hat{\boldsymbol{\beta}}$ é un punto crítico da función obxectivo. Agora comprobamos que é un mínimo acudindo á matriz hessiana, que a obtemos derivando o gradiente con respecto a $\boldsymbol{\beta}$.

Retomando a expresión do gradiente de ϕ que sinalei en (2.17), vemos que $\boldsymbol{\beta}$ só aparece multiplicando matricialmente a X, polo que para derivalo, basta con aplicar a primeira propiedade para a derivación exposta na Proposición 2.3, e así obtemos a matriz hessiana:

$$\mathcal{H}(\phi(\boldsymbol{\beta})) \;=\; \frac{\partial^2 \phi}{\partial \boldsymbol{\beta}^2}(\boldsymbol{\beta}) \;=\; \frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{\partial \phi}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta})\right) \;=\; 2\, X^T W X.$$

É unha matriz definida positiva, xa que para calquera vector $\boldsymbol{v} \in \mathbb{R}^2, \boldsymbol{v} \neq \{0\}$ se ten:

$$\boldsymbol{v}^T X^T W X \boldsymbol{v} \;>\; 0.$$

Vexamos que efectivamente é estritamente positivo (> 0) vendo que é non negativo (≥ 0) e non nulo ($\neq 0$). Por unha parte, agrupando os termos Xv quédanos unha forma cadrática sobre a matriz W, que como é semidefinida positiva (por ser diagonal con todos os elementos non negativos) tense que o produto matricial será maior ou igual ca cero:

$$\boldsymbol{v}^T X^T W X \boldsymbol{v} \;=\; (X\boldsymbol{v})^T W (X\boldsymbol{v}) \;\geq\; 0.$$

Por outro lado, notemos que non se pode dar a igualdade, pois nese $\boldsymbol{v}^T(X^T W X)\boldsymbol{v} = 0$, o cal contradiría a hipótese de que $X^T W X$ é invertible (que o estamos supoñendo por hipótese). Polo tanto, verifícase a desigualdade estrita en (2.4), co cal $X^T W X$ é definida positiva. Tense entón que a matriz hessiana $\mathcal{H}(\phi(\boldsymbol{\beta}))$ é definida positiva, e en consecuencia, o estimador $\hat{\boldsymbol{\beta}}$ obtido en (2.18) efectivamente é un mínimo.

Igual ca ocorría co estimador da regresión lineal simple, é un mínimo absoluto da función obxectivo ϕ, xa que é o único mínimo relativo e é fácil observar que a función obxectivo ϕ tende ó infinito cando tomamos valores de $\boldsymbol{\beta}$ que se escapan cara o $\pm\infty$ $\square$.

Neste teorema asumíase de xeito explícito a existencia da inversa de $X^T W X$. Non obstante, isto non se verifica automaticamente. No seguinte teorema, que é de elaboración propia, dou a condición baixo a cal para un punto x_0 queda garantida a existencia de dita inversa. Basicamente, esa inversa existirá se collemos un parámetro ventana h que englobe "suficientes observacións" como para que se poida facer regresión lineal local. Esta condición, que se formaliza no Teorema 2.4, esixe a existencia de polo menos dous datos na veciñanza $(x_0 - h, x_0 + h)$ e correspondentes a días diferentes para que $X^T W X$ sexa invertible empregando un kernel con soporte $[-1, 1]$. Para un kernel definido en $\mathbb{R}$, como o de Gauss, as condicións redúcense a unha como veremos nunha observación posterior.

Teorema 2.4. *(Sobre a existencia da inversa de X^TWX)*

Sexa un $x_0 \in \mathbb{R}$ un punto para o que queremos obter unha predición empregando regresión lineal local (que tamén pode ser algún dos puntos observados, $x_0 = x_i$) e sexa un $h > 0$ que determina a veciñanza que queremos empregar para obter dita predición.

Entón, se traballamos cun kernel cuxo soporte é $[-1, 1]$ (epanechnikov, uniforme, biweight, triweight,...), a inversa $(X^TWX)^{-1}$ que inclúe a expresión do estimador $\hat{\boldsymbol{\beta}}$ no Teorema 2.1 sobre a expresión do estimador lineal local existe se e só se:

$$\exists\, i, j \in \{1, ..., n\} \text{ con } i \neq j \quad \text{tales que:} \quad x_i \neq x_j \quad e \quad x_i, x_j \in (x_0 - h, x_0 + h).$$

Proba. Se facemos o produto matricial, resulta a matriz cadrada 2×2 seguinte:

$$X^TWX = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} K_h(x_1 - x_0) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & K_h(x_n - x_0) \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{2.19}$$

$$= \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} K_h(x_1 - x_0) & x_1 \cdot K_h(x_1 - x_0) \\ \vdots & \vdots \\ K_h(x_n - x_0) & x_n \cdot K_h(x_n - x_0) \end{pmatrix} \tag{2.20}$$

$$= \begin{pmatrix} \sum_{i=1}^n K_h(x_i - x_0) & \sum_{i=1}^n x_i \cdot K_h(x_i - x_0) \\ \sum_{i=1}^n x_i \cdot K_h(x_i - x_0) & \sum_{i=1}^n x_i^2 \cdot K_h(x_i - x_0) \end{pmatrix}. \tag{2.21}$$

Unha matriz cadrada é invertible se e só se o seu determinante é non nulo. Vexamos que isto ocorre baixo as condicións do enunciado. En vista do produto da matriz X^TWX que acabo de desenvolver, temos que o seu determinante é:

$$\left|X^TWX\right| = \left(\sum_{i=1}^n K_h(x_i - x_0)\right) \cdot \left(\sum_{i=1}^n x_i^2 \cdot K_h(x_i - x_0)\right) - \left(\sum_{i=1}^n x_i \cdot K_h(x_i - x_0)\right)^2.$$

Se reescribimos isto con sumatorios con índices distintos para posteriormente xuntalo todo no mesmo sumatorio dobre, temos:

$$\begin{aligned} \left|X^TWX\right| &= \sum_{i=1}^n K_h(x_i - x_0) \cdot \sum_{j=1}^n x_j^2 \cdot K_h(x_j - x_0) - \sum_{i=1}^n x_i \cdot K_h(x_i - x_0) \cdot \sum_{j=1}^n x_j \cdot K_h(x_j - x_0) \\ &= \sum_{i,j=1}^n x_j^2 \cdot K_h(x_i - x_0) \cdot K_h(x_j - x_0) - \sum_{i,j=1}^n x_i \cdot x_j \cdot K_h(x_i - x_0) \cdot K_h(x_j - x_0) \\ &= \sum_{i,j=1}^n \left(x_j^2 - x_i \cdot x_j\right) \cdot K_h(x_i - x_0) \cdot K_h(x_j - x_0). \end{aligned}$$

De onde, sacando factor común chegamos a:

$$\left|X^T W X\right| \;=\; \sum_{i,j=1}^{n} x_j \cdot (x_j - x_i) \cdot K_h(x_i - x_0) \cdot K_h(x_j - x_0).$$

Decatémosnos de que os termos correspondentes a eleccións de i e j iguais ($i = j$) resultarán ser nulos, pois nese caso $(x_j - x_i) = 0$. Polo tanto, podemos separar este dobre sumatorio en dous: un que considere os $i < j$ e outro os $i > j$:

$$\sum_{i,j=1,\ i<j}^{n} x_j{\cdot}(x_j{-}x_i){\cdot}K_h(x_i{-}x_0){\cdot}K_h(x_j{-}x_0) \;+\; \sum_{i,j=1,\ i>j}^{n} x_j{\cdot}(x_j{-}x_i){\cdot}K_h(x_i{-}x_0){\cdot}K_h(x_j{-}x_0).$$

Non obstante, renomeando os índices do segundo sumatorio chegamos a un sumatorio que se move nos mesmos índices e ca mesma condición que no primeiro ($i < j$):

$$\sum_{i,j=1,\ i<j}^{n} x_j{\cdot}(x_j{-}x_i){\cdot}K_h(x_i{-}x_0){\cdot}K_h(x_j{-}x_0) \;+\; \sum_{i,j=1,\ j>i}^{n} x_i{\cdot}\underbrace{(x_i - x_j)}_{-(x_j - x_i)}{\cdot}K_h(x_i{-}x_0){\cdot}K_h(x_j{-}x_0).$$

Polo tanto, podemos combinalo todo nun único sumatorio do seguinte xeito:

$$\sum_{i,j=1,\ i<j}^{n} (x_j - x_i) \cdot (x_j - x_i) \cdot K_h(x_i - x_0) \cdot K_h(x_j - x_0),$$

quedando entón o determinante da matriz $X^T W X$ como segue:

$$\left|X^T W X\right| \;=\; \sum_{i,j=1,\ i<j}^{n} (x_j - x_i)^2 \cdot K_h(x_i - x_0) \cdot K_h(x_j - x_0).$$

Efectivamente, como os sumandos son non negativos (pois temos un cadrado e o kernel, que é non negativo), teremos garantido que o determinante será tamén non negativo:

$$\left|X^T W X\right| \;\geq\; 0.$$

Por outra parte, ó ser todos os sumandos non negativos (≥ 0), para que o determinante sexa estritamente positivo abondará con que haxa un sumando non nulo. É dicir, o determinante será distinto de cero se e só se existen un par de índices cuxo sumando correspondente sexa estritamente positivo:

$$\left|X^T W X\right| \;\neq\; 0 \;\iff\; \exists\, i, j \in \{1, ..., n\} : (x_j{-}x_i)^2{\cdot}K_h(x_i{-}x_0){\cdot}K_h(x_j{-}x_0) > 0. \qquad (2.22)$$

Para que isto ocorra, os índices han de ter asociado un valor da variable explicativa distinto $x_i \neq x_j$ e ademais, a función K_h avaliada nas súas diferenzas respecto ó punto de interese x_0 ha de ser non nula. Como neste teorema estamos considerando os kernels que fóra de $(-1, 1)$ son nulos, esta segunda condición é equivalente a que os valores da variable

explicativa estean na veciñanza $(x_0 - h, x_0 + h)$. Polo tanto, baixo estas condicións o determinante será distinto de cero (concretamente, positivo), co cal a matriz será invertible. Pola contra, se non se cumprisen estas dúas condicións, todos os sumandos serían nulos e polo tanto o determinante tamén (e non sería invertible). É dicir, tense a equivalencia:

$$\begin{array}{c} X^TWX \text{é invertible} \\ \left(\left|X^TWX\right| \neq 0\right) \end{array} \iff \exists\, i,j \in \{1,...,n\} \text{ verificando } \begin{cases} x_i \neq x_j, \\ x_i, x_j \in (x_0 - h, x_0 + h). \end{cases}$$

Como queríamos probar, estas condicións equivalen á existencia da inversa $\left(X^TWX\right)^{-1}$.

□

Entón, este resultado interprétase do xeito seguinte, X^TWX será invertible se e só se arredor do punto x_0 no que queremos obter a predición hai polo menos dúas observacións i, j con valores distintos da variable explicativa $(x_i \neq x_j)$ que están a unha distancia de x_0 menor ca h. Notemos que neste teorema empreguei un x_0 xenérico, pero podería ser un valor observado da variable resposta, e dicir, podería ser $x_0 = x_i$ para certo $i \in \{1, ..., n\}$.

Con respecto á aplicar este teorema ó noso caso de datos reais, o das eleccións estadounidenses de 2020, a primeira condición está clara, precisamos dúas enquisas que se correspondan a días diferentes $(x_i \neq x_j)$. A segunda condición tradúcese en que esas enquisas estean a unha distancia menor a h días do día x_0 no que queremos facer a predición.

Este teorema trata o caso de kernels que fóra do intervalo $(-1, 1)$ se anulan, como son o de epanechnikov, o uniforme o biweight ou o triweight entre outros. Se empregamos un kernel que tome valores non nulos en todo $\mathbb{R}$ (como o de Gauss) basta con que haxa polo menos dúas observacións con valores distintos na variable explicativa para a existencia teórica da inversa, como explico na seguinte observación, tamén de elaboración propia.

Observación 2.5. *(Sobre a existencia da inversa $(X^TWX)^{-1}$ co kernel de Gauss)*

Se empregamos un kernel non nulo en todo $\mathbb{R}$, como o de Gauss, as funcións K_h serán non nulas sempre, polo que para o cumprimento da condición (2.22) bastará a existencia de dúas observacións con valores distintos na súa variable explicativa $(x_i \neq x_j)$, conseguindo así que o determinante sexa non nulo, o que equivale á existencia da inversa $(X^TWX)^{-1}$.

Non obstante e aínda tendo máis observacións, se estas se atopan moi afastadas e empregamos un h pequeno, a calidade do axuste é malo. Por iso, non está de máis que se cumpra a condición do Teorema 2.4 tamén para o kernel de Gauss.

Damos por concluído o relativo á expresión do estimador $\hat{\boldsymbol{\beta}}$ e procedemos entón a calcular os axustes, $\hat{Y}_i = m(x_i)$ e a predición nun punto, $\hat{Y}_0 = \hat{m}(x_0)$.

2.5. Predición e axustes

Recuperando a notación inicial, o $\hat{\boldsymbol{\beta}}$ obtido no Teorema 2.1 ten por coeficientes os $\hat{\beta}_0(x_0)$ e $\hat{\beta}_1(x_0)$ da expresión (2.12) que resolvían o problema de minimización (2.10):

$$\hat{\boldsymbol{\beta}} \;=\; \begin{pmatrix} \hat{\beta}_0(x_0) \\ \hat{\beta}_1(x_0) \end{pmatrix}.$$

Deste xeito, se reescribimos a expresión (2.12) en notación matricial temos:

$$\hat{m}(x_0) \;=\; \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0)\cdot x_0 \;=\; \begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0(x_0) \\ \hat{\beta}_1(x_0) \end{pmatrix} \;=\; \begin{pmatrix} 1 & x_0 \end{pmatrix} \hat{\boldsymbol{\beta}}.$$

Substituíndo o estimador $\hat{\boldsymbol{\beta}}$ obtido en (2.18), obtemos a **predición para $\boldsymbol{x_0}$** como:

$$\hat{Y}_0 \;=\; \hat{m}(x_0) \;=\; \begin{pmatrix} 1 & x_0 \end{pmatrix} \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}. \tag{2.23}$$

Para maior comodidade, denotaremos por X_0 o vector fila:

$$X_0 \;:=\; \begin{pmatrix} 1 & x_0 \end{pmatrix}, \tag{2.24}$$

e deste xeito podemos escribir a predición $\hat{Y}_0$ como segue:

$$\boxed{\hat{Y}_0 \;=\; \hat{m}(x_0) \;=\; X_0 \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}.} \tag{2.25}$$

Esta expresión depende de h e da función kernel que van incorporados na matriz diagonal W como vimos en (2.14). No fondo, a predición non é máis ca un promedio dos valores observados Y_i que premia a aqueles cuxa variable explicativa x_i está máis preto de x_0.

Entón, a partir da expresión (2.23) obtemos o axuste para a observación i-ésima:

$$\hat{Y}_i \;=\; \hat{m}(x_i) \;=\; X_i \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}, \qquad i \in \{1, \dots, n\}. \tag{2.26}$$

Nótese a diferenza entre x_0 e $X_0 = \begin{pmatrix} 1 & x_0 \end{pmatrix}$ e entre x_i e $X_i = \begin{pmatrix} 1 & x_i \end{pmatrix}$. Mentres os primeiros denotan un valor (escalar) da variable explicativa, os segundos denotan un vector fila que ten por entradas un 1 e o valor da variable explicativa.

Recordando a definición de X dada en (2.14), vemos que X_i é precisamente a fila i da matrix X, polo que podemos reescribir $X_i \;=\; e_i^T X$, quedando o axuste i-ésimo como:

$$\hat{Y}_i \;=\; \hat{m}(x_i) \;=\; e_i^T X \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}, \qquad i \in \{1, \dots, n\}. \tag{2.27}$$

Esta expresión non se pode xeneralizar para un vector de axustes pois a matriz W, definida en (2.14), depende do punto x_0 no que se quere facer a predición (neste caso $x_0 = x_i$).

2.6. Aspectos computacionais

En todo proceso computacional aínda que son importantes a precisión e na medida do posible a simplicidade, é fundamental prestar especial atención á eficacia. Movido pola intención de alixeirar os procesos (cálculo do parámetro ventana óptimo, obtención de predicións,...) xorde a reformulación que explicarei nesta sección, que **reduce os tempos de execución**. Posteriormente empregarei as expresións explícitas que obteremos para relacionar a regresión lineal simple ca regresión lineal simple.

Chegaremos a expresións para $\hat{\boldsymbol{\beta}}$ e $\hat{m}(x_0)$ que suporán unha redución de tempo significativa. Para desenvolver esta sección partín da idea que se propón na páxina 95 de Fan e Gibels (1996) para este mesmo fin. Non obstante, aínda que a idea de base é a mesma, para o desenvolvemento desta sección non seguín exactamente a mesma formulación.

Recordemos que a expresión de $\hat{\boldsymbol{\beta}}$ que deducimos en (2.18) viña dada por:

$$\hat{\boldsymbol{\beta}} \;=\; \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}.$$

Pois ben, reescribiremos $\hat{\boldsymbol{\beta}}$ como o produto matricial da inversa dunha matriz simétrica $S_n \in \mathbb{R}^{2\times 2}$ por un vector $\boldsymbol{T}_n \in \mathbb{R}^2$:

$$\hat{\boldsymbol{\beta}} \;=\; S_n^{-1}\, \boldsymbol{T}_n. \tag{2.28}$$

De xeito evidente, S_n e $\boldsymbol{T}_n$ correspóndense con cada unha das seguintes expresións:

$$S_n \;=\; X^T W X, \tag{2.29}$$

$$\boldsymbol{T}_n \;=\; X^T W\, \boldsymbol{Y}. \tag{2.30}$$

Vexamos como podemos reescribir ambas matrices. A matriz $S_n = X^T W X$, xa foi obtida na demostración do Teorema 2.4 nos pasos (2.19), (2.20) e (2.21) como:

$$S_n \;=\; X^T W X \;=\; \begin{pmatrix} \sum_{i=1}^n K_h(x_i - x_0) & \sum_{i=1}^n x_i \cdot K_h(x_i - x_0) \\ \sum_{i=1}^n x_i \cdot K_h(x_i - x_0) & \sum_{i=1}^n x_i^2 \cdot K_h(x_i - x_0) \end{pmatrix}.$$

Como vemos é unha matriz simétrica que escribiremos como:

$$S_n \;=\; \begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2} \end{pmatrix},$$

e cuxas compoñentes se poden obter de modo xenérico como:

$$S_{n,j} = \sum_{i=1}^{n} x_i^j \cdot K_h(x_i - x_0), \qquad j \in \{0, 1, 2\}. \tag{2.31}$$

Para unha matriz 2×2 simétrica, $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$, a inversa é fácil de obter:

$$A^{-1} \;=\; \frac{1}{a_{11} \cdot a_{22} - a_{12}^2} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{12} & a_{11} \end{pmatrix}.$$

Entón, como S_n está nesas condicións (matriz 2×2 simétrica), a súa inversa é:

$$S_n^{-1} \;=\; \frac{1}{S_{n,0} \cdot S_{n,2} - S_{n,1}^2} \begin{pmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0} \end{pmatrix},$$

onde o denominador é o determinante de S_n.

Por outra parte, reescribimos tamén a expresión que indiquei en (2.30) para $\boldsymbol{T}_n$:

$$\begin{aligned} \boldsymbol{T}_n \;&=\; X^T W \boldsymbol{Y} \;=\; \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} K_h(x_1 - x_0) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & K_h(x_n - x_0) \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\ &=\; \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} Y_1 \cdot K_h(x_1 - x_0) \\ \vdots \\ Y_n \cdot K_h(x_n - x_0) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \cdot K_h(x_i - x_0) \\ \sum_{i=1}^n x_i \cdot Y_i \cdot K_h(x_i - x_0) \end{pmatrix}. \end{aligned}$$

Co cal, $\boldsymbol{T}_n$ é un vector de $\mathbb{R}^2$ que escribiremos como:

$$\boldsymbol{T}_n \;=\; \begin{pmatrix} T_{n,0} \\ T_{n,1} \end{pmatrix},$$

cuxas compoñentes teñen como expresión xenérica:

$$T_{n,j} \;=\; \sum_{i=1}^{n} x_i^j \cdot Y_i \cdot K_h(x_i - x_0), \qquad j \in \{0, 1\}. \tag{2.32}$$

Por conseguinte, se retomamos a expresión de $\hat{\boldsymbol{\beta}}$ dada en (2.28) e temos en conta o visto ata agora, podemos reescribir este estimador do seguinte xeito:

$$\hat{\boldsymbol{\beta}} \;=\; S_n^{-1} \boldsymbol{T}_n \;=\; \frac{1}{S_{n,0} \cdot S_{n,2} - S_{n,1}^2} \begin{pmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0} \end{pmatrix} \begin{pmatrix} T_{n,0} \\ T_{n,1} \end{pmatrix},$$

que, unha vez efectuamos o produto matricial, queda expresado como segue:

$$\hat{\boldsymbol{\beta}} \;=\; \frac{1}{S_{n,0} \cdot S_{n,2} - S_{n,1}^2} \begin{pmatrix} S_{n,2} \cdot T_{n,0} - S_{n,1} \cdot T_{n,1} \\ S_{n,0} \cdot T_{n,1} - S_{n,1} \cdot T_{n,0} \end{pmatrix}. \tag{2.33}$$

Así mesmo, conseguimos expresar a predición para x_0 en función das cinco compoñentes da matriz e vector e do valor da variable explicativa:

$$\hat{Y}_0 \;=\; \hat{m}(x_0) \;=\; \begin{pmatrix}1 & x_0\end{pmatrix}\hat{\boldsymbol{\beta}} \;=\; \frac{1}{S_{n,0}\cdot S_{n,2} - S_{n,1}^2}\cdot \begin{pmatrix}1 & x_0\end{pmatrix}\begin{pmatrix}S_{n,2}\cdot T_{n,0} - S_{n,1}\cdot T_{n,1}\\ S_{n,0}\cdot T_{n,1} - S_{n,1}\cdot T_{n,0}\end{pmatrix}.$$

Facendo o produto matricial chégase á seguinte expresión para o cálculo da predición para x_0 que só depende das cinco compoñentes da matriz e do vector:

$$\hat{Y}_0 \;=\; \frac{S_{n,2}\cdot T_{n,0} - S_{n,1}\cdot T_{n,1} \;+\; x_0\cdot(S_{n,0}\cdot T_{n,1} - S_{n,1}\cdot T_{n,0})}{S_{n,0}\cdot S_{n,2} - S_{n,1}^2}, \tag{2.34}$$

onde as compoñentes de S_n e de $\boldsymbol{T}_n$ teñen por expresións xenéricas as que sinalamos en (2.31) e en (2.32).

Deste xeito, poderemos obter a ventana óptima h e calcular a predición dunha forma máis eficiente e rápida. O inconveniente desta reescritura é que non se ve con tanta claridade o que hai detrás dos pasos que se van efectuando, pero pola súa contra obtemos tempos de execución moito máis competitivos como veremos no último capítulo.

Cabe preguntase que relación garda a expresión do estimador de mínimos cadrados para a regresión lineal local, (2.18), cos da regresión lineal simple. Pois ben, están intimamente relacionados xa que os estimadores da regresión lineal local veñen sendo os estimadores da regresión lineal simple cando a matriz de pesos é a identidade. Isto débese a que con $W = I$, estanse a ter en conta todas as observacións e todas co mesmo peso, co cal estaríamos collendo unha veciñanza que contén a todas os valores da variable explicativa e asignándolle a todas as observacións o peso unidade. A comprobación é inmediata empregando esta última notación de S_n e $\boldsymbol{T}_n$ como mostro na seguinte observación, na que partindo dun parámetro ventana h que abarque todas as observacións e un kernel uniforme vemos que efectivamente estamos facendo regresión lineal simple:

Observación 2.6. *(Regresión lineal simple) Tomando un h suficientemente grande, que abarque todas as observacións e un kernel que asigne peso 1 a todas elas, estamos baixo regresión lineal simple, é dicir a expresión dada para $\hat{\boldsymbol{\beta}}$ usando regresión lineal local recollida no Teorema 2.1, coincide ca dada para regresión lineal simple dada no Teorema 1.1.*

Proba. Efectivamente, cun h que abarque todas as observacións e cun kernel que lles asigne a todas peso 1, teremos que $K_h(x_i - x_0) = 1$ para todo $i \in \{1, ..., n\}$, polo tanto as compoñentes de S_n e T_n dadas en (2.31) e (2.32) quedarían expresadas do seguinte xeito:

$$S_{n,j} = \sum_{i=1}^{n} x_i^j, \qquad T_{n,j} = \sum_{i=1}^{n} x_i^j \cdot Y_i.$$

Polo tanto temos:

$$S_{n,0} = n, \qquad S_{n,1} = \sum_{i=1}^{n} x_i, \qquad S_{n,2} = \sum_{i=1}^{n} x_i^2, \qquad T_{n,0} = \sum_{i=1}^{n} Y_i, \qquad T_{n,1} = \sum_{i=1}^{n} x_i \cdot Y_i.$$

Logo, retomando a expresión para $\hat{\boldsymbol{\beta}}$ dada en (2.33) que empregaba S_n e $\boldsymbol{T}_n$, e substituíndo as expresións de $S_{n,0}$,$S_{n,1}$,$S_{n,2}$,$T_{n,0}$ e $T_{n,1}$ que acabo de presentar, chégase a:

$$\hat{\boldsymbol{\beta}} = \frac{1}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \left(\sum_{i=1}^{n} x_i^2\right) \cdot \left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} x_i \cdot Y_i\right) \\ n \cdot \left(\sum_{i=1}^{n} x_i \cdot Y_i\right) - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} Y_i\right) \end{pmatrix}. \tag{2.35}$$

Así, obtemos o estimador $\hat{\beta}_0$ como:

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) \cdot \left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} x_i \cdot Y_i\right)}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Dividindo o numerador e o denominador entre o cadrado do número de datos, obtemos:

$$\hat{\beta}_0 = \frac{\frac{\left(\sum_{i=1}^{n} x_i^2\right) \cdot \left(\sum_{i=1}^{n} Y_i\right)}{n^2} - \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} x_i \cdot Y_i\right)}{n^2}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}.$$

E nesta expresión hai termos que se corresponden cas medias mostrais da variable explicativa, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e da variable resposta, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Tendo isto en conta, reescribimos a expresión anterior como segue:

$$\hat{\beta}_0 = \frac{\frac{\sum_{i=1}^{n} x_i^2}{n} \cdot \overline{Y} - \overline{x} \cdot \frac{\sum_{i=1}^{n} x_i \cdot Y_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2}.$$

Restamos e sumamos o termo $\overline{x}^2 \cdot \overline{Y}$ para simplificar:

$$\hat{\beta}_0 = \frac{\frac{\sum_{i=1}^{n} x_i^2}{n} \cdot \overline{Y} - \overline{x}^2 \cdot \overline{Y} + \overline{x}^2 \cdot \overline{Y} - \overline{x} \cdot \frac{\sum_{i=1}^{n} x_i \cdot Y_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2} = \overline{Y} + \frac{\overline{x}^2 \cdot \overline{Y} - \overline{x} \cdot \frac{\sum_{i=1}^{n} x_i \cdot Y_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2}$$

$$= \overline{Y} - \frac{\overline{x} \cdot \frac{\sum_{i=1}^{n} x_i \cdot Y_i}{n} - \overline{x}^2 \cdot \overline{Y}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2}.$$

Agora sacando factor común $\overline{x}$ do numerador do cociente obtemos:

$$\hat{\beta}_0 \;=\; \overline{Y} - \frac{\frac{\sum_{i=1}^n x_i \cdot Y_i}{n} - \overline{x} \cdot \overline{Y}}{\frac{\sum_{i=1}^n x_i^2}{n} - \overline{x}^2} \cdot \overline{x}.$$

Que, en efecto, coincide ca expresión dada para o estimador lineal simple no Teorema 1.1, xa que o numerador correspóndese ca covarianza entre x e Y: $S_{xy}^2 = \frac{\sum_{i=1}^n x_i \cdot Y_i}{n} - \overline{x} \cdot \overline{Y}$; e o denominador ca varianza de x: $S_x^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \overline{x}^2$ como vimos en (1.6) e (1.7) respectivamente.

Polo tanto, efectivamente chegamos á mesma expresión que tíñamos no Teorema 1.1:

$$\hat{\beta}_0 \;=\; \overline{Y} - \frac{S_{xY}}{S_x^2} \cdot \overline{x}.$$

Analogamente, comprobamos que o estimador $\hat{\beta}_1$ coincide co dado no Teorema 1.1. Recuperando a segunda compoñente en (2.35) que se corresponde co estimador de $\hat{\beta}_1$:

$$\hat{\beta}_1 \;=\; \frac{n \cdot (\sum_{i=1}^n x_i \cdot Y_i) - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n Y_i)}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}.$$

Dividimos novamente no numerador e no denominador entre n^2 e volvemos a expresalo en función das medias mostrais:

$$\hat{\beta}_1 \;=\; \frac{\frac{\sum_{i=1}^n x_i \cdot Y_i}{n} - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n^2}}{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} \;=\; \frac{\frac{\sum_{i=1}^n x_i \cdot Y_i}{n} - \overline{x} \cdot \overline{Y}}{\frac{\sum_{i=1}^n x_i^2}{n} - \overline{x}^2}.$$

Igual ca vimos para $\hat{\beta}_0$, o numerador $\hat{\beta}_1$ correspóndese ca covarianza entre x e Y e o numerador ca varianza de x, polo tanto, efectivamente nos queda a mesma expresión que obtíñamos para a regresión lineal simple no Teorema 1.1:

$$\hat{\beta}_1 \;=\; \frac{S_{xY}}{S_x^2}.$$

□

A consecuencia inmediata e previsible disto é que a que comentaba antes: a regresión lineal local con matriz de pesos $W = I$ é precisamente a regresión lineal simple. Que a matriz de pesos sexa a identidade reflexa que para todas as observacións, a avaliación do kernel é a unidade: $K_h(x_i - x_0) = 1$. Así, estamos considerando que todas as observacións da explicativa, x_i están no intervalo $x_i \in (x_0 - h, x_0 + h)$ e ademais que o kernel que empregamos considera a todas as observacións co mesmo peso, 1; o cal é xustamente regresión lineal simple, ca que se perde toda a esencia e todo o atractivo da regresión lineal local que premiaba aqueles puntos máis próximos na medida da súa proximidade.

Capítulo 3

Elección do parámetro ventana

Neste capítulo abordarei o importante papel do parámetro ventana no modelo lineal local, que determina a amplitude dos intervalos nos que se fai a regresión lineal local controlando cantas observacións interveñen no promedio local. Este parámetro é o que se emprega no problema de minimización (2.10), ou dado matricialmente (2.13) mediante a matriz diagonal W que incorpora o h na diagonal ponderando as observacións.

En primeiro lugar, introducirei os conceptos de nesgo e varianza na Sección 3.1, que formalizaremos na Sección 3.2, onde tamén veremos súas expresións exactas, que non seremos capaces de aplicar para seleccionar o h a partir da mostra posto que dependerán de cantidades poboacionais descoñecidas. Isto levaranos a traballar con aproximacións asintóticas que permiten unha análise teórica máis sinxela. Ó mesmo tempo tratarei un concepto de gran importancia a nivel estadístico, como é o erro cadrático medio, dando ademais a expresión do parámetro ventana que o minimiza. Finalmente, presentarei unha técnica moi socorrida para o cálculo de estimacións de parámetros, que é o método de validación cruzada, que a partir dunha mostra escolle como ventana óptima aquela que mellor axusta os datos da mostra asumindo que cada dato non intervén na súa propia predición.

3.1. Nesgo e Varianza

Nun modelo non paramétrico é moi importante ter en conta como de preto han de estar as observacións para consideralas. Isto débese a que, como amosarei ó largo desta sección e da seguinte, reducir moito o intervalo que estamos a considerar conleva unha redución do nesgo, pero a costa dun aumento da variabilidade da estimación, mentres que se o ampliamos conseguimos reducir a varianza, pero teremos que asumir un nesgo maior. Veremos como a elección do parámetro ventana resulta decisiva á hora de obter un equilibrio en-

tre o nesgo e a varianza. Nesta sección, ilustrarei estas realidades facendo uso do conxunto de datos `economics` da libraría `ggplot2` (Wickham, 2016), que xa empreguei na Sección 1.3.

En primeiro lugar, o nesgo dun estimador é unha medida da diferenza do valor esperado do estimador con respecto ó valor real do parámetro que se está a estimar, mentres que a varianza cuantifica a dispersión que posúe o estimador con respecto ó valor esperado. Non obstante, definireinos rigorosamente na seguinte sección.

Na procura de resultados innesgados (aqueles cuxo valor esperado coincide co que se pretende estimar), un pode facilmente intuír que, reducindo o valor do parámetro ventana alcanzamos o noso obxectivo. E así é, pero tamén debemos ter en conta é que estaremos asumindo unha variabilidade maior. Ademais, deste xeito quédanos un modelo moi sensible a calquera ruído local nos datos, xa que ó reducir o intervalo no que consideramos as observacións para facer os nosos promedios, estamos usando moitas menos observacións e polo tanto se algunha fose non representativa ou contase con moito ruído estaríamos modelando tamén ese ruído, cando a nosa pretensión é simplemente modelar os valores esperados. Esta situación recibe o nome de sobreaxuste e ten como consecuencia modelos con moita variabilidade e moi sensibles ás pequenas flutuacións no conxunto de datos.

Ilustro este feito na Figura 3.1, na cal escollín unha ventana moi pequena para explicar o desemprego medio en función do mes, concretamente unha ventana de dous meses.

Figura 3.1: Axuste lineal local con unha ventana de 2 meses explicando `uempmed` (eixo y) en función de `data` (eixo x) do conxunto de datos `economics`.

Como vemos na gráfica da Figura 3.1, a curva de axuste oscila moito e iso débese a que o axuste é demasiado "local" xa que estamos restrinxindo máis do debido o intervalo de datos que consideramos. Así, e tal como veremos nos cálculos teóricos posteriores, se ben é certo que conseguimos unha mellora en termos de nesgo, isto é a costa de pagar unha variabilidade moi elevada como se pode intuír ó observar a figura.

Outro posible enfoque consiste en aumentar o valor do parámetro ventana, ensanchando así o intervalo no que consideramos as observacións pois deste xeito conseguimos un axuste que conta cunha variabilidade moito menor. E a consecuencia é previsible, estaremos afrontando un nesgo maior xa que, ó ser o valor de h elevado, o modelo estará considerando en cada punto moitas máis observacións, incluso aquelas máis afastadas que non teñen que ver co modelo no punto, desembocando así nun desaxuste. Deste xeito prodúcese un efecto de "sobresuavizado" sobre a curva de regresión uniformizando máis o modelo e levándonos a un modelo máis simple e incapaz de captar regularidades importantes.

Na Figura 3.2 mostro o que obtemos con un h elevado (concretamente dun ano e medio):

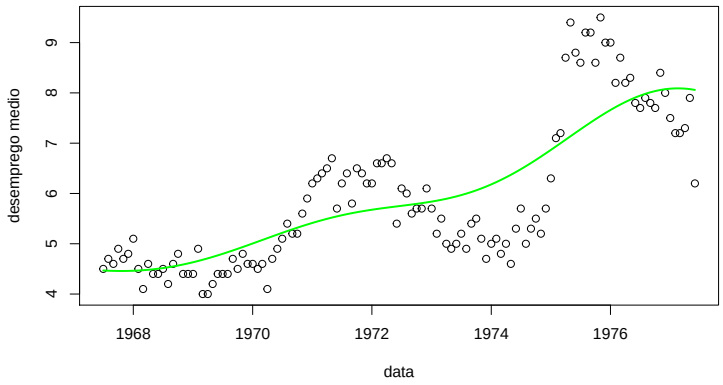


Figura 3.2: Axuste lineal local con unha ventana de 1 ano e medio explicando `uempmed` (eixo y) en función de `data` (eixo x).

Apreciamos que as observacións afastadas conseguen facer certo efecto panca provocando un suavizado excesivo da curva de regresión, e deste xeito menor varianza. Porén, podemos intuír que a curva está claramente sufrindo en termos de nesgo.

Como vemos, atopámosnos cun dilema entre minimizar o nesgo e minimizar a varianza. Ante este panorama no que non é posible reducir ambos simultaneamente preséntasenos unha solución moi atractiva a este conflito, a minimización do erro cadrático medio que ten en conta tanto o nesgo como a varianza, e permite en consecuencia reducir o erro xeral levándonos así a un modelo que consigue un equilibrio entre ambos.

É importante sinalar que o visto ata agora no relativo ó nesgo e á varianza nas sucesivas gráficas simplemente nos permite intuír a existencia destes conceptos, pero realmente só se ven cambiando de mostra e considerando axustes distintos para analizar o seu comportamento. Iso será o que fagamos na primeira parte da seguinte sección recorrendo a un exemplo de simulación. Aínda así, pareceume interesante introducir estes conceptos con un exemplo real, antes de facelo con un simulado como veremos a continuación.

3.2. Erro Cadrático Medio

Como adiantei na sección anterior, a idea que presento nesta sección para obter o parámetro h é moi fácil de comprender, propoño tomar como h óptimo aquel que minimice o erro cadrático medio do estimador da función de regresión, $\hat{m}(x_0)$.

Primeiramente introducirei un exemplo de simulación que afondará nos conceptos de nesgo e a varianza, que posteriormente definirei con rigor. A continuación, faremos un análise exhaustivo dos mesmos, para os que darei tanto expresións teóricas como tamén aproximacións asintóticas. Chegados a ese punto, tomarei como h óptimo aquel que alcance o menor valor para o erro cadrático medio, que denotarei por $\hat{h}_{ECM}$.

Todos os exemplos que irei presentando nesta sección serán relativos a unha función de regresión que tomará unha forma por todos coñecida, a función seno. Concretamente tomarei como función de regresión a estimar a función $m(x) = \text{sen}(2\pi x)$ no intervalo $[0, 1]$. Isto axudará ó lector a visualizar mellor o nesgo e a varianza. En particular, comprobaremos que ó facer regresión lineal local realmente non estamos estimando a función de regresión m, senón que estamos estimando un promedio local dos valores de dita función de regresión nos valores observados da variable explicativa. Neste feito interveñen o nesgo e a varianza e será ilustrado con claridade nas vindeiras páxinas.

Colleremos n puntos equiespaciados do intervalo $[0, 1]$ e xeraremos as observacións cun erro de media nula e certa varianza. É dicir, xeraremos $Y_i = \text{sen}(2\pi x_i) + \varepsilon_i$, onde os erros ε_i son independentes e normais con media $\mathbb{E}[\varepsilon_i] = 0$ e varianza $\text{Var}(\varepsilon_i) = \sigma^2$, $\mathcal{N}(\mu = 0, \sigma^2)$.

Concretamente, tomamos 50 puntos equiespaciados do intervalo $[0, 1]$ que darán valor á variable explicativa, $x_i = i/49,\ i \in \{0, ..., 49\}$ e deixaremos fixos eses valores durante todo o exemplo. A continuación, xero os valores da variable resposta tal como expliquei antes, sumándolles un erro de media cero e varianza $\sigma^2 = 0{,}1$ ca axuda de `rnorm` , que é un comando de R que devolve erros dunha distribución normal con certa media e certa varianza.

Pois ben, interesámosnos por saber a que se aproximan os axustes lineais locais. Para isto xeraremos $M = 1000$ conxuntos de observacións da variable resposta que denotaremos por $Y^{(1)}, ..., Y^{(j)}, ..., Y^{(M)}$, sumándolle a cada un deles un erro de media 0 e varianza 0,1 de modo totalmente independente, de xeito que para cada $j \in \{1, ..., 1000\}$, $Y^{(j)}$ é un conxunto de 50 observacións da variable resposta Y (para cada un dos 50 puntos equidistantes da variable x), $Y^{(j)} = \left\{Y_1^{(j)}, ..., Y_{50}^{(j)}\right\}$, xerados tal e como especifiquei anteriormente.

Repetindo ese proceso 1000 veces o que pretendo é para cada un deses conxuntos de valores da variable resposta, $Y^{(1)}, ..., Y^{(j)}, ..., Y^{(1000)}$, axustar o modelo lineal local para $h = 0{,}25$, obtendo 1000 curvas de axuste diferentes. Non obstante, só represento as vintecinco primeiras en cor cian na Figura 3.3 para facernos unha idea.

A continuación fago un promedio desas 1000 curvas para aproximar $\mathbb{E}\left[\hat{m}(x)\right]$, e finalmente como estamos traballando sobre un modelo coñecido, podemos calcular o promedio lineal local da propia función de regresión, $m(x) = \text{sen}(2\pi x)$ (insisto en que esta función en xeral é descoñecida, pero neste caso estamos probando sobre o caso coñecido do seno). Entón fago o promedio lineal local tal e como se fai para os Y_i pero agora para os auténticos valores da función de regresión sobre os puntos x_i (estes son $m(x_i) = \text{sen}(2\pi x_i)$) e represéntoa en cor vermello na Figura 3.3. A curva que obtemos é efectivamente á que se aproximan as curvas de axuste, representadas en cian na mesma figura. En conclusión, en regresión lineal local non estamos aproximando a propia función m, senón o seu promedio avaliado sobre os valores mostrais da variable explicativa, os x_i.

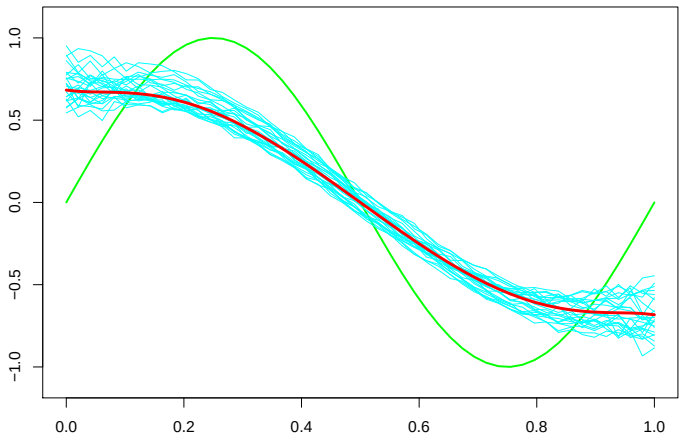


Figura 3.3: Axuste lineal local para 25 cálculos de valores de Y_i a partir de $\text{sen}(2\pi x_i)$ cun erro normal $\mathcal{N}(\mu = 0, \sigma^2 = 0{,}1)$ (en cian) e o promedio local los valores reais de $m(x_i)$ (en vermello) empregando unha ventana $h = 0{,}25$.

Efectivamente, como apreciamos na Figura 3.3, non se está a aproximar a función de regresión (representada en verde), senón que se está a aproximar outra curva que é precisamente a curva de promedios dos valores da función de regresión nos puntos da mostra (que é a curva que debuxo en vermello). É dicir, a regresión lineal local aproxima un promedio local dos $m(x_i) = \text{sen}(2\pi x_i)$.

Ademais, tamén chama a atención algo que xa se adiantara cando formalizabamos a regresión lineal local e é o feito de que ó estimador cústalle captar os máximos e os mínimos, xa que tal e como se pode apreciar na Figura 3.3, non chega ben ós picos e vales nos máximos e mínimos, algo que se agudiza conforme aumentamos o valor do parámetro ventana.

Por outro lado, nesta mesma Figura 3.3 podemos apreciar un efecto fronteira, pois como vemos nos puntos proximos ó 0 e ó 1, as curvas de estimación dispérsanse máis, contrastando ca maior concentración que atopamos nos puntos centrais.

Na Figura 3.3 estase a empregar un parámetro ventana $h = 0{,}25$ e vemos que a regresión lineal local ten dificultades para achegarse ó valor real da función de regresión. Isto débese a un alto nesgo, que xa viamos na sección anterior e que definirei posteriormente de forma rigorosa. Para paliar este alto valor do nesgo, podemos reducir o valor do parámetro, por exemplo ata un $h = 0{,}08$, e repetir o proceso de xerar as 1000 curvas para ver como se comportan agora o nesgo e a varianza e obtemos a Figura 3.4:

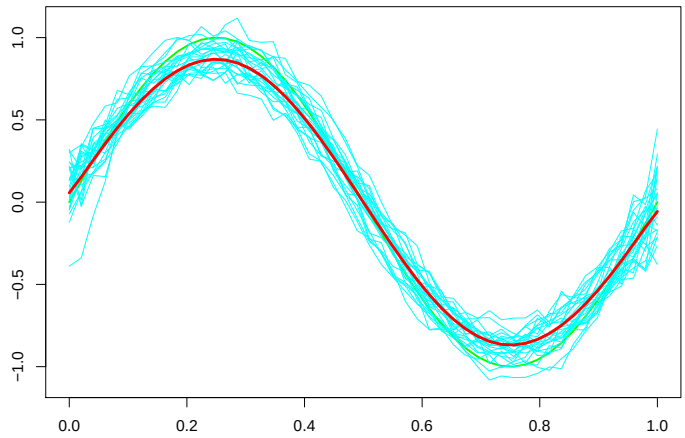


Figura 3.4: Axuste lineal local para 25 cálculos de valores de Y_i a partir de $\text{sen}(2\pi x_i)$ cun erro normal $\mathcal{N}(\mu = 0, \sigma^2 = 0{,}1)$ (en cian) e o promedio local los valores reais de $m(x_i)$ (en vermello) empregando unha ventana $h = 0{,}08$.

Como se pode ver, reducir o parámetro ventana h consegue, como pretendíamos, unha redución importante do nesgo (xa que as curvas de axuste oscilan máis cerca da función do seno), mentres que aumentando o h incorremos nun incremento do nesgo, que se agudiza nos puntos extremos (máximos e mínimos) nos que o axuste ten dificultades. Non obstante, reducir a ventana tamén ten inconvenientes, pois conleva un aumento da variabilidade que se percibe no engrosamento da nube de curvas en cian, dun xeito bastante moderado pero si se produce este engrosamento. É dicir, reducir o h aumenta a variabilidade das estimacións.

É aquí onde aparece o dilema de escoller un parámetro ventana ou outro, pois escoller un parámetro ventana baixo reducirá o nesgo, pero aumentará a varianza. Analogamente, considerar un h elevado reducirá a varianza, pero provocará un incremento no nesgo. Deberemos entón escoller entón un h que consiga un equilibrio entre o nesgo e a varianza.

Por outro lado, as dificultades para axustar os máximos e os mínimos que se poden apreciar nas Figuras 3.3 e 3.4 débense a que son puntos onde a curvatura toma valores moi elevados, e polo tanto a segunda derivada alcanza valores máis extremos. Así, tal e como expliquei na Sección 2.1, a aproximación lineal que se considera no modelo lineal local implica un erro en termos da derivada segunda, polo que captará ben a monotonía pero non a curvatura. De feito, nótese que no noso caso concreto estamos intentando axustar localmente a función seno (que non é para nada lineal) empregando un polinomio de Taylor de grado 1, e por iso se ten ese alto valor do nesgo nos puntos extremos.

Á vista da Figura 3.4, volvo a facer fincapé no feito de que non estamos estimando a función m, senón un promedio da mesma na veciñanza, ponderando as observacións segundo a súa proximidade de acordo co parámetro ventana escollido. Nesta figura aínda que é certo que a curva á que se aproximan (a de promedios dos puntos $m(x_i)$) está máis cerca da verdadeira curva do seno, segue sen coincidir co valor da verdadeira función de regresión. O feito de que estea máis preto da curva real do seno débese exclusivamente a que o nesgo é menor, precisamente por estar tomando un h menor.

Como viñamos dicindo, debemos buscar un equilibrio entre o nesgo e a varianza, para conseguir un axuste que se aproxime á función de regresión, pero ó mesmo tempo, que non conte con moita variabilidade. Para iso, na seguinte sección recorreremos ó erro cadrático medio que é unha medida conxunta do nesgo e da varianza, e poderemos tomar como h óptimo aquel que o minimiza.

Como acabamos de ver ca axuda das sucesivas figuras, o nesgo e a varianza condicionan a calidade do axuste lineal local, polo que convén prestar especial atención a ambos. A continuación farei unha exposición das definicións e tamén das expresións teóricas destes conceptos, que non poderemos empregar na práctica por depender de cantidades poboacionais descoñecidas e tampouco nos ofrecerán unha interpretación clara. Isto levaranos a considerar expresións asintóticas, que si serán interpretables e servirannos ademais para obter o h que minimiza o erro cadrático medio asintótico e que denotaremos por h_{ECM}.

Definición 3.1. O **erro cadrático medio**, ECM (MSE segundo as súas siglas en inglés: *Mean Squares Error*) defínese como a esperanza do cadrado do erro. Así, o erro cadrático medio dun estimador $\hat{\theta}$ dun parámetro poboacional descoñecido θ queda definido por:

$$\text{ECM}\left(\hat{\theta}\right) \;:=\; \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right], \tag{3.1}$$

onde o operador $\mathbb{E}[\cdot]$ denota a esperanza matemática, que ven sendo o valor esperado dunha variable aleatoria. Recordemos que a esperanza dunha variable aleatoria continua Z con función de densidade $f_Z(z)$ ven dada por:

$$\mathbb{E}[Z] \;:=\; \int z\; f_Z(z)\; dz.$$

Non obstante, nos faremos uso dunha reescritura e propiedades que nos facilitarán as contas para o cálculo do erro cadrático medio do noso estimador $\hat{m}(x_0)$.

Primeiramente, podemos reescribir a expresión do erro cadrático medio dada en (3.1), empregando os conceptos de nesgo e varianza que estiven ilustrando (e que definirei con rigor a continuación):

$$\text{ECM}\left(\hat{\theta}\right) \;=\; \left(\text{Nesgo}\left(\hat{\theta}\right)\right)^2 \;+\; \text{Var}\left(\hat{\theta}\right). \tag{3.2}$$

A comprobación da veracidade desta reescritura é sinxela e pódese ver na Subsección 7.5.3 de Peña (2005). Formalicemos entón os conceptos de nesgo e varianza.

Definición 3.2. O **nesgo** dun estimador $\hat{\theta}$ dun parámetro θ é a diferenza entre a súa esperanza matemática e o valor do parámetro que se está a estimar:

$$\text{Nesgo}\left(\hat{\theta}\right) \;:=\; \mathbb{E}\left[\hat{\theta}\right]-\theta.$$

Definición 3.3. A **varianza** dun estimador $\hat{\theta}$ é unha medida da dispersión que este posúe con respecto ó seu valor esperado. Formalmente, calcúlase como a esperanza do cadrado da diferenza do estimador con respecto ó seu valor esperado:

$$\text{Var}\left(\hat{\theta}\right) \;:=\; \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left(\hat{\theta}\right)\right)^2\right].$$

Aínda que $\hat{m}(x_0)$ é un estimador escalar, como na expresión que obtivemos en (2.23) interveñen vectores e matrices debemos formalizar a esperanza e a varianza de cantidades vectoriais, así como algunhas propiedades dos mesmos que serán útiles para obter o nesgo e varianza de $\hat{m}(x_0)$, que é o que perseguimos.

Consideramos agora un vector aleatorio $\boldsymbol{Z} = (Z_1, ..., Z_m)^T$. A esperanza deste vector aleatorio será un vector $\mathbb{E}[\boldsymbol{Z}]$ cuxas entradas son as esperanzas de cada unha das súas compoñentes, mentres que para medir a dispersión xa non teremos unha varianza escalar, senón unha matriz cadrada $\mathrm{Cov}(\boldsymbol{Z}) \in \mathbb{R}^{m\times m}$ que recibe o nome de matriz de varianzas-covarianzas e que terá na diagonal cada unha das varianzas das compoñentes do vector aleatorio, mentres que o resto son as covarianzas entre cada par de compoñentes.

$$\mathbb{E}[\boldsymbol{Z}] = \begin{pmatrix} \mathbb{E}[Z_1] \\ \vdots \\ \mathbb{E}[Z_m] \end{pmatrix}, \qquad \mathrm{Cov}(\boldsymbol{Z}) = \begin{pmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_m) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Var}(Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(Z_m, Z_1) & \mathrm{Cov}(Z_m, Z_2) & \cdots & \mathrm{Var}(Z_m) \end{pmatrix},$$

onde $\mathrm{Cov}(Z_i, Z_j)$ denota a covarianza entre os estimadores escalares $\hat{Z}_i$ e $\hat{Z}_j$, que ven sendo a variación conxunta dos dous con respecto ás súas medias:

$$\mathrm{Cov}(Z_i, Z_j) \;=\; \mathbb{E}\left[(Z_i - \mathbb{E}[Z_i])\,(Z_j - \mathbb{E}[Z_j])\right].$$

A continuación procedo a presentar as fórmulas exactas para o nesgo e para a varianza do estimador da función de regresión $\hat{m}(x_0)$. Posteriormente teremos que recorrer a aproximacións asintóticas, xa que as expresións analíticas exactas involucran cantidades descoñecidas que dependen de h dun xeito complicado de analizar.

Recordemos a expresión do estimador $\hat{m}(x_0)$ á que chegaramos en (2.23)

$$\hat{m}(x_0) = X_0(X^T W X)^{-1} X^T W\, \boldsymbol{Y}, \tag{3.3}$$

onde X_0 denotaba a matriz fila de dúas compoñentes $X_0 = \begin{pmatrix} 1 & x_0 \end{pmatrix}$.

Para poder empregar esta expresión no cálculo do nesgo e da varianza de $\hat{m}(x_0)$ necesitaremos aplicar as propiedades que presento na seguinte proposición.

Proposición 3.4. *(Propiedades da esperanza e varianza) Sexan $\boldsymbol{Z}$ un vector aleatorio de dimensión m e unha matriz $\mathcal{A} \in \mathbb{R}^{p\times m}$ entón:*

1. *Para a esperanza:* $\mathbb{E}[A\boldsymbol{Z}] \;=\; A\,\mathbb{E}[\boldsymbol{Z}]$,
2. *Para a matriz de varianzas-covarianzas:* $\mathrm{Cov}(A\boldsymbol{Z}) \;=\; A\,\mathrm{Cov}(\boldsymbol{Z})A^T$.

Estas propiedades pódense consultar na páxina 238 de Peña (2005). A súa demostración é consecuencia directa das propiedades da esperanza e varianza univariante.

Recalco que aínda que $\hat{m}(x_0)$ é un estimador escalar, para deducir o seu nesgo e varianza empregando a expresión (3.3) deberemos traballar co vector de esperanzas e matriz de varianzas-covarianzas de magnitudes vectoriais; en concreto cas do vector de valores observados da variable resposta $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$.

Empezamos calculando o nesgo do estimador $\hat{m}(x_0)$, para o que debemos calcular antes a súa esperanza, que é fácil de obter tendo en conta que estamos baixo deseño fixo (isto é, os x_i veñen dados de antemán), polo que podemos aplicar a Proposición 3.4 chegando a:

$$\mathbb{E}\,[\hat{m}(x_0)] \;=\; X_0(X^TWX)^{-1}X^TW\;\mathbb{E}\,[\,\boldsymbol{Y}\,] \;=\; X_0(X^TWX)^{-1}X^TW\boldsymbol{m}, \tag{3.4}$$

sendo $\boldsymbol{m}$ o vector dos valores reais da función de regresión nos puntos observados da variable explicativa, $\boldsymbol{m} = (m(x_1), ..., m(x_n))^T$, e polo tanto unha cantidade descoñecida. Vemos na expresión (3.4) que $\mathbb{E}\,[\hat{m}(x_0)]$ é efectivamente o promedio local que define o estimador pero aplicado a $\boldsymbol{m}$ en vez de a $\boldsymbol{Y}$, confirmando o que viñamos intuíndo nas Figuras 3.3 e 3.4.

Denotando agora por $\boldsymbol{e}$ ó vector de residuos nos valores observados, $\boldsymbol{e} = (e_1, ..., e_n)^T$ (de novo, unha cantidade descoñecida), temos a relación:

$$\boldsymbol{m} \;=\; X\boldsymbol{\beta} + \boldsymbol{e},$$

e polo tanto, podemos reexpresar a esperanza de $\hat{m}(x_0)$ que obtivemos en (3.4) como:

$$\mathbb{E}\,[\hat{m}(x_0)] \;=\; X_0(X^TWX)^{-1}X^TW\,(X\boldsymbol{\beta} + \boldsymbol{e}) \;=\; X_0\boldsymbol{\beta} + X_0(X^TWX)^{-1}X^TW\boldsymbol{e}. \tag{3.5}$$

Finalmente, como para o punto x_0 no que queremos obter a predición se ten que:

$$m(x_0) \;=\; X_0\boldsymbol{\beta} \;+\; e(x_0),$$

sendo $e(x_0)$ o erro de predición cometido no punto x_0. Despexando $X_0\boldsymbol{\beta}$ nesta expresión chegamos a $X_0\boldsymbol{\beta} \;=\; m(x_0) - e(x_0)$, co cal a esperanza de $\hat{m}(x_0)$ dada en (3.5) queda como:

$$\mathbb{E}\,[\hat{m}(x_0)] = \; m(x_0) - e(x_0) + X_0(X^TWX)^{-1}X^TW\boldsymbol{e}.$$

Co cal, vemos que é un estimador nesgado cuxo nesgo é:

$$\text{Nesgo}(\hat{m}(x_0)) = \; \mathbb{E}\,[\hat{m}(x_0)] - m(x_0) \;=\; X_0(X^TWX)^{-1}X^TW\boldsymbol{e} - e(x_0). \tag{3.6}$$

Como vemos, ambos os dous sumandos son escalares (aínda que o primeiro involucre produtos matriciais). Ademais é importante notar tamén a diferenza entre $\boldsymbol{e}$ e $e(x_0)$, mentres o primeiro é un vector que recolle os erros nos puntos da mostra x_i, o segundo é un escalar que se corresponde co erro no punto x_0 no que queremos facer a predición.

Para calcular a varianza do estimador $\hat{m}(x_0)$ traballaremos de novo con expresións vectoriais e matriciais. Tendo en conta a expresión de $\hat{m}(x_0)$ que recordei en (3.3), podemos escribir a súa varianza como a matriz de covarianzas da expresión $X_0(X^TWX)^{-1}X^TW\,\boldsymbol{Y}$ para así poder aplicar a segunda propiedade da Proposición 3.4 que nos leva a:

$$\begin{aligned}\operatorname{Var}(\hat{m}(x_0)) &= \operatorname{Cov}\left(X_0(X^TWX)^{-1}X^TW\,\boldsymbol{Y}\right)\\ &= X_0(X^TWX)^{-1}X^TW\ \operatorname{Cov}(\boldsymbol{Y})\,W^TX\left((X^TWX)^T\right)^{-1}X_0^T.\end{aligned}$$

Recordando que W é unha matriz diagonal, temos por unha parte $W^T = W$ e por outra que $\left((X^TWX)^T\right)^{-1} = (X^TWX)^{-1}$, quedando a varianza expresada como segue:

$$\operatorname{Var}(\hat{m}(x_0)) = X_0(X^TWX)^{-1}X^TW\ \operatorname{Cov}(\boldsymbol{Y})\,WX(X^TWX)^{-1}X_0^T.$$

Como estamos supoñendo que as observacións se toman de xeito independente, non estarán relacionadas unhas cas outras, polo que a matriz de varianzas-covarianzas de $\boldsymbol{Y}$ será diagonal, é dicir, $\operatorname{Cov}(\boldsymbol{Y}) = \operatorname{diag}\left\{\sigma^2(x_i)\right\}_{i=1}^n$.

Polo tanto, denotando por $\Sigma = W\ \operatorname{Cov}(\boldsymbol{Y})\ W = \operatorname{diag}\left\{\sigma^2(x_i)\cdot K_h(x_i-x_0)^2\right\}_{i=1}^n$, obtemos a varianza de $\hat{m}(x_0)$ como:

$$\operatorname{Var}(\hat{m}(x_0)) = X_0(X^TWX)^{-1}X^T\Sigma X(X^TWX)^{-1}X_0^T. \tag{3.7}$$

Unha vez obtidos o nesgo en (3.6) e a varianza en (3.7), e tendo en conta (3.2), podemos expresar o erro cadrático medio para $\hat{m}(x_0)$ como a súa suma:

$$\operatorname{ECM}(\hat{m}(x_0)) = X_0(X^TWX)^{-1}X^TW\boldsymbol{e} - e(x_0) + X_0(X^TWX)^{-1}X^T\Sigma X(X^TWX)^{-1}X_0^T.$$

Pero estas expresións dependen de cantidades descoñecidas como o vector $\boldsymbol{e}$ que recolle os erros exactos nos puntos da mostra, o escalar $e(x_0)$ que recolle o erro no punto x_0 ou a matriz diagonal Σ que depende da varianza nos puntos da mostra, $\sigma^2(x_0)$. Isto impídenos interpretar con facilidade estas expresións, polo que nos vemos na obriga de recorrer a aproximacións asintóticas do nesgo e da varianza de cara a poder empregar o erro cadrático medio para obter un h óptimo que o minimice e analizar a influencia de h sobre estas medidas estadísticas.

A continuación mostro un resultado que se extrae do libro de Fan e Gijbels (1996), no que se proporcionan aproximacións asintóticas do nesgo e varianza de $\hat{m}(x_0)$, ás que se chega na páxina 66 deste libro. De acordo ca notación deste resultado, denotaremos por $\sigma^2(x_0)$ á varianza condicional de Y dado x_0 e por $f(\cdot)$ á densidade de datos en dito punto. Este resultado tense baixo deseño aleatorio (isto é, os x_i non veñen fixados, senón que son aleatorios).

Proposición 3.5. *Nun modelo de regresión lineal local con m suave no que $f(\cdot)$ e $\sigma^2(\cdot)$ son continuas nunha veciñanza de x_0 e $f(x_0) > 0$ se se verifica que $h \to 0$ e $n \cdot h \to \infty$ teremos as seguintes expresións para o nesgo e varianza asintóticos:*

$$\begin{aligned} \text{Nesgo}\left(\hat{m}(x_0)\right) &= \frac{1}{2}\left(\int t^2\ K(t)\ dt\right) \cdot m''(x_0) \cdot h^2 \ +\ o_P(h^2), \\ \text{Var}\left(\hat{m}(x_0)\right) &= \left(\int K^2(t)\ dt\right) \cdot \frac{\sigma^2(x_0)}{f(x_0) \cdot n \cdot h} \ +\ o_P\left(\frac{1}{n \cdot h}\right), \end{aligned}$$

onde supoñemos que ambas integrais son finitas e onde $o_P(\cdot)$ denota unha variable aleatoria tal que $o_P\left(h^2\right) = h^2 \cdot\ o_P(1)$ e $o_P\left(\frac{1}{n \cdot h}\right) = \frac{1}{n \cdot h} \cdot o_P(1)$, veficándose $o_P(1) \xrightarrow{P} 0$.

Á vista destas aproximacións asintóticas, confirmamos o que viñamos dicindo, reducir h reduce o nesgo, pero aumenta a varianza e viceversa. Ademais, notemos que o nesgo tamén depende do valor de $m''(x_0)$, que será mínimo nos puntos de máxima curvatura. Finalmente, se nunha rexión hai poucos datos ($f(x_0)$ pequeno) entón a varianza asintótica do esimador incrementarase. Veremos despois con máis detalle o efecto destes termos sobre o h que minimiza o erro cadrático medio.

Como o erro cadrático medio é a suma do nesgo ó cadrado e a varianza tal como sinalamos en (3.2), en virtude da Proposición 3.5, temos a seguinte aproximación asintótica:

$$\text{ECM}\left(\hat{m}_0(x_0)\right) \ \simeq\ \frac{1}{4}\left(\int t^2\ K(t)\ dt\right)^2 \cdot (m''(x_0))^2 \cdot h^4 + \left(\int K^2(t)\ dt\right) \cdot \frac{\sigma^2(x_0)}{f(x_0) \cdot n \cdot h}\ . \quad (3.8)$$

Propoñémosnos achar o parámetro ventana h que minimiza o erro cadrático medio aproveitando esta aproximación asintótica que acabamos de obter en (3.8). En primeiro lugar derivamos e igualamos a cero para obter o mínimo:

$$4 \cdot \left\{\frac{1}{4}\left(\int t^2\ K(t)\ dt\right)^2 \cdot (m''(x_0))^2\right\} \cdot h^3 - \left\{\left(\int K^2(t)\ dt\right) \cdot \frac{\sigma^2(x_0)}{f(x_0) \cdot n}\right\} \cdot \frac{1}{h^2} \ =\ 0 \iff$$

$$\iff \quad h^5 \ =\ \frac{\left(\int K^2(t)\ dt\right) \cdot \frac{\sigma^2(x_0)}{f(x_0) \cdot n}}{\left(\int t^2\ K(t)\ dt\right)^2 \cdot (m''(x_0))^2}\ .$$

Polo tanto, a expresión da **ventana que minimiza o erro cadrático medio** é:

$$h_{ECM} \ =\ \sqrt[5]{\frac{\left(\int K^2(t)\ dt\right) \cdot \sigma^2(x_0)}{\left(\int t^2\ K(t)\ dt\right)^2 \cdot (m''(x_0))^2 \cdot f(x_0) \cdot n}}\ . \quad (3.9)$$

Para comprobar que efectivamente é un mínimo recorremos á segunda derivada.

A segunda derivada do erro cadrático medio é:

$$3 \cdot \left\{ \left(\int t^2 \, K(t) \, dt \right)^2 \cdot (m''(x_0))^2 \right\} \cdot h^2 \; + \; \left\{ \left(\int K^2(t) \, dt \right) \cdot \frac{\sigma^2(x_0)}{f(x_0) \cdot n} \right\} \cdot \frac{1}{h^3} > 0.$$

Efectivamente, todos os termos son non negativos xa que hai termos ó cadrado, integrais de termos positivos e a función f que ó ser unha función de densidade é non negativa. Polo tanto, a segunda derivada é non negativa. Pero ademais é estritamente positiva xa que no segundo sumando temos $\sigma^2(x_0) > 0$ e unha integral do cadrado da función kernel, que non é identicamente nula, co cal, tomará un valor estritamente positivo tamén. Polo tanto, o estimador (3.9) efectivamente é o que minimiza o erro cadrático medio.

Notemos que a expresión que dá o h que minimiza o erro cadrático medio asintótico, (3.9), depende de cantidades descoñecidas como son a varianza, a función de densidade e a segunda derivada da función de regresión. Isto fará que na práctica non o poidamos empregar para seleccionar h a partir dos datos. Non obstante, a expresión é de gran interese para ver a influencia que estes tres factores sobre o h_{ECM} óptimo que obteñamos:

Empecemos vendo a **influencia da varianza**. Decatémosnos de que canto maior sexa a varianza $\sigma^2(x_0)$, maior será tamén a ventana h_{ECM} que teremos que escoller para minimizar o erro cadrático medio. Isto era esperable pois se temos uns datos con moita varianza, deberemos acudir a un modelo máis suavizado, é dicir un no que abarquemos unha veciñanza maior compensando así a maior variabilidade mostral á que nos enfrontamos.

En segundo lugar, estudemos a **influencia da función de densidade** no punto, $f(x_0)$. Un valor pequeno de $f(x_0)$ significará que hai poucos datos nesa veciñanza (de feito o número esperado de datos é $n \cdot h \cdot f(x_0)$), o cal deriva nun aumento da variabilidade, xa que estaremos axustando usando poucos datos mostrais. Para compensar isto, haberá que ensanchar a veciñanza, englobando máis datos, é dicir aumentar h. Por esa razón $f(x_0)$ vai no denominador na expresión (3.9), pois así queda expresada esa dependencia inversa.

Finalmente, é especialmente interesante analizar o **efecto de** $m''(x_0)$ na expresión (3.9). Recordemos que como xa expliquei, a función ten dificultades para alcanzar os máximos e mínimos, especialmente a medida que h aumenta, pois imos englobando máis observacións, agudizando así o efecto panca que os puntos próximos teñen sobre o axuste. Isto refléxase nesta expresión do h_{ECM} pois en puntos nos que a curvatura é elevada, é dicir, o valor de $m''(x_0)$ é grande, o h_{ECM} que minimiza o erro cadrático medio será menor, restrinxíndonos así a unha veciñanza máis pequena, na cal ese punto extremo se captará mellor.

Por outro lado, a expresión (3.9) tamén nos permite comprobar que **a elección do kernel apenas supón cambios** na estimación da función de regresión. Para velo, agrupamos os termos que dependen da función kernel baixo unha constante C:

$$C \;=\; \sqrt[5]{\frac{\int K^2(t)\ dt}{\left(\int t^2\ K(t)\ dt\right)^2}}\ ,$$

quedando entón a expresión de h_{ECM} que presentabamos en (3.9) como segue:

$$h_{ECM} \;=\; C \cdot \sqrt[5]{\frac{\sigma^2(x_0)}{(m''(x_0))^2 \cdot f(x_0) \cdot n}}\ . \tag{3.10}$$

Esa constante C inflúe directamente no h_{ECM}, aínda que só implica un cambio de escala pois na estimación de $\hat{m}(x_0)$ apenas se nota a diferenza. Hai táboas que nos dan os valores desta constante C como a que se mostra na páxina 67 de Fan e Gibels (1996).

Como acabo de indicar, esta influencia de C non produce un gran cambio na estimación de $\hat{m}(x_0)$, pois se collemos esta expresión de h_{ECM} dada en (3.10) e a substituímos na expresión do erro cadrático medio que vimos en (3.8) despois de simplificar obtemos a seguinte expresión asintótica para o erro cadrático medio asociado a $h = h_{ECM}$:

$$\mathrm{ECM}(\hat{m}(x_0)) \;\simeq\; \frac{5}{4} \cdot \sqrt[5]{\left(\int K^2(t)\ dt\right)^4 \cdot \left(\int t^2 \cdot K(t)\ dt\right)^2} \cdot \sqrt[5]{\frac{\sigma^8(x_0) \cdot (m''(x_0))^2}{f^4(x_0) \cdot n^4}}, \tag{3.11}$$

onde, como vemos a efectos de erro cadrático medio, a función kernel só intervén a través da constante $\sqrt[5]{\left(\int K^2(t)\ dt\right)^4 \cdot \left(\int t^2 \cdot K(t)\ dt\right)^2}$. Calculei o valor de dita constante para os cinco kernels que se presentaron na Sección 2.3 e recólloos na Táboa 3.1 redondeados a tres decimais.

Táboa 3.1: Valores da constante $\sqrt[5]{\left(\int K^2(t)\ dt\right)^4 \cdot \left(\int t^2 \cdot K(t)\ dt\right)^2}$ para varios kernels.

Gauss	Epanechnikov	Uniforme	Biweight	Triweight
0.363	0.349	0.370	0.351	0.353

Como podemos apreciar na Táboa 3.1, a constante $\sqrt[5]{\left(\int K^2(t)\ dt\right)^4 \cdot \left(\int t^2 \cdot K(t)\ dt\right)^2}$ apenas varía dun kernel a outro, polo tanto a diferenza de escoller un ou outro kernel será minúscula en termos de erro cadrático medio. En definitiva, elixir un ou outro kernel apenas provocará diferenzas na estimación de $\hat{m}(x_0)$.

3.3. Validación Cruzada

Na práctica, o parámetro ventana que minimiza o erro cadrático medio h_{ECM} non se pode calcular, xa que depende de cantidades descoñecidas. Nesta sección propoño unha alternativa que si se pode usar na práctica: a validación cruzada, que se pode consultar na Subsección 4.10.2 de Fan e Gibels (1996). Tamén hai outros métodos para seleccionar o parámetro ventana (pero que non se trataran neste traballo) tales como a minimización do erro cadrático medio integral, o método "plug-in", a validación cruzada nesgada ou a validación cruzada suavizada que se poden consultar no Capítulo 3 de Wand e Jones (1995).

Propoñemos para obter o parámetro ventana h óptimo recorrer ó método de validación cruzada, que é unha técnica moi empregada en estadística que ten como idea fundamental ver cal é o h que consegue adaptarse mellor ás observacións que xa temos.

Baséase en partir dun conxunto hs de parámetros ventana e ver cal ten asociado o menor erro de validación cruzada, que medirá a discrepancia dos valores da resposta con respecto ós axustes dados por un modelo no que cada observación non intervén no seu propio axuste.

Formalmente, para cada parámetro do conxunto de parámetros ventana, $h \in hs$, calculamos a predición para cada dato $i \in \{1, \dots, n\}$, sen telo en conta. É dicir, tendo en conta soamente os valores da explicativa $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ cos seus correspondentes valores na variable resposta, $Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n$ obteremos a predición para x_i que denotaremos por $Y^h_{i(i)}$, que é, como veño comentando, a predición para o dato i-ésimo sen telo en conta e para ese valor de h.

Así, definiremos o erro de validación cruzada asociado á ventana h, $CV(h)$ (segundo as súas siglas do termo en inglés, *Cross Validation*) como o promedio dos erros de predición:

$$CV(h) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y^h_{i(i)}\right)^2. \tag{3.12}$$

Deste modo, partindo dun conxunto de posibles parámetros ventana hs consideramos o problema de minimización asociado a este erro de validación cruzada:

$$\min_{h\in hs}\; CV(h) \;=\; \min_{h\in hs}\; \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y^h_{i(i)}\right)^2.$$

Co cal, o parámetro ventana óptimo segundo o método de validación cruzada, que denotaremos por $\hat{h}_{CV}$, é aquel que minimiza este erro:

$$h_{CV} \;=\; \arg\min_{h\in hs} CV(h).$$

Vexamos pois como dar con este h óptimo.

Recordemos a expresión obtida en (2.26) para o axuste i-ésimo:

$$\hat{Y}_i \;=\; X_i \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}, \qquad\qquad i \in \{1, \dots, n\},$$

sendo $X_i = (\;1 \quad x_i\;)$.

Observemos entón que o axuste para o dato i-ésimo sen telo en conta é:

$$\hat{Y}_{i(i)} \;=\; X_i \left(X^T \tilde{W}_{(i)} X\right)^{-1} X^T \tilde{W}_{(i)}\, \boldsymbol{Y}, \qquad\qquad i \in \{1, \dots, n\},$$

onde, para cada $i \in \{1, \dots, n\}$, $\tilde{W}_{(i)}$ é a matriz diagonal $W \in \mathbb{R}^{n \times n}$ definida no capítulo anterior pero co i-ésimo valor da diagonal nulo, impedindo así que interveña o dato i-ésimo.

Pero razoemos con rigor que escollendo esas matrices $\tilde{W}_{(i)}$ efectivamente conseguimos anular o efecto de cada dato i-ésimo:

Por unha parte, como a matriz W é diagonal, a matriz $\tilde{W}_{(i)}$ terá a fila i e a columna i nulas. Na matriz X^T, a observación i-ésima correspóndese ca columna i:

$$X^T \;=\; \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \\ 1 & \cdots & x_i & \cdots & x_n \end{pmatrix}.$$

Como ó facer o produto matricial $X^T \tilde{W}_{(i)}$, os elementos da columna i de X^T se están multiplicando polos elementos da fila i de $\tilde{W}_{(i)}$, que é nula. Polo tanto, a observación i na práctica non está participando nese produto matricial $X^T \tilde{W}_{(i)}$. Ademais, como a columna i de $\tilde{W}_{(i)}$ é nula, a columna i do produto $X^T \tilde{W}_{(i)}$ será tamén nula.

Por outro lado, na matriz X, a observación i-ésima correspóndese ca fila i:

$$X \;=\; \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Entón, no produto matricial $X^T \tilde{W}_{(i)} X$, os elementos da fila i de X estanse multiplicando polos elementos da columna i de $X^T \tilde{W}_{(i)}$; que como acabo de indicar, é nula, logo a observación i, de novo, non intervén no produto $X^T \tilde{W}_{(i)} X$. Analogamente, como a columna i de $X^T \tilde{W}_{(i)}$ é nula, ó facer $X^T \tilde{W}_{(i)}\, \boldsymbol{Y}$, o termo Y_i tampouco estará participando.

En definitiva, nos sucesivos produtos matriciais de $\left(X^T \tilde{W}_{(i)} X\right)^{-1} X^T \tilde{W}_{(i)}\, \boldsymbol{Y}$, a observación i-ésima non está intervindo. Por esa razón, ó multiplicar matricialmente pola esquerda pola matriz fila X_i, estarase obtendo a predición para a observación i-ésima sen tela en conta, que é o que se denota como $\hat{Y}^h_{i(i)}$.

Unha vez obtidas todas as predicións $\hat{Y}^h_{i(i)}$, procedemos a calcular o erro de validación cruzada $CV(h)$ como a suma dos cadrados das diferenzas das predicións de validación cruzada $\hat{Y}^h_{i(i)}$ con respecto ós valores reais da explicativa Y_i, tal e como indiquei en (3.12). Escolleremos como parámetro ventana óptimo $\hat{h}_{CV}$ aquel que minimice dito erro.

En resumo, o proceso de validación cruzada consiste en calcular para cada h do conxunto de parámetros ventanas hs, a diferenza entre a predición obtida para cada dato sen telo en conta, $\hat{Y}_{i(i)}$ e o valor observado da variable resposta Y_i e toma como h óptimo aquel parámetro ventana para o que a suma desas diferenzas ó cadrado sexa menor.

Deste xeito estamos garantindo certa independencia polo feito que evitar que na propio axuste dunha observación estea intervindo a propia observación.

Pero, chegados a este punto decatémosnos de que a validación cruzada calcula a predición de cada un dos valores observados da variable explicativa, polo tanto a secuencia hs deberá ter parámetros ventanas para os que poidamos facer predición en todos os x_i. O que veño a dicir é que non poderemos considerar valores de h excesivamente pequenos pois impediranos aplicar o modelo lineal local en zonas de baixa densidade de datos.

Recordemos que no Teorema 2.4 vimos que para obter a predición nun punto x_0 usando kernels con soporte $[-1, 1]$ debíamos considerar unha ventana h para a que houbese polo menos dúas observacións con valores distintos da variable explicativa x e a unha distancia de x_0 menor ca h. Polo tanto, para poder aplicar validación cruzada deberemos considerar valores de h que garantan esta condición para cada x_i. Formalízoo na seguinte observación.

Observación 3.6. *Dado un conxunto de observacións $\{x_i, Y_i\}_{i=1}^n$ no que queremos determinar o h óptimo mediante validación cruzada e empregando un kernel que fora do intervalo (-1,1) asigna pesos nulos. A un conxunto de parámetros ventana $hs \subset \mathbb{R}^+$, poderáselle aplicar validación cruzada se e só se, para todo $h \in hs$ se verifica:*

$$\textit{Para cada } i \in \{1, ..., n\},\ \exists\ j, k \in \{1, ..., n\} \setminus \{i\} \textit{ tal que: } \begin{cases} x_j \neq x_k, \\ x_j, x_k \in (x_i - h, x_i + h). \end{cases}$$

Basicamente, a condición indica que para que a un conxunto de parámetros ventana hs se lle poida aplicar validación cruzada con un kernel definido en $[-1, 1]$, deberá haber, para todo h deste conxunto e toda observación i, cando menos outras dúas observacións j, k a unha distancia de x_i menor de h e con distintos valores da variable explicativa ($x_j \neq x_k$).

Esta condición débese cumprir para aplicar validación cruzada con kernels que fóra de $(-1, 1)$ asignan peso nulo. Para os kernels definidos en todo $\mathbb{R}$, como o de Gauss, non é necesario, aínda que si nos podemos atopar con erros computacionais se empregamos valores de h moi baixos.

Chegados a este punto, é importante pensar na programación do proceso de validación cruzada. Nótese que a elección dun h óptimo mediante validación cruzada leva asociada un alto tempo de execución (en termos relativos) xa que está axustando un modelo lineal local para cada h do conxunto de parámetros ventana hs. Polo tanto sería interesante que, xa que se van a facer moitos axustes, tantos como parámetros ventana, que sexamos o máis eficaces posible na obtención destes axustes.

A reescritura que presento na Sección 2.6 ven a cumprir este obxectivo, alixeirando o tempo de espera no proceso de validación cruzada. De feito, foi esta a razón que me levou a plantexarme abordar unha reescritura que optimizase os tempos de execución, pois percibía que o tempo de espera era moi alto e confiaba en que se puidese reducir. Veremos no seguinte capítulo que efectivamente, dita reescritura acada tempos máis competitivos.

Observación 3.7. *Para facer validación cruzada empregando a reescritura presentada na Sección 2.6, emprégase a expresión que vimos para a predición en (2.34) ignorando (anulando) o sumando correspondente á observación i-ésima nos sumatorios que dan o valor das compoñentes $S_{n,0}$, $S_{n,1}$, $S_{n,2}$, $T_{n,0}$ e $T_{n,1}$, pois así estaremos garantindo que a observación i-ésima non intervén na súa propia predición. Computacionalmente é moi sinxelo e reduce tempos de execución como veremos na Sección 4.4.*

Capítulo 4

Aplicación a datos reais

Neste último capítulo aplicaremos todo o desenvolvemento teórico visto ó largo do traballo ó caso real de varios estados nas eleccións de Estados Unidos do 2020 facendo uso do software estadístico R (R Core Team, 2021). Modelaremos a porcentaxe de votos estimada polas enquisas mediante regresión lineal local, xa que encaixa á perfección neste exemplo, pero non é o único pois tamén poderíamos aplicar outros modelos, tales como suavizados tipo splines, que se poden consultar nos Capítulos 4 e 5 de Eubank (1999).

Aplicaremos os contidos vistos a tres estados, Georgia, Pensilvania e Florida, xa que nestes, a diferenza doutros estados, o sentido do voto non estaba claro e polo tanto eran decisivos en ditas eleccións, polo que o seu estudo é de especial interese.

Cómpre ter en conta que nas eleccións presidenciais de Estados Unidos, o reparto de escanos non é para nada proporcional. Nelas o candidato que máis votos ten en cada estado leva todos os compromisarios dese estado en bloque (excepto en dous estados que non analizaremos por non seren significativos no resultado final). Este feito permítenos aplicar facilmente o noso modelo, pois estimaremos a intención de votos para Biden. Mais concretamente, traballaremos con unha porcentaxe de votos reescalada, como se só se presentasen Biden e Trump, xa que na práctica son os únicos candidatos factibles.

A continuación aplicaremos o modelo ó conxunto de datos (Bycoffe e outros, 2020), obtido o 29 de novembro do 2020 do sitio web FiveThirtyEight, que se adica a agregar enquisas e ó seu análise. O arquivo ".csv" que utilizaremos ten por nome *president_polls.csv*, e contén unha gran cantidade de enquisas dun amplo rango de datas e de todos os estados que asignan unha estimación da porcentaxe de voto para cada candidato. En canto ó código empregado ó longo deste capítulo, non se empregou ningunha función predefina de R, senón que se programaron as propias fórmulas obtidas no traballo.

4.1. Preparando os datos

Cargamos o arquivo de datos (Bycoffe e outros, 2020) e interesámosnos inicialmente pola variable `created_at`, que contén as datas nas que se elaboran as enquisas, e convertémola a formato "`Date`", que é o empregado en R para datas.

As outras variables que nos interesan son o estado sobre o que trata a enquisa (`state`), a data na que se realizou (`created_at`), o candidato (`answer`) e a porcentaxe estimada para dito candidato (`pct`). Ademais, tamén precisaremos a variable que identifica as enquisas (`question_id`) xa que neste conxunto de datos as porcentaxes de cada candidato danse por separado, e necesitaremos esta variable para preparar o conxunto de datos sobre o que traballaremos, un que para cada enquisa, só recollerá a súa estimación para Biden.

Quedámosnos cas variables que nos interesan e facendo uso do comando `head` vemos como é o conxunto de datos:

```
  question_id         state   created_at      answer    pct
1      136283          Iowa   2020-11-02       Biden   49.0
2      136283          Iowa   2020-11-02       Trump   48.0
3      136322  Pennsylvania   2020-11-02       Biden   48.4
4      136322  Pennsylvania   2020-11-02       Trump   49.2
5      136322  Pennsylvania   2020-11-02   Jorgensen    1.4
6      136491       Florida   2020-11-02       Biden   47.0
```

Como podemos ver na quinta observación, este conxunto de datos ofrece estimacións para todos os candidatos, non só para Biden e Trump. Porén, como na práctica estes dous son os únicos factibles, quedámosnos só cas filas correspondentes a estes candidatos.

A continuación, definimos como variable resposta a porcentaxe de votos para Biden reescalada, que se obtén dividindo a estimación para Biden entre a suma das estimacións para Biden e Trump. Na construcción desta variable é necesaria a variable `question_id` que identifica cada enquisa. Esta porcentaxe reescalada correspóndese ca porcentaxe de votos que obtería Biden nun contexto bipartidista, que na práctica é o que ocorre xa que os outros candidatos non son factibles, de aí o interese de estudar esta variable que ademais tamén nos permite saber o porcentaxe para Trump (simplemente restándolle a 100 a estimación de Biden). Finalmente defino un `data.frame` cas variables que nos interesan e ordenamos cronolóxicamente as observacións e considerando as enquisas a partir do 1 de maio de 2020 (xa que para datas anteriores a densidade de enquisas é moito menor).

Escollemos como variable explicativa x o número de días desde o 1 de maio (que é o primeiro día que estamos considerando), sendo este o día 1. E como variable resposta Y a porcentaxe estimada pola enquisa para Biden (reescalada), como xa indiquei.

Explicarei o proceso para facer a estimación no estado de Georgia e empregando o kernel de Gauss exposto na Sección 2.3. Posteriormente mostrarei os resultados para os outros kernels e outros estados (Pensilvania e Florida), pois así teremos visto o funcionamento en estados onde o modelo se comporta diferente ou nos que o resultado diferiu.

Antes de empezar a aplicar a regresión lineal local, convén ter unha idea de como son os datos. Na Figura 4.1 preséntase unha gráfica dos datos brutos.

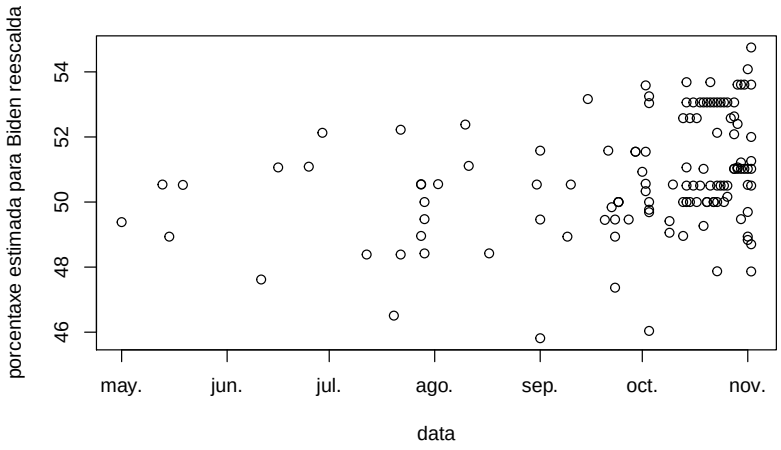


Figura 4.1: Observacións para o estado de Georgia, no eixo x a data da enquisa e no eixo y a porcentaxe reescalada que dita enquisa lle outorga a Biden.

Efectivamente, apréciase unha notable diferenza na densidade dos datos. Isto corrobórase dun xeito claro no histograma de densidade de observacións recollido na Figura 4.2.

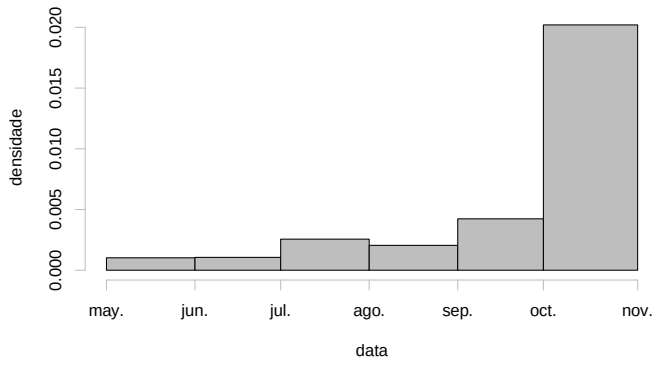


Figura 4.2: Histograma que mostra a densidade de datos en cada mes.

Finalmente para acabar de preparar os datos, construímos a matriz X que se indicaba en (2.14) na Sección 2.4, que ten tantas filas como observacións e de dúas columnas, a primeira con 1's e a segunda co valor da explicativa correspondente á observación, que é o número de día no que se fai a enquisa contado dende o 1 de maio do 2020.

4.2. Elección da ventana óptima

Para achar o parámetro ventana óptimo procedemos por **validación cruzada**, que se estudou na Sección 3.3, e cuxa idea básica consiste en partir dun conxunto hs de posibles parámetros ventana e escoller como óptimo aquel que mellor prediga as observacións, sen empregar a propia observación na súa predición.

En primeiro lugar definimos unha secuencia hs de parámetros ventana. Non podemos empezar dende h baixos xa que como se amosou nas Figuras 4.1 e 4.2, nos primeiros meses hai menos datos, polo que se tomamos ventanas baixas non se poderá axustar por ausencia de observacións en veciñanzas pequenas nos primeiros meses. Veremos despois que este impedimento non supón ningún problema xa que os valores de h máis baixos terán un alto erro e polo tanto non van a ser óptimos. Na Observación 3.6 analizabamos a condición que tiña que cumprir o conxunto hs para que se lle puidese aplicar validación cruzada (con kernels definidos en $[-1, 1]$ aínda que sempre é conveniente que se cumpra) e esta era que para todo $h \in hs$ e toda observación i houbese cando menos, outras dúas observacións a unha distancia de x_i menor de h e con distintos valores da variable explicativa. Propoñémosnos estudar as ventanas comprendidas entre medio mes (15 días) e 6 meses (180 días).

Empregando un bucle, calculamos para cada h do conxunto de parámetros ventana hs o erro de validación cruzada, que se basea no promedio do cadrado das diferenzas de cada valor da variable resposta con respecto ó seu valor da súa predición sen ter en conta a propia observación na súa predición, como se detallou na Sección 3.3. Representamos estes erros na Figura 4.3, na cal se ve que efectivamente para os h pequenos o erro de validación cruzada é claramente maior, co cal, como xa adiantei é evidente que os valores máis baixos de h non van ser os óptimos. Neste caso o h óptimo é $h_{VC} = 55$.

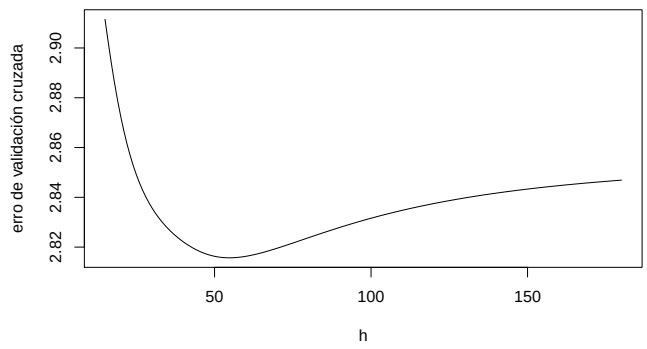


Figura 4.3: Gráfica dos erros de validación cruzada (eixo y) para os h entre 15 e 180 (eixo x).

4.3. Axustes e predición

Unha vez que temos o h óptimo, xa podemos calcular o axuste de cada observación, cuxa expresión (2.27) se deduciu na Sección 2.5:

$$\hat{Y}_i \;=\; \hat{m}(x_i) \;=\; e_i^T X \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y}, \qquad\qquad i \in \{1, \dots, n\}.$$

Realizamos o axuste utilizando R que representamos ca axuda da función `plot`, xunto cos puntos da mostra (x_i, Y_i), obtendo a Figura 4.4, na que ademais represento a curva $y = 50\,\%$, a partir da cal a estimación de votos para Biden é maior que para Trump.

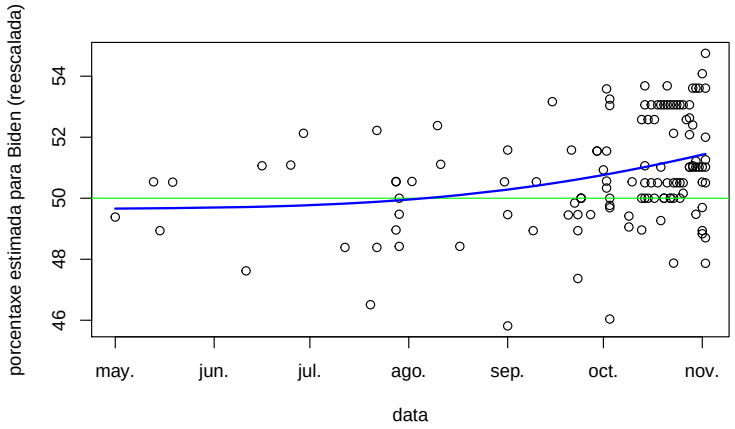


Figura 4.4: Observacións e curva de axuste para o estado de Georgia empregando o kernel de Gauss e o h óptimo segundo validación cruzada ($h_{VC} = 55$). No eixo x recóllese a data e no eixo y a porcentaxe reescalada que dita enquisa lle outorga a Biden. En cor verde a recta $y = 50$.

Como vemos, hai unha tendencia crecente que se vai acelerando co paso dos meses e que alcanza o máximo os días previos ós comicios. Isto, unido ó feito de que dende agosto se superou a barreira do 50 %, mantendo unha tendencia cada vez máis crecente, fainos esperar unha vitoria para Biden en Georgia, aínda que sexa axustada.

Recordamos a expresión para a predición $\hat{m}(x_0)$ á que se chegou en (2.25):

$$\hat{m}(x_0) = X_0 \left(X^T W X\right)^{-1} X^T W\, \boldsymbol{Y},$$

onde X_0 denota o vector fila $X_0 \;:=\; \begin{pmatrix} 1 & x_0 \end{pmatrix}$.

Executando o código correspondente en R obtemos a **predición** para o día dos comicios, o día 3 de novembro de 2020, sendo entón x_0 o número de días contados dende que empezou maio, que neste caso é $x_0 = 187$. A predición que obtemos é de $\hat{Y}_0 = 51{,}47$, que se corresponde ca porcentaxe de votos estimada para Biden (reescalada). Isto interprétase como un 51,47 % de votos estimados para Biden nun contexto puramente bipartidista.

Repetimos o proceso para os outros kernels e obtemos para cada un deles o h óptimo de validación cruzada, h_{CV}, que empregamos para obter a predición o día das eleccións $\hat{Y}_0$. Na Táboa 4.1 recollo estas predicións $\hat{Y}_0$ xunto co erro de validación cruzada asociado a escoller ese h_{CV}, que se denota por $CV(h_{CV})$ e cuxa expresión se viu en (3.12).

Táboa 4.1: Na primeira columna, predición para o día das eleccións no estado de Georgia empregando distintos kernels e o parámetro ventana óptimo segundo validación cruzada, h_{VC}. Na segunda columna, o erro de validación cruzada asociado a dito h_{CV} redondeado a tres decimais.

	$\hat{Y}_0$	$CV(h_{CV})$
Gauss	51,47 %	2,816
Epanechnikov	51,50 %	2,812
Uniforme	51,58 %	2,797
Biweight	51,51 %	2,815
Triweight	51,48 %	2,816

Efectivamente, os valores da porcentaxe reescalada móvense en torno ó 51,5 %. Isto interprétase como unha estimación dun 51,5 % de votos para Biden nun contexto bipartidista, co cal en principio esperaríamos unha vitoria deste, pero non demasiado folgada. Así mesmo, o erro de validación cruzada tamén se mantén en torno ó 2,81, confirmando así que a elección da función kernel apenas implica variacións na estimación nin no erro da mesma.

A razón pola cal a predición $\hat{Y}_0$ apenas varía é a que adiantabamos ó final da Sección 3.2. Víamos en (3.11) que o erro cadrático medio asintótico do parámetro ventana que o minimiza, h_{ECM}, só dependía dunha constante, $\sqrt[5]{\left(\int K^2(t)\ dt\right)^4 \cdot \left(\int t^2 \cdot K(t)\ dt\right)^2}$, que tiña practicamente o mesmo valor para os cinco kernels que tratamos como se recollía na Táboa 3.1, o que implicaba que a estimación sobre un punto x_0 sufrise unha variación mínima.

Nas eleccións, Biden gañou cun 50,1 % dos votos fronte ó 49,9 % de Trump (valores reescalados), co cal, a vitoria foi axustada, máis incluso do que parece intuírse co noso modelo, o que nos fai pensar que as enquisas foron lixeiramente optimistas respecto a Biden.

A continuación, repetimos o proceso para os outros dous estados que propuxen, Pensilvania e Florida. En ambos farei uso do kernel de Gauss, aínda que como vimos no capítulo anterior e acabo de recordar, a repercusión elección dun ou outro kernel na predición é mínima. Na Táboa 4.2 recóllense os h óptimos que devolve a validación cruzada, así como a predición para o día das eleccións nos estados de Pensilvania e Florida.

Táboa 4.2: Parámetros ventana óptimos por validación cruzada e estimación da porcentaxe de votos reescalada para Biden en Pensilvania e Florida, empregando o kernel de Gauss.

	h_{VC}	$\hat{Y}_0$
Pensilvania	136	52,82 %
Florida	29	50, 90 %

Por outra parte, na Figura 4.5 represéntase a curva de axuste xunto cas observacións, traballando co kernel de Gauss e o h óptimo elixido por validación cruzada.

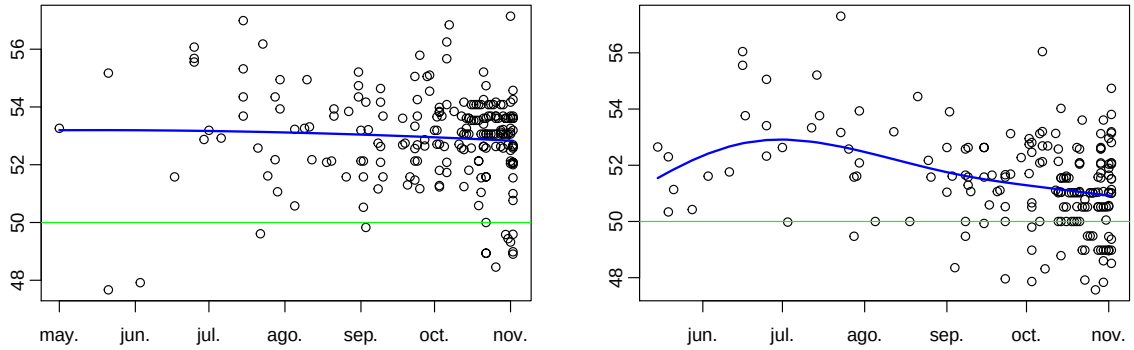


Figura 4.5: Observacións e curva de axuste para Pensilvania (esquerda) e Florida (dereita) co kernel de Gauss e o h óptimo de validación cruzada ($h_{VC} = 136$ para Pensilvania e $h_{VC} = 29$ para Florida). No eixo x, a data da enquisa e no eixo y a porcentaxe de votos (reescalada) para Biden.

Podemos apreciar na primeira gráfica da Figura 4.5 que a curva de axuste apenas sufriu alteracións, manténdose a porcentaxe estimada para Biden (reescalada) por encima do 50 % dende maio. Se nos fixamos na predición Y_0 que se recolle na Taboa 4.2, esta alcanza un valor máis dun punto porcentual superior ó que tomaba caso de Georgia, o que nos fai esperar unha vitoria de Biden máis folgada ca no caso de Georgia. Efectivamente, neste estado gañou Biden cun 50,6 % dos votos fronte ó 49,4 % que obtivo Trump (valores reescalados), unha diferenza máis folgada có 0,2 % que distaron os resultados en Georgia.

No caso de Florida, a situación é completamente diferente que nos casos de Georgia e Pensilvania, xa que en primeiro lugar, a porcentaxe de votos que o promedio local de enquisas axusta vaise reducindo a partir de xullo, como percibimos na segunda gráfica da Figura 4.5. Por outro lado, aínda que a estimación que obtemos para o día das eleccións é dun 50,90 %, e polo tanto superior ó 50 %, neste estado gañou Trump. Isto tampouco nos debe sorprender pois estamos analizando estimacións puntuais que teñen variabilidade, pero o que si sorprende é que gañou dun xeito relativamente contundente cun 51,7 % dos votos fronte ó 48,3 % que obtivo Biden (valores reescalados), o que nos leva a conxecturar que as enquisas sobrevaloraron as posibilidades de Biden en Florida.

4.4. Aspectos computacionais

Unha vez aplicados os resultados que fomos tratando ó longo do traballo ó caso real das eleccións de EEUU de 2020, só falta por revisar a reescritura do modelo proposta na Sección 2.6. A súa programación é moi sinxela pois redúcese a aplicar a expresión dada en (2.25) e a continuación darei unhas indicacións para unha execución aínda máis eficaz.

Fixémosnos que nas expresións dos elementos $S_{n,0}$, $S_{n,1}$, $S_{n,2}$, $T_{n,0}$ e $T_{n,1}$ presentados na Sección 2.6, (concretamente en 2.31 e 2.32), os termos son moi parecidos:

$$S_{n,0} = \sum_{i=1}^{n} K_h(x_i - x_0), \qquad S_{n,1} = \sum_{i=1}^{n} x_i \cdot K_h(x_i - x_0), \qquad S_{n,2} = \sum_{i=1}^{n} x_i \cdot K_h(x_i - x_0),$$

$$T_{n,0} = \sum_{i=1}^{n} Y_i \cdot K_h(x_i - x_0), \qquad T_{n,1} = \sum_{i=1}^{n} x_i \cdot Y_i \cdot K_h(x_i - x_0),$$

polo que de cara a unha implementación en R (R Core Team, 2021) máis eficiente, definiremos dous vectores auxiliares; un que denotarei por `aux` e cuxas entradas serán $K_h(x_i - x_0)$ e outro, que denotarei por `x_aux`, que terá por entradas $x_i \cdot K_h(x_i - x_0)$ e que resultará de facer o produto compoñente a compoñente do vector de valores observados $\boldsymbol{x}$ por `aux`.

$$\begin{aligned} \texttt{aux} &= \Big(K_h(x_1 - x_0) \quad \cdots \quad K_h(x_n - x_0)\Big), \\ \texttt{x\_aux} &= \Big(x_1 \cdot K_h(x_1 - x_0) \quad \cdots \quad x_n \cdot K_h(x_n - x_0)\Big). \end{aligned}$$

Nótese que $S_{n,0}$ é a suma das entradas de `aux` e $S_{n,1}$ das de `x_aux`. Por outra parte, $S_{n,2}$ é o produto escalar de `x` por `x_aux`, $T_{n,0}$ o de $\boldsymbol{Y}$ por `aux` e $T_{n,1}$ o de $\boldsymbol{Y}$ por `x_aux`. Entón unha vez calculados estes valores, só queda aplicar a fórmula da predición (2.34).

Obviamente, obtéñense os mesmos resultados, xa que estamos facendo o mesmo pero dun xeito máis eficiente, pois evitamos facer todas as operacións matriciais que esixía o método estándar para calcular $\hat{\boldsymbol{\beta}}$ en (2.18) e que elevan o tempo de execución.

Pero chegados a este punto, cabe preguntarse se compensa empregar esta reformulación ou se apenas é un cambio estético. Porén, como amosarei agora, nada máis lonxe da realidade xa que si se consegue unha redución de tempos de execución, que é especialmente notable no proceso de validación cruzada pois é o proceso máis lento.

Ca axuda do comando `proc.time()` de R e repetindo o proceso en bucle 100 veces fixen un promedio dos tempos de execución do proceso de validación cruzada, por unha parte usando a formulación nas primeiras seccións e por outra parte ca reformulación exposta na Sección 2.6 e mostro na Táboa 4.3 eses tempos.

Táboa 4.3: Comparativa de tempos medios de execución (en segundos) do proceso de validación cruzada para os distintos kernels da formulación estándar fronte a reescritura usando S_n e $\boldsymbol{T}_n$ para o estado de Georgia (número de observacións = 126).

	Formulación estándar	Empregando S_n - $\boldsymbol{T}_n$
Gauss	22,8	11,9
Epanechnikov	14,1	2,6
Uniforme	13,7	2,4
Biweight	13,8	2,8
Triweight	14,9	3,2

Á vista da Táboa 4.3, confirmamos que a reescritura que involucra S_n e $\boldsymbol{T}_n$ bríndanos unha redución de tempos considerable, ó redor de 5 veces menos, a excepción do kernel de Gauss, no que se reduce ata case a metade, o que non deixa de ser unha redución interesante. Porén, aínda que o aforro de tempo é evidente para este estado do que só hai 126 observacións, é moito máis rotundo para o estado de Pensilvania, do que hai 212 observacións e onde os tempos de espera se reducen entre seis e sete veces como se aprecia na Táboa 4.4, salvo o de Gauss, que aínda así mellora, tardando dúas veces e media menos.

Táboa 4.4: Comparativa de tempos medios de execución (en segundos) do proceso de validación cruzada para os distintos kernels da formulación estándar fronte a reescritura usando S_n e $\boldsymbol{T}_n$ para o estado de Pensilvania (número de observacións = 212).

	Formulación estándar	Empregando S_n - $\boldsymbol{T}_n$
Gauss	80,6	32,6
Epanechnikov	53,4	8,3
Uniforme	51,8	6,6
Biweight	52,4	8,7
Triweight	54,6	9,1

En efecto, as Táboas 4.3 e 4.4, lévannos a concluír que a reescritura é moito máis eficaz, e que se notará tanto máis canto maior sexa o tamaño mostral.

4.5. Variable resposta vectorial (varios partidos)

Todo o desenvolvemento abordado neste traballo refírese ó caso dunha variable resposta real, que na práctica se corresponde con eleccións bipartidistas ou nas que só hai dous candidatos factibles. O caso das eleccións estadounidenses axústase a esta situación pois aínda que formalmente é multipartidista, o feito de que o gañador leve todos os delegados electorais, convérteo nun escenario que marxina ós candidatos que non teñen posibilidade de conseguir representación, quedando só dous factibles. Non obstante, o máis habitual son eleccións multipartidistas con máis de dous candidatos que poden conseguir representación.

Aínda así, o desenvolvemento feito neste traballo pódese xeneralizar ó caso de varios partidos, pasando a usar unha variable vectorial no canto dunha escalar como se fixo neste traballo. Ata o momento, $\boldsymbol{Y}$ era unha variable real que daba a porcentaxe de votos para un candidato, pero ó só haber dous factibles, a porcentaxe do outro non é máis cá diferenza ata chegar ó 100 % (por iso reescalabamos). Non obstante, nun escenario con tres ou máis candidatos con posibilidade de conseguir representación, $\boldsymbol{Y}$ será unha variable vectorial.

Na práctica, estaremos facendo un promedio local conxunto dos p partidos a partir de n enquisas que recollerán información sobre os p partidos.

Entón, a función de regresión xa non será unha función real correspondente ca estimación para un candidato, senón que será unha función vectorial que recollerá as estimacións para os p candidatos, $m : \mathbb{R} \to \mathbb{R}^p$, onde a compoñente j da mesma, $m_j : \mathbb{R} \to \mathbb{R}$ recolle as porcentaxes de voto para o partido j.

En consecuencia, estarase facendo regresión lineal local sobre unha variable resposta vectorial $\boldsymbol{Y}$, na que a súa compoñente j, que podemos denotar por Y^j se corresponde ca porcentaxe de votos para o partido j. Así, o vector de valores observados $\boldsymbol{Y}$ pasará a ser unha matriz $\mathcal{Y}$ con tantas filas como observacións e tantas columnas como partidos:

$$\mathcal{Y} = \begin{pmatrix} Y_1^1 & Y_1^2 & \cdots & Y_1^p \\ Y_2^1 & Y_2^2 & \cdots & Y_2^p \\ \vdots & \vdots & \ddots & \vdots \\ Y_n^1 & Y_n^2 & \cdots & Y_n^p \end{pmatrix} \in \mathbb{R}^{n \times p},$$

onde o elemento Y_i^j recolle a porcentaxe de votos que a enquisa i lle outorga ó partido j. Deste xeito, cada fila recolle as estimacións de votos que cada enquisa lle outorga a cada un dos candidatos, que se corresponden cas columnas.

A variable explicativa x seguirá sendo escalar, correspondéndose ca data na que se realiza a enquisa. Así mesmo, as matrices X e W manterán a súa estrutura posto que ambas se definen exclusivamente á variable explicativa tal e como as definimos en (2.14).

Máis alá do feito de que a variable resposta pase a ser vectorial, o desenvolvemento do traballo (excepto a reescritrura do modelo presentada na Sección 2.6) segue sendo válido e os resultados son análogos. Por exemplo, a expresión para a predición á que se chegaba en (2.23) segue sendo válida pois considerando unha matriz $\mathcal{Y}$ en vez dun vector $\boldsymbol{Y}$, obtemos efectivamente un vector ca predición para cada un dos partidos. En canto ó estudo de nesgo e varianza, pasaría a ser en termos vectoriais candidato a candidato, análogos. Pódese consultar o caso vectorial por exemplo na Sección 5.2 de Wand, M.P. e Jones, M.C.

Como vemos, este modelo tamén se pode aplicar a un escenario multipartidista, sen máis que traballar cunha variable resposta vectorial. O feito de considerar unha variable resposta real débese a que o obxectivo era aplicar o modelo lineal local á predición de votos nas eleccións presidenciais de EEUU do 2020, xa que foron as que se celebraron nos primeiros meses de elaboración deste traballo. Ademais, a elección dunha variable resposta real tamén nos brindou unha maior facilidade para a comprensión, interpretación e visualización dos contidos presentados ó longo dos capítulos e seccións deste traballo.

Bibliografía

Bycoffe, A., King, R., Koeze, E., Mehta, D., Mithani, J. e Wolfe, J. (2020). *president_polls.csv*. FiveThirtyEight. Recuperado o 29 de novembro do 2020 de `https://projects.fivethirtyeight.com/polls/president-general/national/`.

Eubank, R.L. (1999). *Nonparametric regression and spline smoothing*. Marcel Dekker. New York.

Fan, J. e Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC. London.

Graefe, A. (2015). German Election Forecasting: Comparing and Combining Methods for 2013. *German Politics*, *24(2)*, 195-204, DOI: `10.1080/09644008.2015.1024240`.

Peña, D. (2002). *Regresión y diseño de experimentos*. Alianza Editorial. Madrid.

Peña, D. (2005). *Fundamentos de Estadística*. Alianza Editorial. Madrid.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Recuperado o 16 de febreiro de 2021 de `https://www.R-project.org/`.

Trench, W.F. (2013), *Introduction to Real Analysis*. Trinity University. Ed. 2.04. Recuperado o 22 de xullo de `https://digitalcommons.trinity.edu/mono/7/`.

Wand, M.P. e Jones, M. C. (1995) *Kernel Smothing*. Chapman and Hall. London.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: `https://ggplot2.tidyverse.org`.