Traballo Fin de Máster

Modelos de Regresión Aditiva Estruturada. Aplicacións en biomedicina

Jenifer Espasandín Domínguez

Máster en Técnicas Estadísticas

Curso 2014/2015

Proposta de Traballo Fin de Máster

Título en galego: Modelos de Regresión Aditiva Estruturada. Aplicacións en biomedicina
Título en español: Modelos de Regresión Aditiva Estruturada. Aplicaciones en biomedicina
English title: Structured Additive Regression Models. Applications in Biomedicine
Modalidade: A
Autor/a: Jenifer Espasandín Domínguez
Director/a: Carmen Cadarso Suárez
Breve resumo do traballo: Os modelos de regresión aditiva estruturada son un tipo de regresión moderna que permiten modelar de forma flexible posibles efectos non lineais das covariables contínuas ademais de incluír por exemplo efectos xeográficos ou espazo-temporais, ou mesmo modelos de supervivencia. O obxectivo deste traballo é revisar a literatura existente sobre este tipo de modelos e aplicalos en datos biomédicos.

Dona Carmen Cadarso Suárez, informa que o presente Traballo Fin de Máster titulado:

Modelos de Regresión Aditiva Estruturada. Aplicacións en biomedicina

foi realizado baixo a súa dirección por Dona Jenifer Espasandín Domínguez para o Máster en Técnicas Estadísticas. Estimando que o traballo está terminado, dá a súa conformidade para a súa presentación e defensa ante un tribunal.

Santiago de Compostela, 7 de Xullo 2015

A directora:	A autora:
Dona Carmen Cadarso Suárez	Dona Jenifer Espasandín Domínguez

How can it be that mathematics, being after all a product of human thought independent of experience, is so admirably adapted to the objects of reality!

Albert Einstein

Índice xeral

Resumo xiii

Abstract xv

Prefacio xvii

1. Conceptos básicos de suavización en regresión 1

1.1. Suavización Univariante 2

1.1.1. Splines Polinómicos 4

1.2. Splines Penalizados (P-splines) 19

1.3. Suavización bivariante 23

1.3.1. Funcións base radiais e *Thin Plate Splines* 30

1.4. Técnicas de suavización q-dimensionais 34

1.5. Técnicas de suavización espacial 34

1.5.1. Cadeas aleatorias de Markov 35

1.6. Resumo sobre os diferentes enfoques das aproximacións de penalización 40

2. Modelos de Regresión Aditiva Estruturada **41**

2.1. Introdución 42

2.2. Distribucións previas 45

2.2.1. Modelado dos efectos das covariables continuas e escalas temporais . . 47

2.2.2. Modelado dos efectos espaciais 48

2.2.3. Indicadores de grupos e efectos espaciais non estruturados 50

2.2.4. Modelado de interacións 51

2.3. Representación como modelos mixtos 52

2.4. Inferencia baseada na metodoloxía dos modelos GLMM 54

2.5. Modelos de regresión estruturada de risco 57

2.5.1. Capacidade de discriminación do modelo 59

2.5.2. Índice C de concordancia 62

2.6. Implementación de modelos STAR: Bayes X 63

3. Patróns espaciais na taxa de abstinencia do alcohol **65**

3.1. Introdución 65

3.2. Descripción da base de datos 67

3.3. Metodoloxía estadística 70

3.4. Resultados 71

3.5. Discusión 75

4. Supervivencia do síndrome coronario agudo na área sanitaria de Santiago de Compostela **77**

4.1. Introdución 77
4.2. Descrición da base de datos 78
4.3. Formulación do modelo 80
4.4. Resultados 81
4.5. Capacidade de discriminación do modelo 90
4.5.1. Curvas ROC tempo dependentes 90
4.5.2. Índice C de concordancia 93

5. Comentarios finais **97**

Bibliografía **101**

Resumo

Os modelos de regresión aditiva estruturada (STAR, Structured Addittive Regression Models, Fahrmeir e Kneib, 2013) permiten modelar de forma flexible posibles efectos non lineais das covariables continuas, interaccións complexas e efectos espazo-temporais, ou mesmo realizar estudos de supervivencia con datos censurados.

Neste Traballo Fin de Máster, revisaremos a literatura existente sobre a metodoloxía estatística dos modelos STAR, e demostraremos a utilidade e versatilidade de ditos modelos na práctica, coa aplicación en dous estudios biomédicos. No Capítulo 1 presentaremos os conceptos básicos de suavizado en regresión necesarios para desenvolver a teoría fundamental dos modelos STAR, que será presentada no Capítulo 2 deste traballo.

Nas últimas décadas, son moitos os estudos nos que se considera a área xeográfica como un factor importante a ter en conta nos estudos clínicos, posto que a análise do impacto destes efectos é especialmente importante para capturar posibles heteroxeneidades espaciais. Neste sentido, no Capítulo 3 investigaremos as tendencias espaciais na taxa de abstinencia do alcohol en Galicia empregando os modelos STAR cunha resposta de Poisson.

No Capítulo 4, presentaremos unha extensión dos modelos clásicos de supervivencia de Cox (1972) baseándonos nos modelos STAR que nos permitirán investigar as desigualdades xeográficas na supervivencia dos pacientes diagnosticados de síndrome coronario agudo na área sanitaria de Santiago de Compostela.

Para finalizar, no Capítulo 5, a modo de conclusión exporemos algunhas limitacións dos modelos STAR e introduciremos posibles liñas de investigación futuras.

Abstract

Structured Additive Regression models (STAR Models, Fahrmeir and Kneib, 2013) allows for modeling flexible non-linear effects of continuous covariates including, for example, temporal-spatial effects and survival studies.

In this master thesis, as indicated in the title, we will review exisiting literature about these STAR models and we will apply them to various biomedical data. In addittion, in the first chapter we will present the basic concepts of smooth regression that are necessary to develop the fundamental theory about STAR models.

Within the last decades, there have been a lot of studies which consider geographical areas as a decisive factor to include in clinical studies because the analysis of the impact of these effects is especially important to capture possible spatial heterogeneities. In this sense, in Chapter 3, we will investigate spatial tendencies of the rate of alcohol withdrawal syndrome in Galicia using star models with Poisson response.

In Chapter 4, we will present an extension of the classic Cox models of survival (Cox, 1972), based on STAR models that allow us to investigate the geographical inequalities of patients survival who were diagnosed with acute coronary syndrome in the sanitary area of Santiago de Compostela.

To sum up, in the last chapter, we will show some limitations of STAR models and we will introduce possible future investigations.

Prefacio

En biomedicina e outros campos de aplicación das metodoloxías estatística, as técnicas de regresión son moi útiles pois permiten modelar unha variable resposta de interese, y, en función dun conxunto de variables, $x_1, .., x_n$. Dependendo do tipo de variable resposta que pretendamos analizar (continua, binaria, categórica ou de conteo) e das covariables que posúa o noso estudo, existen diferentes modelos. Neste Traballo Fin de Máster, faremos fincapé en exemplos máis complexos nos que é posible incluír efectos temporais, ou mesmo variables que nos permiten describir distribucións espaciais ou localizacións xeográficas.

Unha das familias de modelos de regresión máis empregada na práctica, baséanse nos modelos lineais xeneralizados (GLM, McCullagh e Nelder, 1989) para respostas da familia exponencial e preditores lineais. Unha das principais limitacións destes modelos, é a asunción da linearidade dos efectos das covariables. En moitas ocasións, supoñer un efecto estritamente linear para as variables preditoras continuas pode non ser apropiado. Unha maneira de solucionar este inconveniente baséase na utilización dos modelos xeneralizados aditivos (GAM, Hastie e Tibshirani, 1990; Wood, 2006). Nestes modelos substitúese o preditor lineal por un aditivo semiparamétrico, que nos permite estimar os efectos non lineais das covariables. Non obstante, estes modelos teñen algunha limitación, por exemplo, non permiten incorporar de forma doada efectos aleatorios, ou espazo-temporais...

Nos últimos anos, os modelos de regresión aditiva estruturada (Structured Additive Regression, STAR, Fahrmeir et al., 2013) están a acadar moito interés por investigadores de diversas áreas, pois xeneralizan aos GLM e GAM. Os modelos STAR permiten incorporar dunha maneira unificada efectos suaves das covariables continuas (utilizando splines penalizados), efectos aleatorios, datos clúster, ou efectos espazo-temporais entre outros.

Como xa comentamos, o principal obxectivo deste Traballo Fin de Máster é presentar a

metodoloxía dos modelos STAR (Capítulo 2). Non obstante, antes de introducir estes modelos, precisamos desenvolver algunhas técnicas básicas de suavización en regresión, necesarias para poder formular os aspectos fundamentais dos modelos de regresión aditiva estruturada. A inferencia dos modelos STAR pódese realizar mediante métodos puramente Baiesianos, (Full Bayes, FB) ou aproximacións empíricas (Empirical Bayes, EB). Neste traballo centraremonos na inferencia empírica, na cal, tanto a varianza como os parámetros de suavización se consideran constantes descoñecidas e estímanse mediante aproximacións REML (Restricted Maximun Likelihood). Neste contexto, os efectos non lineais das covariables continuas, modelaranse mediante versións baiesianas dos splines penalizados (P-splines; Fahrmeir, Kneib e Lang, 2004), mentres que os efectos espaciais estimaránse empregando Cadeas Aleatorias de Markov (Rue e Held, 2005). Técnicas que presentaremos no vindeiro Capítulo.

Ademais do modelado das covariables, debemos ter especial coidado coa distribución da variable resposta. Os modelos de regresión clásicos, restrínxense a determinadas distribucións. Os modelos STAR son moi flexibles permitindo incorporar unha ampla variedade de variables resposta (da familia exponencial, respostas categóricas, ou memo tempos de supervivencia con datos censurados. Nos Capítulos 3 e 4, mostraremos a versatilidade dos modelos STAR, mediante dúas aplicacións biomédicas, cuias bases de datos correspondentes foron proporcionadas pola Unidade de Epidemioloxía Clínica do Hospital Clínico Universitario de Santiago de Compostela. No Capítulo 3 investigaremos as tendencias espaciais na taxa de abstinencia ao alcohol (AWS) en Galicia e a súa relación con varios factores socioeconómicos, empregando un modelo STAR con resposta unha distribución de Poisson.

Unha gran vantaxe destes modelos STAR e que nos permiten realizar estudos de supervivencia incluíndo datos censurados. Os modelos clásicos de supervivencia, adoitan modelar tanto a taxa de risco coma os efectos das covariables continuas de forma linear. Na práctica, tal e como podemos supoñer, isto pode chegar a ser moi restritivo, (Cadarso-Suárez et al., 2010). Neste traballo, presentaremos unha xeneralización dos modelos de supervivencia de Cox (1972) empregando a regresión aditiva estruturada. Esta formulación STAR permitirá modelar de forma flexible a taxa de risco e os efectos non lineais das covariables continuas, ademais de incluír de forma sinxela efectos espazo-temporais. A modo de ilustración, no Capítulo 4, analizaremos as desigualdades xeográficas na supervivencia dos pacientes ingresados cun diagnóstico de síndrome coronario agudo na área sanitaria de Santiago de Compostela.

Xa para rematar, no Capítulo 5, a modo de conclusión exporemos algunhas limitacións dos modelos STAR e introduciremos posibles liñas de investigación futuras.

Capítulo 1

Conceptos básicos de suavización en regresión

Como veremos no Capítulo 2, nos modelos STAR a resposta de interese explícase a través das covariables, permitindo que a forma destas relacións sexa flexible. Utilizando técnicas de suavización como os splines con penalizacións (P-Splines; Eliers e Marx, 1996) modelaremos os efectos das variables continuas. Ademais incorporaremos interaccións entre variables mediante produtos tensoriais; os efectos espaciais modelaranse empregando Cadeas Aleatorias de Markov (Rue e Held, 2005). Presentar estas técnicas será o obxectivo fundamental deste Capítulo.

En primeiro lugar, para dar comezo a este capítulo, na Sección 1.1 presentaremos as principais técnicas de suavización univariante e na Sección 1.2 presentaremos os splines con penalizacións para, finalmente, estender estes conceptos ao mundo multivariante (Sección 1.3 e 1.4). Na Sección 1.5, presentaremos varias técnicas de suavización espacial.

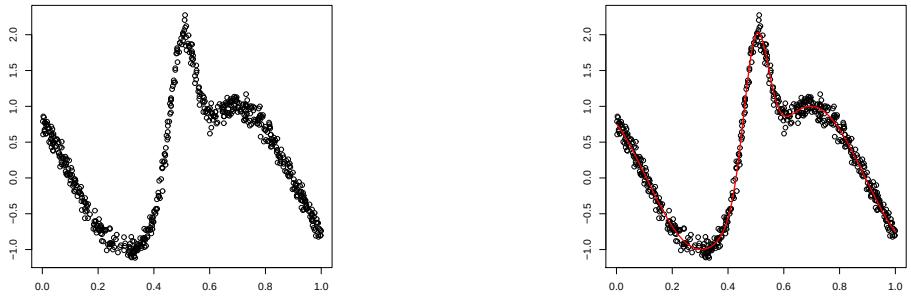
En moitas aplicacións prácticas un modelo puramente lineal non sempre é suficiente, por exemplo, naquelas situacións nas que os efectos das covariables sobre a resposta sexan dunha forma específica, non linear.

Consideraremos dúas primeiras aproximacións para modelar os efectos non lineais das covariables continuas baseadas en transformacións simples ou en polinomios (Fahrmeir et al., 2013). A pesar de que, nalgúns casos, estes métodos poden ser suficientes, en estudos comple-

xos ou mesmo cando hai moitas covariables involucradas, é inviable o seu uso. Por exemplo, no tocante ás posibles transformacións que podemos realizar, na práctica son limitadas, e en consecuencia os modelos resultantes non serán moi flexibles.

Para motivar este estudo, empregaremos os seguintes datos simulados con R, [1] que representamos na Figura 1.1. O panel da esquerda mostra o diagrama de dispersión dos datos simulados, mentres que no dereito, incorporamos o verdadeiro efecto da covariable.

Neste caso, o diagrama de dispersión suxire bastante ben a verdadeira estrutura dos datos, e como podemos observar a relación non é linear e parece bastante difícil aproximar esta función mediante simples polinómicos.



a) b)

Figura 1.1: Na Figura a) representamos o gráfico de dispersión dos datos simulados segundo o modelo $y = f(z) + \epsilon$, onde $f(z) = \sin(8z - 4) + 2\exp\left(-256(z - 0.5)^2\right) + \epsilon$, $\epsilon \in N(0, 0.3^2)$. En b) representamos ademais o verdadeiro efecto da covariable.

1.1. Suavización Univariante

Nesta sección presentaremos algunhas técnicas de suavización en regresión que permiten modelar de xeito flexible o efecto dunha covariable continua en función doutra variable tamén continua.

[1] R é unha linguaxe e un entorno especialmente desenvolvido para o traballo estatístico e representación gráfica de datos. Trátase dun programa moi empregado no campo da estadística totalmente gratuíto, que se pode descargar da seguinte páxina web: www.r-project.org.

En xeral, na regresión non paramétrica univariante suponse que podemos explicar a variable resposta, y, mediante unha función determinista en termos da covariable, z, engadindo un termo relativo ao erro ϵ:

$$y_i = f(z_i) + \epsilon_i, i = 1, \ldots, n.$$

Suporemos ademais, que os erros son independentes e identicamente distribuídos con:

$$\mathbb{E}(\epsilon_i) = 0, \text{ e } Var(\epsilon_i) = \sigma^2, i = 1, \ldots, n.$$

Da mesma forma que nos modelos lineais, séguese que: $\mathbb{E}(y_i) = f(z_i)$ e $Var(\epsilon_i) = \sigma^2, i = 1, \ldots, n$. É dicir, o valor esperado da variable resposta modélase a través de f.

Finalmente, por simplicidade, supoñemos ademais que a función f é continua e diferenciable.

Actualmente é indubidable que as técnicas de suavización teñen un papel moi relevante, (Durbán, 2008). Esta popularidade débese, en boa parte, a complexidade dos datos dos que se dispón actualmente, imaxes, microarrays, etc., que fan que un modelo paramétrico sexa inviable. Ademais, gracias aos actuais avances informáticos redúcense, cada vez máis, os custos de computación que supoñen axustar os modelos de suavizado. É habitual empregar o termo *non paramétrico* para referirse a estes modelos, pero este nome tan só é adecuado cando se empregan técnicas tipo núcleo, (en inglés, *kernel*). Debemos ter en conta, que en ocasións, algunhas técnicas de suavización non paramétricas con splines, a pesares do seu nome, son puramente paramétricas, xa que se determinan empregando moitos parámetros. Existen dous grandes enfoques no eido dos modelos de suavización con splines: splines de suavización (en inglés, *smoothing splines*) e splines de regresión (en inglés, *regression splines*). (Durbán, 2008.)

Os splines de suavización (ver, por exemplo, Green e Silverman (1994)) empregan tantos parámetros como observacións. Polo tanto, cando temos un número moi elevado de datos, a súa implementación non é eficiente. Por outra banda, os splines de regresión poden axustarse empregando o método de mínimos cadrados determinando previamente o número de nodos. Non obstante, esta selección de nodos faise mediante algoritmos bastante custosos computacionalmente. Por outro lado, os splines con penalizacións, (aos que chamaremos

P-splines) combinan o mellor de ambos enfoques: empregan menos parámetros que os splines de suavización, pero a selección dos nodos non é tan determinante como nos splines de regresión.

A razón fundamental para empregar P-splines é que son de baixo rango, é dicir, o tamaño da base empregada é moito menor que a dimensión dos datos. Isto contrasta co que ocorre cos splines de suavización, onde por cada dato hai un nodo polo que se debe traballar con matrices de grandes dimensións. No caso dos P-splines, o número de nodos non supera os 40, e polo tanto son eficientes computacionalmente, incluso cando se traballa con grandes cantidades de datos. Ademais, a introdución de penalizacións relaxa a importancia da elección do número de nodos e a súa localización, cuestión de gran importancia nos splines de baixo rango sen penalizacións, (ver, por exemplo Rice e Wu, 2001).

1.1.1. Splines Polinómicos

Como primeira aproximación á regresión non paramétrica, presentamos os splines polinómicos (*polynomial splines*) baseados nas ideas de regresión polinómica.

Nos modelos de regresión polinómica suponse que o efecto da covariable, z, sobre a variable resposta, y, é un polinomio de grao l:

$$f(z_i) = \gamma_0 + \gamma_1 z_i + \cdots + \gamma_l z_i^l. \tag{1.1}$$

Os coeficientes de regresión, $\gamma_i, i = 1, \ldots, l$, poden estimarse de xeito semellante aos modelos lineais, empregando o método de mínimos cadrados.

Na Figura 1.2, representamos varios modelos de regresión polinómica para os datos simulados anteriormente. Tal e como podemos observar, en ocasións, os modelos puramente polinómicos non son suficientes para estimar funcións non lineais. A modo de exemplo, presentamos os resultados obtidos para as regresións polinómicas de grao 3, 7, 10 e 17.

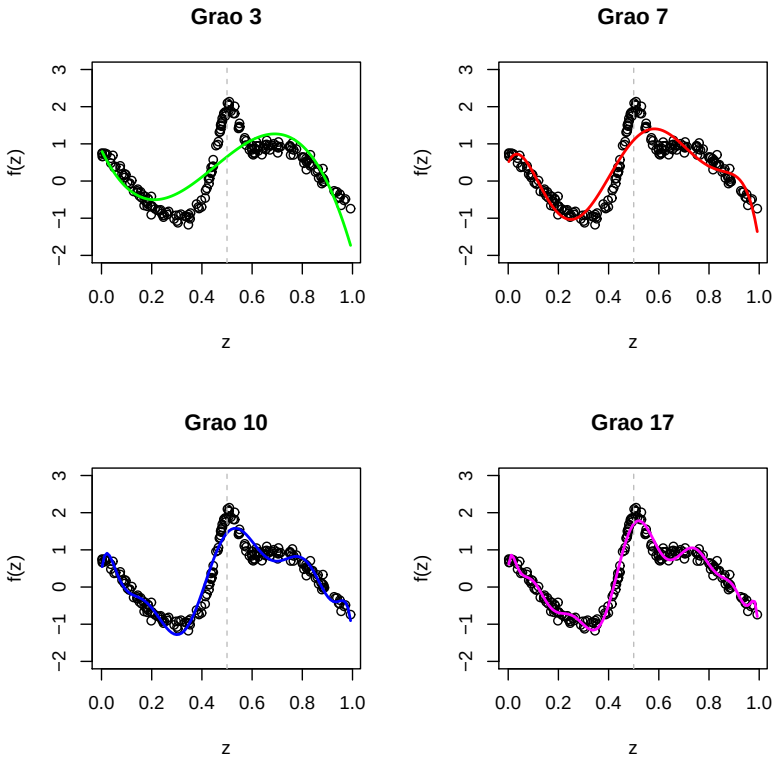


Figura 1.2: Modelos de regresión polinómica para o conxunto de datos simulados anteriormente, $y = f(x) + \epsilon$ onde $f(z) = \sin 8z - 4 + 2\exp -256(z - 0.5)^2 + \epsilon$, onde $\epsilon \sim N(0, 0.3^2)$.

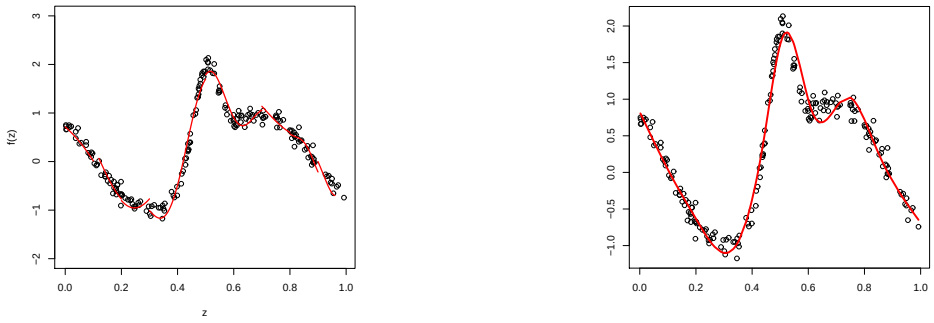
Á hora de realizar unha regresión polinómica, debemos ter en conta que o parámetro fundamental a elixir é o grao do polinomio, l. En moitos casos, esta escolla é esencial no resultado final da regresión. A modo de exemplo, na Figura 1.2, presentamos catro regresións polinómicas baseadas no conxunto simulado ao principio do Capítulo. Nas dúas primeiras gráficas, empregamos polinomios de grao 3 e 7, mentres que nas outras dúas presentamos os resultados obtidos considerando polinomios de graos superiores ($l = 10$ e $l = 17$, respectivamente).

Non obstante, tal e como observamos na Figura 1.2, en ocasións, a regresión polinómica non nos permite captar a verdadeira estrutura dos datos. Neste exemplo, os modelos considerados non son capaces de localizar exactamente o máximo local que se sitúa en $z = 0.5$.

Se ben é certo, a medida que aumentamos o grao do polinomio considerado, esta estimación mellora. Considerando un polinomio de grao 17 ($l = 17$), case se logra modelar este máximo local pero a estimación no resto de puntos é algo abrupta e irregular.

É preciso comentar que estas gráficas dan unha impresión diferente da estrutura dos datos. De feito, se comparamos os resultados obtidos para $l = 3$ e $l = 17$ é sorprendente que se traten de representacións diferentes dos mesmos datos. En xeral, polinomios de graos elevados dan lugar a estimacións relativamente abruptas, e con "picos". Por outra banda, a aparencia da curva estimada cun l pequeno é máis "suave" á vista. É por este motivo polo que se coñece como un parámetro de suavización, xa que controla a "suavización" que se aplica aos datos. A dificultade será atopar un valor de l axeitado que respecte a estrutura dos datos, pero aínda así, ás veces isto non é suficiente.

Unha forma intuitiva de aumentar a flexibilidade da regresión polinómica é realizar unha regresión polinómica local. É dicir, dividir o intervalo total en intervalos máis pequenos e realizar, en cada subintervalo por separado, unha estimación polinómica. Deste xeito, no canto de estimar un polinomio global, considéranse polinomios locais e estímanse os respectivos coeficientes independentemente para cada un dos subintervalos considerados empregando o método de mínimos cadrados. Isto permitiranos controlar o grao de suavización necesario en cada subintervalo.



a) b)

Figura 1.3: Na Figura a) representamos a regresión polinómica local para os datos simulados en comparación coa estimación resultante de empregar splines polinómicos (Figura b).

Na Figura 1.3, ilustramos esta aproximación. Para este exemplo, dividimos o dominio de

definición de z, en dez subintervalos de lonxitude 0.1 e realizamos a estimación polinómica en cada un deles independentemente. Efectivamente, obtense unha estimación máis flexible, que reflexa a verdadeira estrutura dos datos.

Non obstante, neste exemplo, tamén se pon de manifesto un dos maiores inconvenientes deste tipo de regresión: a función estimada non é continua, pois os polinomios foron axustados de xeito independente e non coinciden os valores dos extremos de cada intervalo. Polo tanto sería útil, impoñer certas restricións de suavización sobre a función e os extremos de cada intervalo de forma que puideramos obter unha función similar á representada na parte dereita da Figura 1.3.

A idea principal consiste en definir polinomios locais nos intervalos do dominio de definición da covariable, e para garantir a suavización da curva estimada, imponse que a función resultante sexa $(l-1)-$veces continuamente diferenciable. Estas ideas dan lugar a seguinte definición de Splines Polinómicos:

Un Spline Polinómico de grao $l \geq 0$ e nodos $\{k_1, \ldots, k_m\}$ é unha función, $f : [k_1, k_m] \mapsto \mathbb{R}$ que verifica as seguintes condicións (Fahrmeir et al., 2013):

1. $f(z)$ é $(l-1)$-veces continuamente diferenciable. No caso no que $l = 1$, esíxese a continuidade de $f(z)$ pero non a diferenciabilidade. Cando $l = 0$, non se necesita ningunha condición de suavización para $f(z)$.

2. $f(z)$ é un polinomio de grao l en cada un dos intervalos definidos polos nodos, $[k_j, k_{j+1}), j = \{1, \ldots, m-1\}$.

Deste xeito, considéranse funcións polinómicas definidas a cachos, de tal forma que a partición do dominio da variable determínase a través dos diferentes nodos $k_1, \ldots, k_m$. Ademais, esíxese que a curva estimada sexa $(l-1)$-veces continuamente diferenciable para garantir a suavización.

A suavización global da curva depende fundamentalmente do grao l elixido, mentres que a aparencia da curva estimada varía en función dos nodos empregados. Cantos máis nodos empreguemos, máis polinomios estaremos estimando e polo tanto atopámonos de novo co problema de escoller correctamente estes dous parámetros.

Fixado un conxunto de nodos e un determinado grao, para poder empregar os splines

polinómicos necesitamos necesitamos dispor dunha **base de regresión** formada por un conxunto de splines polinómicos. Estas bases poden calcularse de diferentes xeitos, entre os usuarios de P-splines, existen dous grandes grupos: os que empregan bases de polinomio truncadas e, por outro lado, os que se fundamentan en bases de B-splines. Ademais destes mecanismos, existen outras alternativas das que non falaremos no presente traballo (Durbán, 2008).

Polinomios truncados

Consideremos de novo que dispoñemos de n pares; $(z_i, y_i), i = 1, \ldots, n$. Para simplificar supoñemos que z está no intervalo $[0, 1]$. Tomamos m nodos equidistantes en ese intervalo $t_j = \frac{(j-1)}{m}, j = 2, \ldots, m+1$. Unha base de polinomios truncados de grao l vén dada por:

$$1, z, z^2, \ldots, z^l, \left\{(z - t_1)_+\right\}^l, \ldots, \left\{(z - t_m)_+^l\right\}^l$$

onde $z_+ = \text{máx}(0, x)$. Reciben o nome de polinomios truncados porque a partir dun certo punto valen 0.

A función $\{(z - t)_+\}^l$ ten $l - 1$ derivadas continuas, deste xeito canto maior é l, máis suaves serán as función da base.

Na Figura 1.4, representamos bases de polinomios truncados de graos 0, 1, 2, e 3 respectivamente tomando 10 nodos equidistantes para a covariable $z \in [0, 1]$.

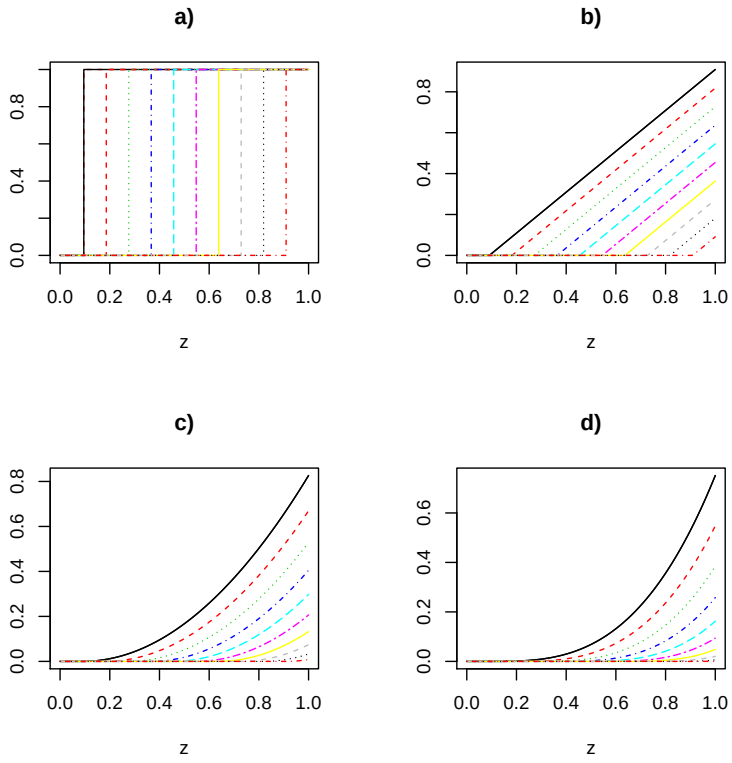


Figura 1.4: Bases de polinomios truncados de grao 0 (Figura a), 1 (Figura b), 2 (Figura c) e grao 3 (Figura d).

A continuación, consideremos o seguinte modelo de regresión:

$$y_i = \gamma_1 + \gamma_2 z_i + \cdots + \gamma_{l+1} z_i^l + \gamma_{l+2}(z_i - k_2)_+^l + \cdots + \gamma_{l+m-1}(z_i - k_{m-1})_+^l + \epsilon_i,$$

onde

$$(z - k_j)_+^l = \begin{cases} (z - k_j)^l & \text{se} \quad z \geq k_j. \\ 0 & \text{noutro caso.} \end{cases}$$

A primeira parte do modelo correspóndese cun polinomio global de grao l, tal e como presentabamos ao principio do Capítulo en (1.1) (a única diferencia reside na notación do

intercepto, que aquí chamamos γ_1 no canto de γ_0, por razóns que explicaremos máis adiante). Pero a diferencia dos modelos de regresión presentados, o coeficiente do polinomio máis alto vai cambiando en cada nodo $k_2, \ldots, k_{m-1}$.

Desta forma, ao mesmo tempo que realizamos unha estimación polinómica local en cada un dos intervalos que definen os nodos, garantimos as condicións globais de suavización.

Na Figura 1.5, ilustramos este concepto para un spline polinómico de grao 1, ($l = 1$). Na gráfica (a) representamos a función que define o modelo, é dicir, a función polinómica global de grao $l = 1$ (liña punteada) xunto cos polinomios truncados (liñas sólidas). Na gráfica (b) escalamos estas funcións cos coeficientes de regresión estimados segundo os datos considerados. (Para definir as funcións base empregamos 10 nodos equidistantes no intervalo unidade).

A liña horizontal en $y \approx 0.8$ correspóndese coa constante global γ_l. No primeiro intervalo $[0, 0.1)$, a función decrece dende este nivel global, representado polo parámetro γ_2. Dende o nodo, $k_2 = 0.1$ en adiante, o parámetro γ_3 impón a pendente. Neste exemplo, γ_3 é negativo e polo tanto fai que a función decreza de forma máis pronunciada. O coeficiente positivo γ_4 diminúe a pendente negativa dende $k_3 = 0.2$ en adiante. Non obstante, seguimos tendo unha tendencia negativa. Cando consideramos as tendencias adicionais da función, os coeficientes γ_j indican o cambio de pendente que ocorren no correspondente nodo k_{j-1}.

Grazas a introdución de cada termo, $(z-k_j)_+^l$, podemos garantir a suavización no cambio de pendente e desta forma mantéñense as propiedades dos splines polinómicos. Finalmente, cando engadimos todas as funcións escaladas, obtemos o axuste de $f(z)$ representado na Figura 1.5c.

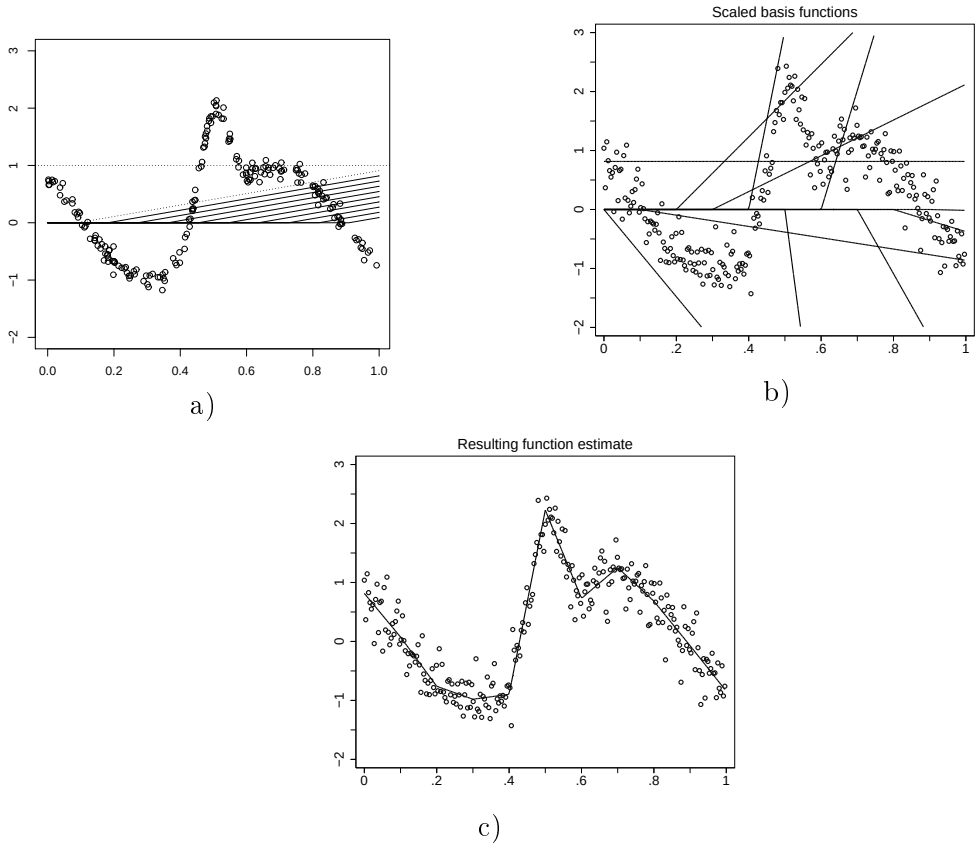


Figura 1.5: Axuste spline polinómico empregando polinomios lineais truncados. (a) Funcións base, (b) Funcións base escaladas. (c) Suma de funcións base escaladas.

De forma máis formal, pode demostrarse que cada spline polinómico de grao l con nodos $k_1 < \cdots < k_m$ pode ser determinado unicamente como combinación linear de $d = m + l - 1$ funcións (Fahrmeir et al., 2013):

$$B_1(z) = 1, B_2(z) = z, \ldots, B_{l+1}(z) = z^l B_{l+2}(z) = (z - k_2)_+^l, \ldots, B_d(z) = (z - k_{m-1})_+^l$$

Desta forma o problema de regresión non paramétrica queda representado da seguinte forma:

$$y_i = f(z_i) + \epsilon_i = \sum_{i=1}^{d} \gamma_j B_j(z_i) + \epsilon_i.$$

As funcións $B_1, \ldots, B_d$, forman unha base á que chamaremos base TP, *do inglés, truncated power series basis*, xa que nos permiten representar os splines polinómicos.

Modelar $f(z)$ como spline polinómico, permítenos interpretar o modelo de regresión non paramétrico como linear, aínda que, iso si, posiblemente cun gran número de parámetros.

Se denotamos por y ao vector de observacións, ϵ os erros, e Z á matriz de deseño,

$$Z = \begin{pmatrix} B_1(z_1) & \ldots & B_1(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \ldots & B_d(z_n) \end{pmatrix} = \begin{pmatrix} 1 & z_1 & \ldots & z_1^l & (z_1 - k_2)_+^l & \ldots & (z_1 - k_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & z_n & \ldots & z_n^l & (z_n - k_2)_+^l & \ldots & (z_n - k_{m-1})_+^l \end{pmatrix}.$$

obtemos a ecuación:

$$y = Z\gamma + \epsilon.$$

Sendo $\gamma = (\gamma_1, \ldots, \gamma_d)'$ o vector de coeficientes. Ao tratarse dun modelo linear con coeficientes de regresión, γ, poderemos empregar o método de mínimos cadrados para estimalos:

$$\widehat{\gamma} = \left(Z'Z\right)^{-1} Z'y.$$

Non obstante, a diferencia do que acontece cos modelos lineais, interpretar os coeficientes individualmente non é informativo, senón que o interese reside en analizar a forma da curva estimada, calculada en base ás estimacións dos coeficientes. É dicir:

$$\widehat{f}(z) = z'\widehat{\gamma},$$

sendo $z = (B_1(z), \ldots, B_d(z))'$ dependendo do valor da covariable z elixida. Finalmente, coa axuda dun gráfico de dispersión dos datos podemos comprobar a calidade do modelo axustado.

Influencia do número de nodos elixidos

Como podemos ver nas Figuras 1.7 e 1.6, nos modelos de regresión spline é esencial o grao do spline considerado e a posición e o número de nodos elixidos. En xeral, por defecto adóitanse empregar splines cúbicos (de grao 3) xa que nos permiten obter unha función "suave" e dúas veces continuamente diferenciable. Non obstante, construír unha regra que nos permita calcular o número de nodos óptimos que debemos empregar é moito máis difícil.

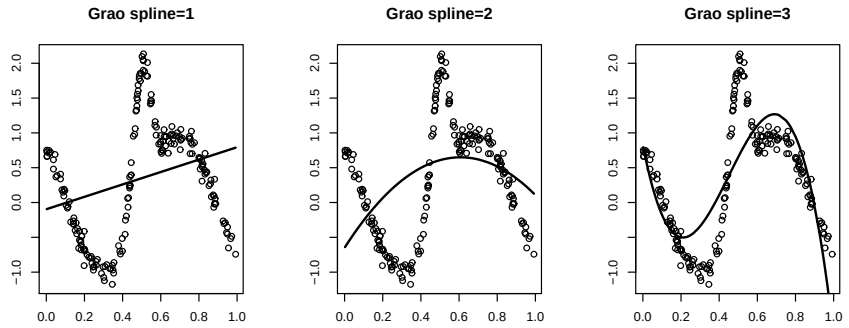


Figura 1.6: Estimacións non paramétricas para os datos simulados ao inicio do capítulo, baseadas en splines polinómicos de distintos graos. En todas as estimacións se tomaron os mesmos nodos.

Na Figura 1.7 empregamos splines cúbicos para amosar as diferencias existentes nas funcións estimadas segundo o número de nodos empregado, tal e como podemos observar, cantos máis (menos) nodos empreguemos máis (menos) flexible será a función estimada. Neste exemplo, vemos que cando se empregan poucos nodos, resulta unha función moi suave, que neste caso non capta a verdadeira estrutura dos datos. A medida que aumentamos o número de nodos, esta estimación mellora, neste exemplo parece que con 20 nodos é bastante exacta. Non obstante, se seguimos aumentando o número de nodos considerados, obtemos unha estimación irregular, con *picos*, moi difícil de interpretar. A dificultade reside precisamente en atopar o número óptimo de nodos necesarios para obter unha estimación que non sexa nin demasiado "suave" nin "irregular."

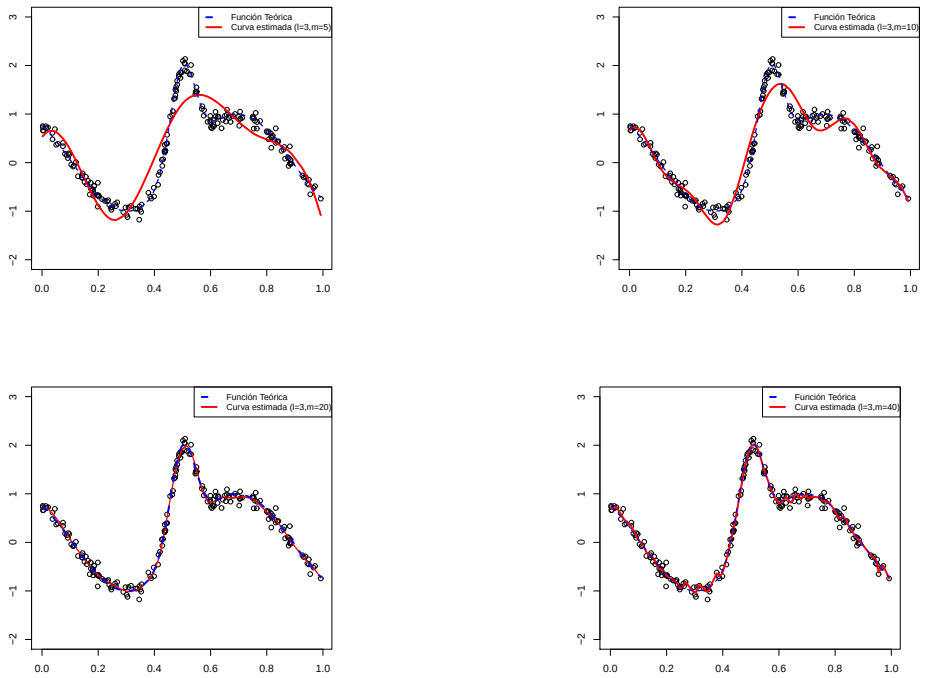


Figura 1.7: Importancia do número de nodos considerados no axuste cúbico spline. A función estimada represéntase en vermello e en azul mostramos a verdadeira función.

No exemplo anterior, é doado comprobar a fiabilidade de cada modelo pois coñecemos a función teórica, pero en xeral, na práctica non a coñeceremos polo que a dificultade reside en obter un mecanismo ou unha regra que nos permita decidir o número de nodos que empregaremos en cada caso. Pero ademais disto, tamén debemos decidir a posición destes nodos. Habitualmente, empréganse nodos equidistantes, ou mesmo, se calculan en función do diagramas de dispersión dos datos ou dos cuantiles. (Para maior detalle véxase pp. 426 de Fahrmeir et al., 2013). Non obstante, ningunha destas estratexias resolve o problema do número de nodos. Para solventalo, existen dúas alternativas, introducir unha "penalización" que fai que esta escolla sexa menos importante ou adaptar algún criterio de selección de modelos para determinar o número de nodos óptimos. (Véxase por exemplo, pp: 491-512 de Fahrmeir et al., 2013). Neste Traballo Fin de Máster centrarémonos nas aproximacións con penalizacións. Pero antes, presentaremos unha representación alternativa dos splines polinómicos que pode ser útil para construír os métodos baseados en penalizacións que

presentaremos na seguinte sección.

B-Splines

A parte das bases TP presentadas anteriormente, existen outras bases de splines polinómicos que posúen mellores propiedades numéricas. As referencias básicas para B-splines son De Boor (1977) e Diercks (1993).

Tal e como comentamos, as bases TP constrúense a partir de polinomios truncados e polo tanto poden orixinar inestabilidades numéricas cando empregamos covariables con valores grandes. Ademais, as funcións das bases TP son case colineais, especialemente cando dous nodos están máis próximos que o resto. Por este motivo, empregaremos bases B-spline (*basic spline*) como alternativa aos splines polinómicos. Ademais, os B-splines, non "padecen" dos efectos fronteira tan comúns noutros métodos de suavización como algúns suavizadores tipo núcleo, nos que ao estender a curva axustada fóra do dominio dos datos, esta tende a cero (Durbán, 2008).

Antes de dar unha definición puramente matemática dos B-splines, motivaremos a súa construción de maneira intuitiva. Tal e como comentabamos ao principio do capítulo, para construír as bases TP empréganse polinomios definidos a cachos impoñendo certas condicións de suavización sobre a función $f(z)$. Do mesmo xeito, un B-spline tamén está formado por cachos de polinomios que se unen "suavemente" nos nodos para garantir as condicións de suavización necesarias para realizar unha estimación coherente. Máis especificamente, unha función de bases B-spline está formada por $(l+1)$ polinomios de grao l que se unen $(l-1)$ veces continua e diferenciablemente. (Fahrmeir et al., 2013).

Un exemplo moi simple dun B-spline de grao 1 aparece na parte superior esquerda da Figura 1.8. Na parte dereita aparecen máis B-splines, cada un dos cales está baseado en tres nodos. Na parte inferior representamos un B-spline de grao 3. Está formado por 4 anacos de polinomios unidos entre si. Podemos observar que todas as funcións da base teñen a mesma forma pero están desprazadas horizontalmente (Durbán, 2008).

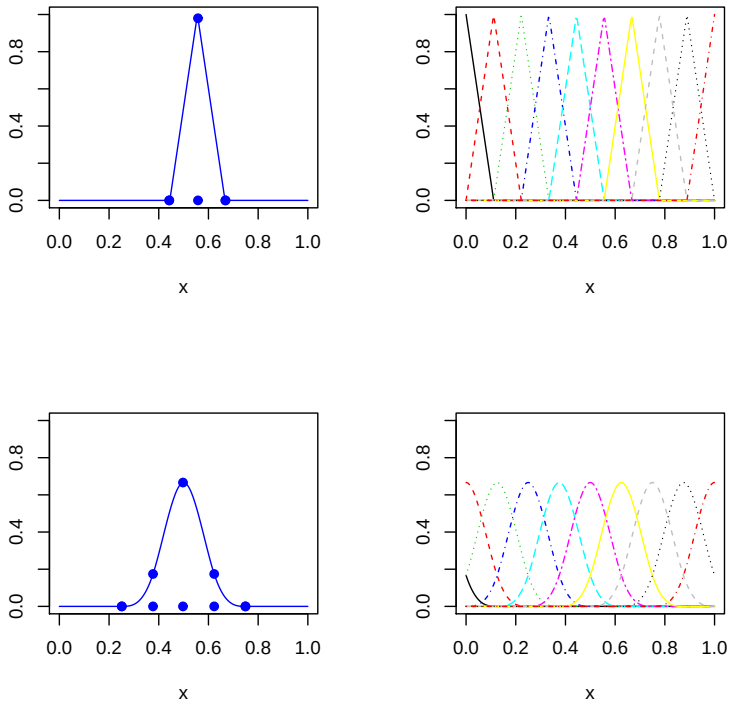


Figura 1.8: Bases de B-spline de orde 1 e 3.

En xeral, un B-spline de grao l ten as seguintes características (Durbán, 2008):

- Consiste en $l+1$ anacos de polinomios de orde l que se unen en l nodos internos.
- As derivadas ata $l-1$ son continuas nos puntos de unión.
- É positivo no dominio expandido por $l+2$ nodos e 0 no resto.
- Excepto nos extremos, solápase con $2l$ anacos de polinomios dos seus veciños.
- Para cada valor de x, $l+1$ B-splines son non nulos.

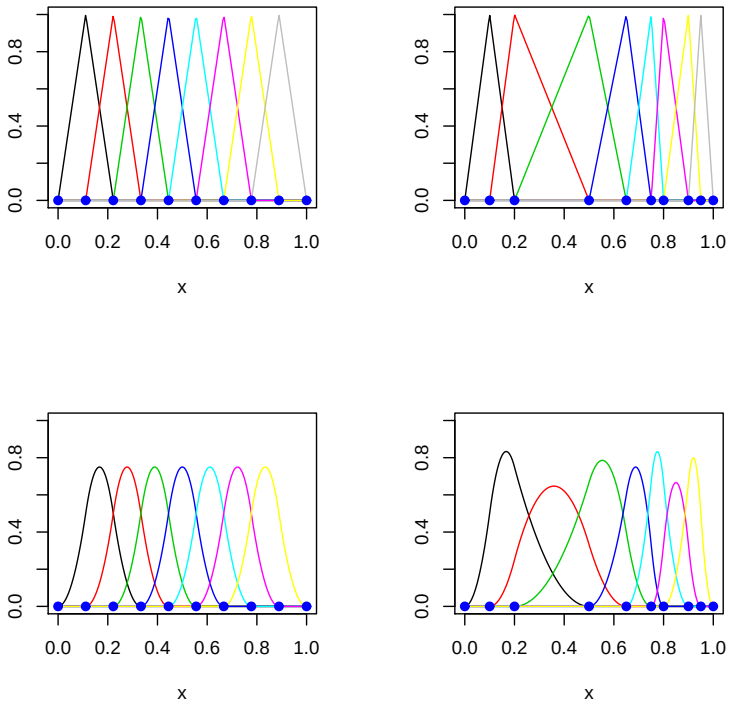


Figura 1.9: B-splines de grao $l = 1$ (Parte superior) e $l = 2$ (Panel inferior), fixando nodos equidistantes (1ª Columna) e aleatorios (2ª Columna)

Na Figura 1.9, mostramos bases B-splines empregando nodos equidistantes e outros distribuídos aleatoriamente. Ao igual que no caso anterior, empregando a base completa, a función $f(z)$ pode representarse de novo a través da combinación de $d = m + l - 1$ funcións bases, é dicir:

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z).$$

A principal vantaxe que se introduce co uso dos B-splines é que a súa definición é local, a diferencia do que acontece cos polinomios truncados das bases TP as bases B-spline tan só

son positivas no dominio expandido por $l+2$ nodos. Ademais, as funcións están limitadas e polos tanto non presentan os problemas numéricos das bases TP.

Nas páxinas 429-431 de Fahrmeir et al. (2013) pode consultarse a expresión matemática dos B-splines (pp. 429-431).

Na Figura 1.10 ilustramos a estimación dun axuste B-spline para o exemplo simulado dende o comezo do capítulo. En primeiro lugar, fixados os nodos, calcúlase as base B-spline (neste caso de grao 3) (Figura 1.10a)). A estimación de $\widehat{\gamma}$ mediante mínimos cadrados permítenos realizar o escalado que se representa na Figura 1.10b). Finalmente, sumando as funcións base escaladas, obtense a estimación final (Figura 1.10c)). (Fahrmeir et al., 2013).

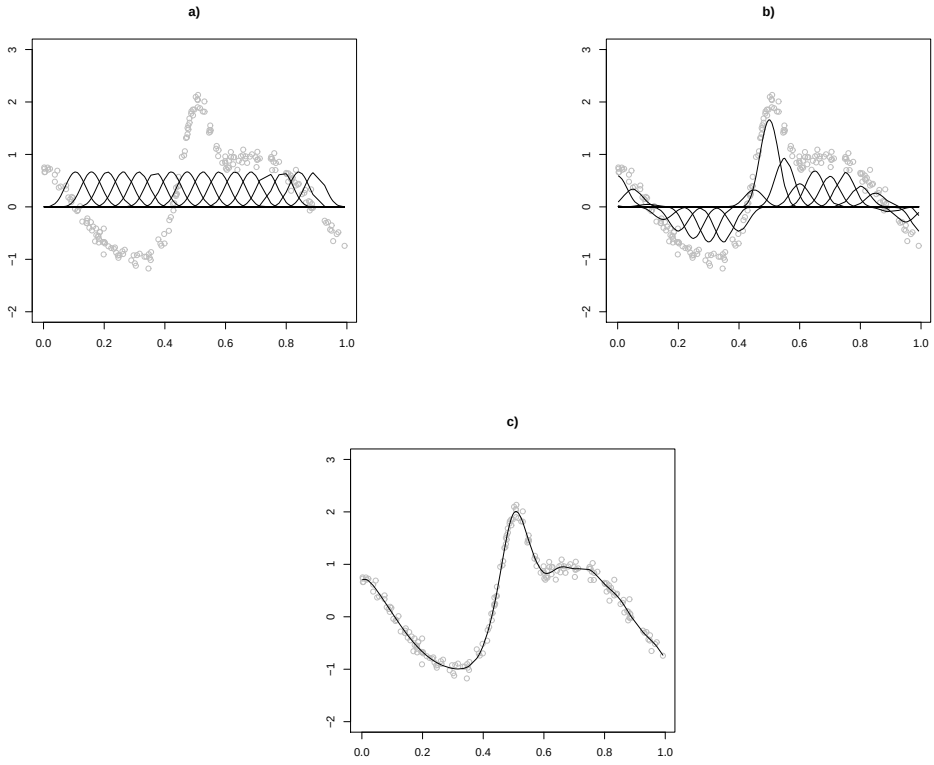


Figura 1.10: Representación dun axuste non paramétrico con B-splines cúbicos. Na primeira fila representamos, en primeiro lugar as bases B-spline, e a continuación as bases B-spline escaladas. Finalmente presentamos a suma de funcións base B-spline escaladas.

1.2. Splines Penalizados (P-splines)

Tal e como comentamos na Sección anterior os modelos de regresión baseados en splines polinómicos dependen en gran medida do número de nodos empregados. Unha forma de solucionar este problema é mediante a introdución de penalizacións.

Para realizar un axuste empregando splines penalizados (P-splines) debemos seguir os seguintes pasos (Durbán, 2008):

- Emprégase un spline polinómico con "bastantes nodos" (normalmente entre 20-40), para estimar $f(z)$ de forma que poida ser aproximada con suficiente flexibilidade para representar a complexidade da devandita función.
- Introdúcese un termo adicional de penalización que evita o sobreaxuste e minimiza o criterio PLS (mínimos cadrados penalizados, do inglés: *penalized least squares)* en vez do criterio usual de mínimos cadrados.

P-splines baseados en bases TP

Comezaremos considerando P-splines baseados nunha base TP, é dicir:

$$f(z) = \gamma_1 + \gamma_2 z + \cdots + \gamma_{l+1} z^l + \gamma_{l+2}(z - k_2)_+^l + \cdots + \gamma_{l+m-1}(z - k_{m-1})_+^l$$

Estas bases, como xa comentamos, constan de dúas partes, unha primeira formada por $l+1$ funcións base, que describen un polinomio global en z, mentres que as series truncadas (truncated powers) simbolizan as desviacións ao polinomio considerado. Desta forma, para regularizar a estimación pódese introducir unha penalización dos coeficientes das funcións base. Unha forma de definir esta penalización, é, por exemplo, empregar o cadrado dos coeficientes:

$$\sum_{l+2}^{d} \gamma_j^2,$$

de maneira que se penan os coeficientes asociados ás series truncadas demasiado grandes.

No canto de empregar a suma de cadrados residual usual, [2] minimizaremos a suma de cadrados penalizada.

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2.$$

O obxectivo da penalización é modelar aquelas funcións irregulares, con demasiado "ruído", evitando así o sobreaxuste dos datos.

O parámetro de suavización introducido, $\lambda \geq 0$, controla a influencia da penalización. O papel deste parámetro de suavización é o mesmo que en calquera outro método de suavización: controlar a suavización da curva. O obxectivo dos P-splines é suavizar os coeficientes que están moi separados entre si, polo tanto canto maior é λ, máis se aproximan os coeficientes a cero de forma que se $\lambda \mapsto \infty$, estímase $f(z)$ como un polinomio de grao l. Polo contra, cando $\lambda \mapsto 0$, o efecto da penalización desaparece e estaremos empregando o método de mínimos cadrados ordinario e polo tanto aproximámonos a un axuste linear. Variando o valor de λ, podemos conseguir un compendio entre ambos extremos.

A principal vantaxe que se introduce coa penalización, é que desta forma a suavización non depende da posición e o número de nodos empregados, senón dun único valor real, ao que denotamos como un parámetro de suavización (λ). De xeito que se empregamos un número suficiente grande de nodos, a posición destes é indiferente, por este motivo, na práctica, acostúmase empregar nodos equidistantes ou baseados en cuantiles, por simplicidade. Na Figura 1.11, podemos ver que considerando penalizacións a estimación da función non é flexible cando empregamos poucos nodos, pero se consideramos un número suficientemente grande, elixindo o parámetro de suavización axeitado non existen diferenzas. Pero evidentemente, resulta esencial elixir adecuadamente o parámetro de suavización para cada base de datos, en pp. 479-791 de Fahrmeir et al. (2013), poden consultarse os principais métodos de escolla. Entre os que se encontran os criterios clásicos de selección de modelos como validación cruzada, validación cruzada xeneralizada, ou o AIC (Akaike's information criterion) ou mesmo estimacións baiesianas baseadas en modelos mixtos.

[2] Suma de cadrados residual usual:

$$LS = \sum_{i=1}^{n} (y_i - f(z_i))^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2.$$

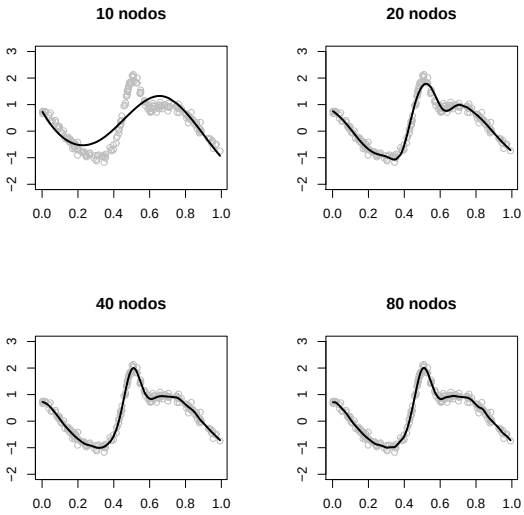


Figura 1.11: Influencia do número de nodos na estimación de P-splines.

P-splines baseados en B-Splines

Outra posibilidade consiste en representar $f(z)$ empregando B-splines, no canto de bases TP.

Neste caso considérase unha penalización baseada na segunda derivada da función, pois permite representar a variabilidade:

$$\lambda \int \left(f''(z)\right)^2 dz.$$

A integral da segunda derivada da curva axustada ao cadrado é unha penalización bastante común, (O'Sullivan, 1986). Non obstante, non hai nada de particular na segunda derivada, poden empregarse derivadas de calquera orde. A novidade que introducen os P-splines é que a penalización é discreta, é dicir, pénanse os coeficientes directamente, en lugar da curva, reducíndose deste xeito a dimensionalidade do problema. (Durbán, 2008.)

En Fahrmeir et al. (2013) podemos atopar os detalles desta estimación (pp. 433-441). Este tipo de suavización foi proposta en Eilers e Marx (1996) e converteuse nunha das

técnicas de suavización máis populares. Consiste en empregar unha penalización baseada nas diferencias de orde d entre os coeficientes adxacentes da base B-spline. Este tipo de penalización é máis flexible xa que é independente do grao do polinomio empregado para construír os B-splines.

Na Figura 1.12 representamos diferentes estimacións splines penalizadas considerando distintos valores do parámetro de suavización, de forma que para valores grandes do parámetro de suavización ($\lambda \mapsto \infty$), orixinan unha estimación linear no caso de considerar diferencias de orde dous. En xeral, cando $\lambda \mapsto \infty$, o axuste aproxima un polinomio de grao $r-1$, sendo r a orde das diferenzas (debemos ter en conta que como mínimo se debe tomar $l \geq r$). [3]

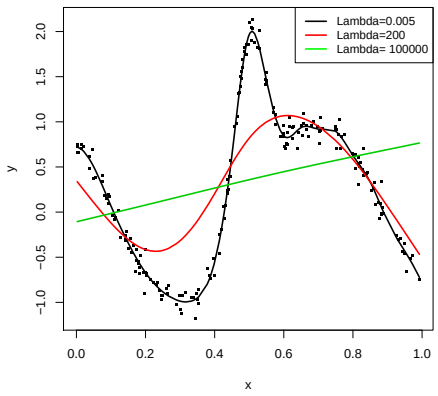


Figura 1.12: Influencia do parámetro de suavización na estimación de P-splines considerando penalizacións de segunda orde.

Na Figura 1.13, mostramos o axuste dunha curva mediante B-splines considerando nun caso penalizacións e noutro non. Xunto coa función estimada, representamos as funcións que forman as bases (as columnas da matriz B) multiplicadas polos coeficientes (representados por un círculo). Na parte esquerda da Figura, vemos a curva obtida é pouco suave. En cambio, cando empregamos penalizacións, imponse aos coeficientes se pase dun ao outro de

[3] No caso de considerar diferenzas de orde un, pode probarse que nese caso a primeira derivada sería 0, e polo tanto a correspondente función sería unha constante se e só se todos os coeficientes da primeira derivada son cero (Fahrmeir et al., 2013).

forma suave, e polo tanto a curva obtida é máis suave, (Durbán, 2008).

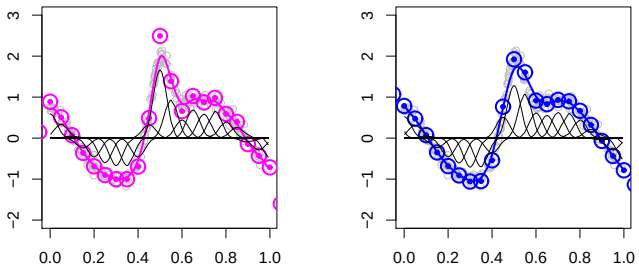


Figura 1.13: Curva estimada con 20 nodos, sen penalizar os coeficientes (esquerda) e penalizando os coeficientes (dereita). Neste gráfico podemos ver que o obxectivo da penalización, é precisamente, impoñer que se pase dun coeficiente ao outro de forma suave.

Entre as propiedades dos P-splines con bases B-splines hai que destacar que non padecen de efecto fronteira (como ocorre cos suavizadores tipo núcleo). O efecto fronteira é o que fai que ao estender fora do dominio da covariable a curva decreza rapidamente cara 0. Ademais, no caso no que as curvas sexan polinomios un P-spline consegue estimalas exactamente. Finalmente, conservan os momentos, é dicir, a media e a varianza dos valores axustados é a mesma que a dos datos, independentemente do parámetro de suavización elixido, ao contrario do que acontece cos estimadores tipo núcleo, que tenden a aumentar a varianza canto maior é a suavización. (Wand e Jones, 1995).

1.3. Suavización bivariante

Nesta sección presentaremos varias aproximacións de suavizado bivariantes (modelos con dúas variables de regresión continuas).

As técnicas de suavización presentadas anteriormente permítennos analizar o efecto dunha variable explicativa continua mediante métodos de regresión non paramétrica. Pero, en moitas ocasións, necesitamos dispor de modelos con dúas ou máis variables. Empregando os

modelos anteriores tan só se podería analizar o efecto non paramétrico de cada variable por separado, pero desta forma, asumiríase que non existen interaccións entre as variables.

Outras das limitacións dos modelos anteriores é que non nos permiten incorporar efectos espaciais. A miúdo, en moitos campos de aplicación estadística (epidemioloxía, economía, ciencias sociais, ...), disponse de datos que conteñen información xeográfica ou espacial (Hennerfeind el al., 2005), por exemplo, a dirección de residencia (país, provincia, código postal ...) dos pacientes ou individuos que forman o estudo. En moitos destes estudos, como é lóxico, analizar o impacto destes efectos xeográficos é de vital importancia pois permítennos captar posibles heteroxeneidades espaciais que non se reflicten co resto de covariables. Nesta sección, estudaremos a posibilidade de incluír os efectos espaciais nos modelos de regresión.

Bases de produtos tensoriais

En superficies bivariantes, podemos estender os conceptos introducidos anteriormente mediante *bases de produtos tensoriais.*

Nesta sección consideraremos o caso no que a variable resposta y se describe en termos de superficies de dúas dimensións $f(z_1, z_2)$, onde z_1 e z_2 poden ser covariables continuas, así como coordenadas no caso de modelos espaciais. En primeiro lugar, construiremos bases univariantes para z_1 e z_2: $B_j^{(1)}(z_1), j = 1, \ldots, d_1$, e $B_r^{(2)}(z_2), r = 1, \ldots, d_2$. A base produto tensorial consiste no produto destas bases:

$$B_{jr}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_r^{(2)}(z_2), j = 1, \ldots, d_1, r = 1, \ldots, d_2.$$

Desta forma, podemos representar $f(z_1, z_2)$ da seguinte forma:

$$f(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{r=1}^{d_2} \gamma_{jr} B_{jr}(z_1, z_2).$$

Para ilustrar a construción das bases produto tensoriais, na Figura 1.14 representamos splines base produto tensorial construídos a partir de bases lineais univariantes TP,

$$B_1^{(1)}(z_1) = 1, B_2^{(1)}(z_1) = z_1, B_3^{(1)}(z_1) = (z_1 - k_1)_+$$

e

$$B_1^{(2)}(z_2) = 1, B_2^{(2)}(z_2) = z_2, B_3^{(2)}(z_2) = (z_2 - k_2)_+.$$

A función constante que aparece na primeira representación do panel esquerdo da Figura 1.14, resulta do produto de $B_1^{(1)}$ e $B_1^{(2)}$. Tanto as gráficas representadas na primeira fila coma as representadas na primeira columna obtéñense multiplicando a función base constante na dirección de z_1, coas funcións base na dirección de z_2 e viceversa. O resto de representacións corresponden aos produtos do resto de funcións base univariantes.

Como no caso dos TP-splines, poderíamos pensar en empregar penalizacións, non obstante, neste caso, atopámonos con moitos máis problemas numéricos que no caso univariante, polo que empregaremos produtos tensoriais de bases B-splines posto que son máis estables. Na Figura 1.15, representamos os produtos individuais de B-splines de graos $l = 0, 1, 2$ e 3. Observamos como a medida que aumenta o grao do spline, a suavización tamén é maior. En Dierckx (1993), pódese consultar unha descrición máis detallada das propiedades dos splines produto tensoriais.

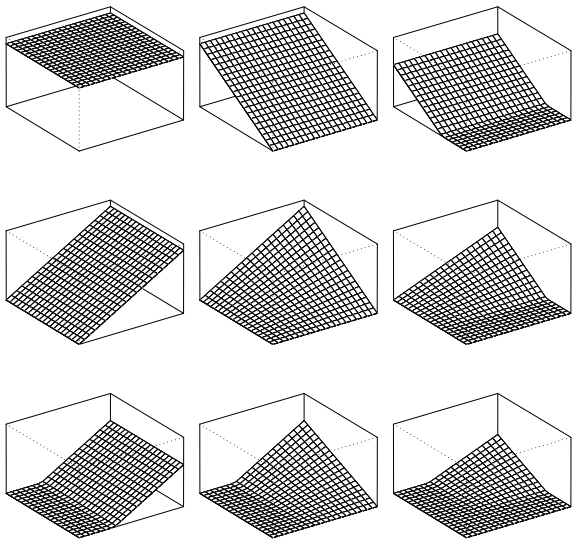


Figura 1.14: Bases produto tensoriais construídas a partir de bases lineais univariantes TP.

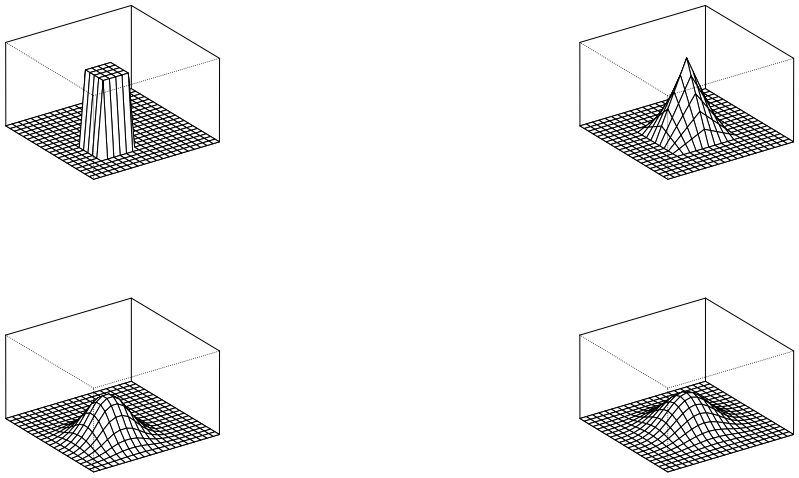


Figura 1.15: Funcións base produto tensoriais calculadas a partir de B-splines univariantes de graos $l = 0, 1, 2$ e 3, respectivamente de esquerda a dereita. Observamos como a medida que aumenta o grao do spline, a suavización tamén é maior, así por exemplo, os spline produtos tensoriais de grao 0, non son continuos, mentres que os de grao 1, son continuos pero non diferenciables.

Se observamos as representacións gráficas das liñas de contorno dos splines produto tensoriais (Figura 1.16), vemos que non son círculos, especialmente os de grao 1, e polo tanto os produtos tensoriais splines non son radiais. (As bases radiais introducirémolas na seguinte Sección.)

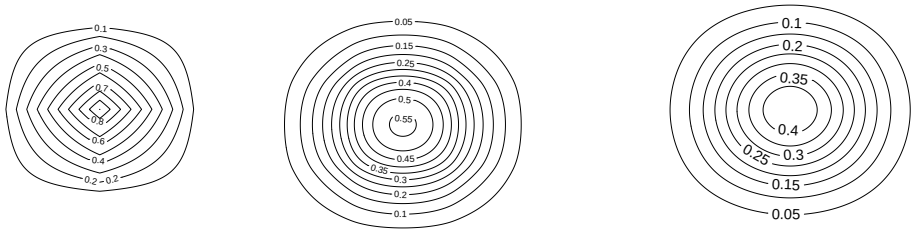


Figura 1.16: Representación gráfica das liñas de contorno de B-splines produto tensoriais de grao $l = 1, 2$ e 3, respectivamente de esquerda a dereita.

A pesar de que estas bases produto tensoriais semellan ser moito máis complexas que as univariantes, pódense representar en forma dos modelos lineais. Con este obxectivo, definimos a matriz Z cuxa fila $i-$ésima vén dada por:

$$z_i' = \left(B_{11}\left(z_{i1}, z_{i2}\right), \ldots, B_{d_1 1}\left(z_{i1}, z_{i2}\right), \ldots, B_{1 d_2}\left(z_{i1}, z_{i2}\right), \ldots, B_{d_1 d_2}\left(z_{i1}, z_{i2}\right)\right)$$

e o vector dos coeficientes de regresión como:

$$\gamma = \left(\gamma_{11}, \gamma_{d_1 1}, \ldots, \gamma_{1 d_2}, \ldots, \gamma_{d_1 d_2}\right)'.$$

Obtendo deste xeito a ecuación estándar de regresión, $y = Z\gamma + \epsilon$.

En principio, poderíamos estimar os coeficientes do mesmo xeito que nos modelos lineais, pero o número de parámetros a estimar é moito maior que no caso univariante polo que, recorreremos a outras técnicas baseadas en B-splines.

Ao igual que cos polinomios splines univariantes, necesitamos determinar o número óptimo e a posición dos nodos para construír splines produto tensoriais. Pero ademais, a miúdo, atoparémonos co problema de que nalgunhas rexións non existen observacións e nestes casos é imposible estimar os coeficientes das funcións base asociadas a esas rexións. Este mesmo

problema tamén pode ocorrer cando empregamos bases B-splines univariantes (por exemplo cando a covariable non toma valores en grandes intervalos) pero non adoita ser tan común coma no caso univariante. Non obstante, estes problemas poden solucionarse engadindo penalizacións.

Penalizacións 2D

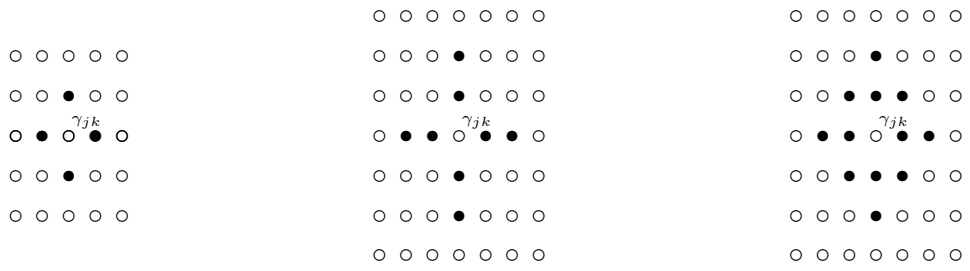


Figura 1.17: *Veciñanzas espaciais.* Representación dos veciños de γ_{jk}.

No caso dos B-splines univariantes, as penalizacións constrúense en función das diferencias dos cadrados dos coeficientes das funcións base dos veciños. Para trasladar este concepto ao caso bidimensional, debemos definir as veciñanzas espaciais. Na Figura 1.17, representamos unha posible definición de veciñanza considerando ben 4, 8 ou 12 veciños, respectivamente. A continuación, presentaremos varias penalizacións baseadas nestes veciños.

Comezamos co caso máis simple no que consideramos 4 veciños. Unha maneira razoable de construír estas penalizacións pode ser empregar os cadrados das diferencias entre γ_{jk} e eses catro veciños. Consideremos, pois: D_1 e D_2 as matrices das diferencias univariantes de primeira orde nas direccións z_1 e z_2, respectivamente.

Por filas, as diferenzas de primeira orde obtéñense aplicando as matrices das diferenzas expandidas $I_{d_2} \otimes D_1$, ao vector γ, onde I_d é a matriz identidade d-dimensional e $\otimes$ denota ao produto de Kronecker (Ver Definición A.10 do Apéndice A.1 en Fahrmeir et al., 2013.) Aplicando esta matriz de diferenzas ao vector formado polos coeficientes de regresión obtemos:

$$\gamma' \left(I_{d_2} \otimes D_1\right)' \left(I_{d_2} \otimes D_1\right) \gamma = \sum_{r=1}^{d_2} \sum_{j=2}^{d1} \left(\gamma_{jr} - \gamma_{j-1,r}\right)^2, \tag{1.2}$$

é dicir, a suma por filas de todos os cadrados das diferencias. Analogamente, por columnas:

$$\gamma' \left(D_2 \otimes I_{d_1}\right)' \left(D_2 \otimes I_{d_1}\right) \gamma = \sum_{j=1}^{d_1} \sum_{r=2}^{d2} \left(\gamma_{jr} - \gamma_{j,r-1}\right)^2. \tag{1.3}$$

Sumando (1.2) e (1.3), obtemos as penalizacións totais:

$$\lambda \gamma' K \gamma = \lambda \gamma' \left[\left(I_{d_2} \otimes D_1\right)' \left(I_{d_2} \otimes D_1\right) + \left(D_2 \otimes I_{d_1}\right)' \left(D_2 \otimes I_{d_1}\right)\right] \gamma. \tag{1.4}$$

Tendo en conta as propiedades do produto de Kronecker (Ver Teorema A.4 do Apéndice A.1 de Fahrmeir et al., 2013) pode demostrarse que (1.4) é equivalente a:

$$\lambda \gamma' K \gamma = \lambda \gamma' \left[I_{d_2} \otimes K_1 + K_2 \otimes I_{d_1}\right] \gamma,$$

sendo $K_1 = D_1' D_1$ e $K_2 = D_2' D_2$ matrices univariantes de penalizacións.

Obtemos deste xeito unha penalización cadrática, $\lambda \gamma' K \gamma$. Para estimar λ, pódense empregar calquera dos métodos discutidos anteriormente (Ver Fahrmeir et al., 2013 para máis detalle, pp: 508-510 e pp: 479-486.)

A partir da penalización anterior podemos derivar unha aproximación baiesiana. Neste caso, ao igual que no caso univariante, podemos interpretar K como unha matriz da distribución completa do vector γ cando consideramos paseos aleatorios bidimensionais de segunda orde. Máis concretamente, obtemos a distribución a priori de γ como:

$$p\left(\gamma \mid \tau^2\right) \propto \left(\frac{1}{\tau^2}\right)^{\frac{rk(K)}{2}} \exp\left(-\frac{1}{2\tau^2} \gamma' K \gamma\right) \tag{1.5}$$

Baseándonos nesta densidade a priori, podemos calcular a distribución condicional de γ_{jr} coñecidos o resto de coeficientes:

$$\gamma_{jr} \mid \cdot \sim N\left(\frac{1}{4}\left(\gamma_{j-1,r}+\gamma_{j+1,r}+\gamma_{j,r-1}+\gamma_{j,r+1}\right), \frac{\tau^2}{4}\right)$$

Deste xeito, o vector γ verifica de forma espacial a propiedade de Markov, xa que a distribución condicional de γ_{jr} só depende dos catro veciños máis próximos. Podemos observar como o valor esperado da distribución condicional é, precisamente, a media dos valores dos 4 veciños máis próximos.

Este mecanismo podémolo aplicar a matrices de diferencias de ordes superiores, obtendo penalizacións da seguinte forma:

$$\lambda\gamma' K\gamma = \lambda\gamma'\left[I_{d_2}\otimes K_1^{(k_1)} + K_2^{(k_2)}\otimes I_{d_1}\right]\gamma,$$

con matrices de penalizacións univariantes $K_1^{(k_1)}$ e $K_2^{(k_2)}$ de orde k_1 e k_2. Por exemplo, con $k_1 = k_2 = 2$, obtemos a penalización en función dos cadrados das diferencias de segunda orde baseada nos 8 veciños máis próximos a longo dos eixos coordenados (Ver Figura 1.17).

En conclusión, acabamos de ver como efectivamente, o vector da función de avaliacións pode representarse como un gran modelo lineal $Z\gamma$ con penalizacións cadráticas $\lambda\gamma' K\gamma$ ou de forma equivalente, a través da súa distribución a priori (1.5). Deste xeito, podemos empregar os métodos de selección de parámetros ventá mencionados na Sección anterior. Non obstante, debemos ter en conta que o número de parámetros nos modelos bidimensionais é moito maior que no caso univariante, polo que será aínda máis crucial empregar métodos numéricos eficaces.

1.3.1. Funcións base radiais e *Thin Plate Splines*

Un método alternativo para construír funcións base bivariantes é empregar bases radiais. Unha función base radial, defínese como unha función da distancia Euclídea usual entre un nodo $k = (k_1, k_2)$ e un punto observado, $z = (z_1, z_2)$, é dicir,

$$B_k(z) = B(\|z-k\|) = B(r),$$

elixindo unha función escalar axeitada, B, e a distancia euclídea,

$$r = \|z-k\| = \sqrt{\left((z_1-k_1)^2 + (z_2-k_2)^2\right)}.$$

O termo de bases radiais, provén do feito de que, por construción, as liñas de contorno son circulares. Todas as funcións base teñen esta forma e ademais cada unha delas asóciase a un único nodo, a diferencia do que acontecía coas funcións B-splines produto tensoriais.

Habitualmente, os nodos dunha base radial son un subconxunto dos puntos observados, é dicir, $\{k_1, \ldots, k_d\} \subset \{z_1, \ldots, z_n\}$, de forma que a distribución das funcións base radiais adáptase a estrutura dos datos. Pola contra, un dos inconvenientes que presentan as funcións produto tensoriais é que adoitan tomar valores en intervalos onde non existe ningunha observación.

Minimizando o seguinte criterio (sobre a clase de todas as funcións $f(z)$ dúas veces continuamente diferenciables), obtemos as funcións base radiais máis coñecidas:

$$\sum_{i=1}^{n} (y_i - f(z_i))^2 + \lambda \int\int \left[\left(\frac{\partial^2}{\partial^2 z_1} + 2\frac{\partial^2}{\partial z_1 \partial z_2} + \frac{\partial^2}{\partial^2 z_2}\right) f(z_1, z_2)\right]^2 dz_1 dz_2 \to \min_f \qquad (1.6)$$

Neste caso,

$$\lambda \int\int \left[\left(\frac{\partial^2}{\partial^2 z_1} + 2\frac{\partial^2}{\partial z_1 \partial z_2} + \frac{\partial^2}{\partial^2 z_2}\right) f(z_1, z_2)\right]^2 dz_1 dz_2 \qquad (1.7)$$

representa ao análogo bivariante ao cadrado integrable da segunda derivada.

Resultado da minimización de (1.6), obtemos os *thin plate splines*, unha xeneralización dos splines cúbicos que se comportan de forma lineal fora do dominio das observacións (é dicir, verifican as condicións naturais de fronteira). Un *thin plate splines*, pode ser representado como:

$$f(z_1, z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \sum_{j=1}^{n} \gamma_j B_j(z_1, z_2),$$

onde

$$B_j(z_1, z_2) = B(\|z - z_j\|) = \|z - z_j\|^2 \log(\|z - z_j\|)$$

ademais debemos considerar certas restricións sobre os coeficientes que presentaremos a continuación.

Os *thin plate spline*, baséanse en efectos lineais nas direccións de z_1 e de z_2 e das funcións base radiais,

$$B(r) = r^2 \log(r)$$

centradas nos n valores das covariables. En Green e Silverman (1993), preséntase de forma máis minuciosa o concepto de thin plate spline, ademais de demostrarse a súa optimalidade.

En notación matricial, a representación dos thin plate spline inducen o seguinte modelo:

$$y = XB + Z\gamma + \epsilon,$$

onde X denota á matriz de deseño que contén as constantes e os efectos lineais de z_1 e z_2; por outro lado $\beta = (\beta_0, \beta_1, \beta_2)'$ é un vector formado polos coeficientes de regresión,

$$Z[i, j] = B_j(z_{i1}, z_{i2})$$

contén as funcións base radiais avaliadas nos valores observados das covariables, e γ é o vector dos coeficientes da base. Se contabilizamos o número de coeficientes da regresión, vemos que hai $n + 3$, polo que existen máis parámetros que ecuacións. Para solventar isto, é suficiente con impoñer a restrición $X'\gamma = 0$, desta forma aseguramos que a parte linear do modelo é ortogonal á parte que representa as funcións base radiais. Ademais, tamén se pode ver que a penalización integral (1.7), equivale a $\gamma' Z \gamma$, onde a matriz de penalización

coincide coa matriz de deseño. En consecuencia, o problema de minimización presentado en (1.6) pode reescribirse da seguinte forma:

$$(y - X\beta - Z\gamma)'(y - XB - Z\gamma) + \lambda\gamma' Z\gamma \to \min_{\beta,\gamma}$$

suxeito a que $X'\gamma = 0$.

Do mesmo xeito que acontecía no contexto dos splines de suavizado, o número de coeficientes asociados cos thin plate splines son demasiados, na práctica habería que resolver $(n+3)(n+3)$ sistemas de ecuacións. Polo tanto é necesario obter outras aproximacións aos thin plate splines, de baixo rango, intentando manterse cerca da solución óptima na medida do posible. Neste sentido, as regras comúns para seleccionar como nodos un conxunto de observacións, a miúdo presentan solucións razoábeis. Non obstante, Wood propuxo unha aproximación óptima baseada nas descomposición espectral da matriz de deseño, Z (Wood, 2003.) Na que en primeiro lugar, se considera a seguinte descomposición espectral:

$$Z = \Gamma\Omega\Gamma',$$

onde Γ é unha matriz ortogonal de autovectores e Ω contén os correspondentes autovalores (non negativos) en orden descendente. Pode demostrarse que:

$$Z_d = \Gamma_d\Omega_d\Gamma_d',$$

onde Γ_d e Ω_d son submatrices de Γ e Ω asociadas cos d autovalores máis grandes, é o mellor aproximación de rango d no sentido da norma espectral $\|z - z_d\|$. [4] A idea é substituír Z por Z_d, de maneira que se traslada o problema orixinal nun subespazo d-dimensional, onde d debe ser elixido de tal forma que a aproximación do erro sexa pequeno.

Antes de introducir os modelos de regresión aditiva estruturada, que son precisamente os que dan o nome ao presente Traballo Fin de Máster é necesario introducir os conceptos básicos das técnicas de suavizado espaciais.

[4] Dada unha matriz semidefinida, A, a norma espectral ($\|A\|$), corresponde á raíz cadrada do autovalor máis grande de A.

1.4. Técnicas de suavización q-dimensionais

En principio, as ideas presentadas na Sección 1.3, poden estenderse para modelar os efectos de superficies de dimensións maiores:

$$y = f(z_1, \ldots, z_q) + \epsilon$$

Por exemplo, para construír os produtos tensoriais $q-$dimensionais, bastaría con considerar todas as posibles interaccións dos splines univalentes para cada unha das covariables, $z_1, \ldots, z_q$. Ao igual que no caso bidimensional, construiríanse as matrices de penalización. Por exemplo, construír as bases radiais tamén é bastante simple. Do mesmo xeito, estendendo o concepto de veciñanza, poden empregarse as metodoloxía das cadeas aleatorias de Markov.

Se ben é certo, que independentemente da metodoloxía elixida, poden ocorrer certos problemas cando intentamos estimar funcións de grandes dimensión non parametricamente. En xeral, o número de parámetros empregados é moi grande e os algoritmos de resolución son moi custosos.

1.5. Técnicas de suavización espacial

A información espacial recompílase en diversos campos científicos e actualmente está cada vez máis presente en fontes de datos científicas e públicas. É máis, son moitas as ramas nas que a análise e suavizado espacial dos datos é precisamente a principal fonte de interese, por exemplo en estudos medioambientais, mapas de enfermidades ou en imaxes médicas.

Entre os diferentes métodos de suavizado espacial, existe unha diferenza básica en función de se se considera a información espacial de forma continua ou discreta (Fahrmeir e Kneib, 2011).

Denotemos por s unha localización espacial. No caso continuo, a información espacial interprétase como coordenadas. Por exemplo, no caso bivariante, a localización represéntase como: $s = (s_1, s_2)$ con coordenadas s_1 e s_2, e s toma valores nun subconxunto, $D \subset \mathbb{R}^2$. Cando dispoñemos deste tipo de variables de localización continua empréganse técnicas de

suavización coñecidas tradicionalmente como *Kriging*. Este nome débese ao enxeñeiro D.G. Krige, que inventou este tipo de modelos para determinar os graos de mineral nas minas de ouro.

No caso discreto, toma valores $s \in \{1, \ldots, d\}$. Neste caso, s_i representa por exemplo un píxel dunha imaxe, unha rexión ou área específica (os concellos de Galicia, países, comunidades autónomas) dentro dun conxunto de d rexións. Con esta información podemos construír, por exemplo, mapas de enfermidades que son moi empregados nalgunhas aplicacións epidemiolóxicas ou xeográficas.

No caso das variables espaciais discretas, non dispoñemos das coordenadas específicas de cada observación senón que a cada unha delas se lle asigna un determinado clúster. Por exemplo, cando dispoñemos de datos de censos, por razóns de confidencialidade non se rexistran as coordenadas exactas da localización de cada vivenda, senón que se clasifican por comunidades, provincias, municipios ou o que máis interese ao investigador.

Nas páxinas 315-325 de Fahrmeir e Kneib (2011) pódense consultar os principais tipos de datos espaciais, non obstante neste traballo basearémonos na información espacial discreta, en concreto, nas Cadeas Aleatorias de Markov (do inglés, *Markov Random Fields, MRF*), como técnica de suavización espacial para as variables de localización discretas. Dependo do tipo de datos espaciais que dispoñamos empregaremos unhas ou outras técnicas para realizar o noso estudo.

1.5.1. Cadeas aleatorias de Markov

O principal obxectivo deste Traballo Fin de máster será, precisamente, presentar un modelo non paramétrico que nos permita non só explicar efectos non lineais de covariables continuas ou discretas senón ir máis alá e explicar comportamentos espaciais. A principal diferenza coa que nos atopamos ao modelar o efecto dun vector de covariables e un vector de coordenadas espaciais, é a escala de cada un dos elementos: mentras que as coordenadas espaciais se expresan normalmente nas mesmas unidades (habitualmente en metros ou quilómetros), en xeral as unidades do resto de covariables son moi diferentes (Fahrmeir e Kneib, 2011). Isto dificulta a construcción dunha medida de distancia tendo en conta as covariables espaciais. Nestes casos, é habitual recorrer aos produtos tensoriais para modelar os efectos das covariables.

Unha das dificultades da regresión espacial é atopar unha medida que nos permita medir a proximidade entre dúas rexións ou localizacións. Mentres que no caso continuo, podemos resolver este problema coa distancia euclídea, por exemplo, no caso discreto necesitamos definir novos conceptos como o de *veciñanza*. En realidade, existen diferentes formas de construír unha veciñanza (Fahrmeir e Kneib, 2011):

- Se a covariable espacial, s, denota o número de observacións rexistradas nunha rexión particular, adoitase considerar como veciños aquelas rexións que comparten fronteira. (Figura 1.18).
- Outra posibilidade sería diferenciar veciños de primeira, segunda ou $n-$ésima orde, segundo a súa proximidade pero por simplicidade, no que segue non consideraremos esta definición de veciñanza.

Empregaremos a notación $\sim$ para indicar que s e r son dúas rexións veciñas, é dicir, comparten fronteira. Ademais, a cada rexión asignarémoslle un coeficiente de regresión, $f_{\text{geo}}(s) = \gamma_s, s = 1, \ldots, d$. Deste xeito introdúcense moitos coeficientes, polo que necesitamos unha estrutura apropiada para modelar os efectos espaciais de suavizado de forma que se reduza o número de parámetros efectivos. Baseándonos na idea intuitiva de que os coeficientes de aquelas rexións que se atopan *cerca* non deberían variar moito, construiremos unha penalización a partir dos cadrados das diferenzas entre os parámetros das rexións veciñas, considerando o seguinte criterio PLS, construído a partir de cadrados de diferenzas de todas as combinacións posibles de rexións veciñas:

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - f_{\text{geo}}(s_i)\right)^2 + \lambda \left[\sum_{s=2} \sum_{r \in N(s), r<s} (\gamma_r - \gamma_s)^2 \right], \tag{1.8}$$

onde $N(s)$ define o conxunto de veciños da rexión s.

Desta forma, penalízanse aquelas rexións cuxos valores disten dos rexistrados nos seus veciños.

Figura 1.18: Veciñanzas de primeira orde dunha cuadrícula regular (esquerda) e no panel dereito considerando un conxunto de datos rexional irregular.

Para incluír esta aproximación dentro dos métodos xerais de regresión, en primeiro lugar defínese a matriz de deseño Z como segue:

$$Z[i,s] = \begin{cases} 1 & \text{se } y_i \text{ é unha observación rexistrada na rexión} s \\ 0 & \text{noutro caso.} \end{cases} \tag{1.9}$$

Isto permítenos expresar, $f_{geo} = (f_{geo}(s_1), \ldots, f_{geo}(s_n))$, como un modelo linear $Z\gamma$. Do mesmo xeito, a penalización pode escribirse de forma compacta como unha forma cadrática, $\lambda\gamma' K\gamma$ con:

$$K[s,r] = \begin{cases} -1 & \text{se } s \neq r, s \sim r \\ 0 & \text{se } s \neq r, s \not\sim r \\ |N(s)| & \text{se } s = r \end{cases}$$

Minimizando o criterio PLS, obtemos de novo a estimación penalizada, $\widehat{\gamma} = (Z'Z + \lambda K)^{-1} Z'y$.

Segundo a definición anterior, a matriz de penalización K ten estrutura dunha matriz de veciñanzas, pois cada un das entradas $K[s,r]$ tan só é distinta de cero cando s e r son veciños. Trátase, pois dunha matriz dispersa, é dicir, a maioría dos seus elementos son cero, polo que se poden usar métodos numéricos eficientes para procesar esta matriz. Ver por exemplo, George e Liu (1981). En Rue e Held (2005) tamén se discuten estes algoritmos

dende o punto de vista da estadística, non obstante debemos de ter en conta que é o número de rexións consideradas é un factor moi importante.

Formulación do Modelo Baiesiano

A continuación, introduciremos de novo as cadeas aleatorias de Markov nun contexto máis xeral. Na maioría da literatura existente, os métodos de penalización para as localizacións espaciais discretas desenvólvese dende unha perspectiva baiesiana; Fahrmeir et al., (2013).

Comezaremos definindo as Cadeas aleatorias de Markov (MRF, do inglés Markov Random Fields).

Sexa $D = \{1, \ldots, s, \ldots, d\}, s = 1, \ldots, d$ o conxunto de todas as rexións. Diremos que $\gamma = \{\gamma_s, s \in D\}$ verifica a condición de MRF se a distribución condicional de γ_s coñecidos o resto de efectos $\gamma_r, r \neq s$, depende só dos seus veciños. A densidade (condicional) correspondente pode escribirse como:

$$p\left(\gamma_s \mid \gamma_r, r \neq s\right) = p\left(\gamma_s \mid \gamma_r, r \in N(s)\right).$$

Se consideramos, agora, o seguinte modelo:

$$y_i = f_{geo}(s_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2).$$

O noso obxectivo será asignar unha MRF a distribución de $f_{geo}(s_i) = \gamma_s$.

Neste caso particular, podemos supoñer que (Fahrmeir et al., 2013):

$$\gamma_s \mid \gamma_r, r \in N(s) \sim N\left(\frac{1}{|N(s)|}\sum_{r:r\sim s}\gamma_r, \frac{\tau^2}{|N(s)|}\right), \tag{1.10}$$

onde $|N(s)|$ denota ao número de veciños da rexión s. Segundo a expresión anterior, a distribución condicional de γ_s asúmese a priori como normal cuxa esperanza vén dada pola media dos valores veciños mentres que a varianza, τ^2, é inversamente proporcional ao

número de veciños e controla a desviación de γ_s á esperanza. Deste xeito podemos obter a distribución conxunta de todos os efectos espaciais:

$$p(\gamma \mid \tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{d-1}{2}} \exp\left(-\frac{1}{2\tau^2}\gamma' K \gamma\right). \tag{1.11}$$

A densidade anterior está definida de forma proporcional, unicamente. A matriz de precisión K, correspóndese exactamente coa matriz de penalización introducida previamente no criterio PLS (1.8), en (1.11), presentamos o equivalente baiesiano.

Por outro lado, na ecuación (1.10) non se ten en conta que cada veciño pode influír dunha forma diferente. Unha forma de estender a ecuación (1.10) é asociar a cada veciño un peso determinado:

$$\gamma_s \mid \gamma_r \in N(s) \sim N\left(\sum_{r:r\sim s} \frac{w_{sr}}{w_{s+}}\gamma_r, \frac{\tau^2}{w_{s+}}\right)$$

tomando pesos simétricos de forma que $w_{sr} = w_{rs}$ e $w_{s+} = \sum_{r:r\sim s} w_{sr}$. Desta forma, a esperanza condicional de γ_s vén dada pola media dos coeficientes das observacións veciñas.

En Fahrmeir et al., (2013) preséntanse varias opcións para definir pesos, w_{sr}:

- Empregar o mesmo peso para todos os veciños, é dicir, $w_{sr} = 1$. Esta consideración, daría lugar á definición de MRF, orixinalmente presentada.
- Considerar pesos inversamente proporcionais a distancia dos centroides, por exemplo, $w_{sr} \propto \exp(-d(s,r))$, onde $d(s,r)$ define a distancia Euclídea entre os centroides das rexións s e r.
- Empregar pesos proporcionais á lonxitude das fronteiras comúns de s e r.

Se consideramos estes pesos, a matriz de precisión ou de penalizacións, K, tamén cambia:

$$K[s,r] = \begin{cases} -w(s,r) & \text{se } s \neq r, s \sim r \\ 0 & \text{se } s \neq r, s \not\sim r \\ w_{s+} & \text{se } s = r \end{cases}$$

A continuación proporcionamos un pequeno resumo sobre os métodos bivariantes discutidos anteriormente.

1.6. Resumo sobre os diferentes enfoques das aproximacións de penalización

Tal e como vimos, os métodos bivariantes de suavizado, e as MRF, poden tratarse tendo en conta os criterios xerais de penalización.

Pero, en todos os casos, os modelo lineais obtidos, $y = Z\gamma+\epsilon$, posúen unha gran cantidade de coeficientes de regresión, xa sexan os relativos as funcións base ou aos efectos espaciais. Para *regularizar* a estimación de γ, introdúcense as penalizacións cadráticas,$\lambda\gamma' K\gamma$.

En función do método empregado, as penalizacións estarán baseadas nas diferencias dos coeficientes empregando operadores de derivadas, en diferenzas dos coeficientes, correlacións, ou estruturas de veciños. Finalmente selecciónase un criterio adecuado para escoller un parámetro de suavizado.

Introducidos os conceptos básicos de suavización en regresión, no Capítulo 2 introduciremos a teoría fundamental dos modelos de regresión aditiva estruturada.

No Capítulo 3, empregaremos os modelos STAR con resposta de Poisson para investigar as tendencias espaciais na taxa de abstinencia ao alcohol (AWS) en Galicia e a súa relación con varios factores socioeconómicos. Finalmente, no Capítulo 4 presentaremos un exemplo de análise de supervivencia empregando os modelos de regresión estruturada. Para elo, analizaremos as desigualdades xeográficas na supervivencia dos pacientes ingresados cun diagnóstico de síndrome coronario agudo na área sanitaria de Santiago de Compostela.

Capítulo 2

Modelos de Regresión Aditiva Estruturada

Ao longo deste capítulo desenvolveremos a teoría fundamental sobre os modelos STAR (do inglés *Structured Additive Regression models*, Fahrmeir et al., 2013). En primeiro lugar, para dar comezo a este capítulo, farase unha introdución aos modelos STAR (Sección 2.1). Como xa comentamos no Capítulo 1, estes modelos permiten incluír efectos non lineais das covariables continuas, interaccións entre covariables, efectos aleatorios, datos clúster ou incluso efectos espaciais ou temporais, entre outros. Na Sección 2.2, explicaremos como modelar cada un dos efectos anteriores.

Para realizar a inferencia dos modelos STAR, partiremos da posibilidade de expresar os modelos STAR como modelos mixtos (Sección 2.3). Isto proporcionaranos as ferramentas necesarias para realizar a inferencia dende unha perspectiva empírica Baiesiana (Sección 2.4).

Ademais da flexibilidade á hora de modelar as covariables, os modelos STAR permítennos incoporporar unha ampla familia de variables resposta (familia exponencial, respostas categóricas, tempos de supervivencia, ou multiestado). Neste Traballo Fin de Máster, faremos fincapé na posibilidade que nos ofrecen de modelar tempos de supervivencia. Para elo na Sección 2.5 introduciremos os modelos de regresión estruturada de risco.

Xa para rematar, na Sección 2.6, presentaremos o software estadístico BayesX, que nos

permitirá estimar os modelos STAR.

2.1. Introdución

No capítulo anterior amosamos diferentes técnicas para modelar de forma flexible o efecto dunha covariable continua z sobre a variable resposta y. Ademais vimos como xeneralizar estes conceptos para dúas covariables continuas, z_1 e z_2, incluíndo ademais unha variable espacial de localización. Pero, en ocasións, dispoñemos de moitas covariables continuas $z_1, \ldots, z_q$, cuxos efectos non se poden modelar como unha forma funcional fixada inicialmente. Senón que estaremos interesados en modelar de forma flexible o efecto destas covariables en forma dunha función $f(z_1, \ldots, z_q)$. Non obstante, a estimación de funcións de grandes dimensións adoita ser moi custosa ademais de necesitar tamaños mostrais moi grandes. Por este motivo, suporemos que se verifica unha estrutura aditiva máis restritiva:

$$f(z_1, \ldots, z_q) = f_1(z) + \cdots + f_q(z_q).$$

Ademais, en moitos estudos necesitamos analizar efectos non lineais de interaccións, incluír efectos aleatorios ou mesmo empregar modelos xeoaditivos para poder captar posibles heteroxeneidades espaciais que non se reflicten mediante outras variables.

Nos últimos anos os modelos de regresión aditiva estruturada, coñecidos como modelos STAR (*Structured Additive Regression models*, Fahrmeir et al., 2013), están acadando gran interese en moitos campos de aplicación estadística, posto que nos permiten incorporar todos os efectos anteriores nun único modelo, de maneira que permiten xeneralizar aos modelos clásicos de regresión, os modelos lineais xeneralizados (GLM, Generalized Linear Models, McCullag e Nelder, 1989) e aos modelos aditivos xeneralizados (GAM, Generalized Additive Models, Hastie e Tibshirani, 1990).

En Thomas (2005), demóstrase a xeneralidade destes modelos partindo da base de que se pode demostrar que os modelos estadísticos clásicos presentes na literatura (Modelos GAM, GAMM, modelos xeoaditivos, modelos de interación tipo ANOVA, modelos de coeficiente variable), son casos especiais dos modelos STAR.

A pesar de que os modelos xeneralizados lineais son flexibles no sentido que nos permiten

empregar diferentes tipos de funcións de distribución como resposta, neles asúmese que a influencia das covariables é lineal e as observacións son independentes. Pero na práctica, non sempre é correcto asumir que todas as covariables continuas son lineais, senón que poden depender doutras formas non lineais descoñecidas. Ademais, poden existir correlacións espaciais e/ou tendencias temporais entre as observacións. Do mesmo xeito, a heteroxeneidade dos individuos ou as semellanzas entre grupos non se explican facilmente mediante as covariables continuas (Thomas, 2005).

Unha forma de solucionar as carencias dos modelos lineais xerais, conséguese substituíndo o predictor linear por outro paramétrico aditivo estruturado. Explicitamente, a fórmula xeral dos modelos STAR toma a forma (Fahrmeir et al.,2013):

$$\eta_i = f_1(v_{i1}) + \cdots + f_p(v_{iq}) + x_i'\beta \tag{2.1}$$

onde η é a variable resposta; i é un índice xenérico que denota a observación i-ésima e v denota as diferentes covariables xenéricas de distintos tipos e dimensións, e $f_i, i = 1, \ldots, p$ son funciones descoñecidas (non necesariamente suaves) que nos permiten modelar efectos non lineais das covariables continuas, tendencias temporais ou efectos espaciais, superficies bidimensionais, modelos de coeficientes variables, interceptos e pendentes aleatorias independentes e identicamente distribuídas ou mesmo efectos espaciais correlacionados. Nunha primeira visual pode resultar estraño que en (2.1) empreguemos unha mesma notación para todas as posibles funcións non lineais que explican as variables continuas. Non obstante, poder tratalos de maneira unificada é unha das principais vantaxes que nos ofrecen este tipo de modelos. Finalmente, a segunda parte da ecuación, $x'\beta$, denota aos efectos paramétricos das covariables estudadas.

Neste contexto, os efectos non lineais das covariables continuas así como posibles tendencias temporais, modelaranse mediante versións baiesianas de splines penalizados (P-splines; Fahrmeir, Kneib e Lang, 2004).

Ademais, poderemos engadir ao modelo, efectos espaciais estruturados, que se estimarán empregando campos aleatorios gaussianos de Markov (Rue e Held, 2005). De xeito adicional, tamén poderán incorporarse efectos espaciais non estruturados, que nos permitirán captar posibles tendencias espaciais de pequenas áreas locais ou heteroxeneidades específicas de determinados individuos. Suporemos que estes efectos non estruturados seguen unha distribución previa gaussiana. Deste xeito, tódolos parámetros e funcións descoñecidas se tratan

de forma similar, é dicir, asígnaselle unha distribución previa coa mesma estrutura xeral pero de diferentes formas e graos de suavizade en función do tipo de efecto que pretendamos modelar.

A inferencia dos modelos STAR pódese realizar mediante métodos puramente Baiesianos (Full Bayes, FB) ou, pola contra, empregando aproximacións empíricas (Empirical Bayes, EB).

Na inferencia puramente Baiesiana, a varianza ou calquera outro parámetro de suavizado considérase unha variable aleatoria que se estimará empregando extensións das técnicas MCMC (Fahrmeir et al., 2005).

En Fahrmeir, Kneib e Lang (2004); podemos atopar un estudo de simulación comparativo de ambas técnicas de inferencia. Pero neste traballo centrarémonos na inferencia EB, na cal, tanto a varianza como os parámetros de suavizado se consideran constantes descoñecidas e estímanse mediante aproximacións REML (Restricted Maximun Likelihood). Cada un dos parámetros de suavizado, así como os efectos das covariables e as funcións descoñecidas, obtéñense maximizando as densidades posteriores. Neste traballo presentaremos unha aproximación EB baseada nos modelos lineais mixtos xeneralizados (GLMM, Breslow e Clayton, 1993) e empregaremos algoritmos REML computacionalmente eficientes, que nos permitirán aplicar a metodoloxía dos modelos GLMM para realizar inferencia nos modelos STAR incluso con bases de datos moi grandes.

Modelos STAR. Definición

Unha das principais vantaxes que nos ofrecen os modelos STAR é que se poden estender para case todo tipo de respostas, en particular, respostas binarias, discretas e categóricas. Do mesmo xeito que nos modelos lineais xeneralizados, consideramos que as variables resposta y_i son (condicionalmente) independentes respecto do preditor η_i. Entón, o preditor estruturado aditivo -onde as variables, $v_1, \ldots, v_q$ son covariables uni ou multidimensionais construídas a partir das variables orixinais- (Fahrmeir et al., 2013),

$$\eta_i^{struct} = f(v_{i1}) + \cdots + f(v_{iq}) + x_i'\beta$$

relaciónase coa media (condicional), $\mathbb{E}(y_i) = \mu_i$:

$$\mathbb{E}(y_i) = \mu_i = h(\eta_i^{strct}),$$

elixindo correctamente a función h. Referirémonos ao modelo resultante como modelo STAR xeneralizado.

Este tipo de modelo xeneralizado, contén como casos particulares os principais modelos estudados nos cursos máis comúns de estadística (os modelos xeneralizados aditivos (modelos GAM), modelos xeoaditivos, modelos con coeficientes variables, modelos de regresión xeográfica baseados en pesos ou modelos ANOVA con interacións) como un caso especial. Por exemplo, se tomamos:

$$\eta_i^{struct} = \eta_i^{add} = f_1(z_{i1}) + \cdots + f_p(z_{iq}) + x_i'\beta$$

resulta un modelo xeneralizado aditivo (GAM). Incorporando un efecto xeográfico,

$$\eta_i^{struct} = \eta_i^{add} + f_{\text{spat}(s_i)}$$

obtemos un modelo xeneralizado xeoaditivo, e poderíamos continuar derivando o resto de modelos.

As densidades ou distribucións previas das funcións, f_j, dependerán do tipo específico de covariables consideradas, v_j, e das suposicións de suavidade sobre cada función f_j. Na seguinte sección introduciremos varias posibilidades de modelado.

2.2. Distribucións previas

Tal e como comentabamos ao comezo deste Capítulo, a fórmula xeral dos modelos STAR vén dada por (Fahrmeir et al., 2013):

$$\eta_i = f_1(v_{i1}) + \cdots + f_p(v_{iq}) + x_i'\beta. \tag{2.2}$$

Dende o punto de vista baiesiano, tanto as funcións descoñecidas $f_1, \ldots, f_p$, como os parámetros de efectos fixos, β, considéranse variables aleatorias, as cales substituiremos en cada caso por distribucións previas adecuadas.

Neste Traballo Fin de Máster, suporemos que os efectos paramétricos posúen a seguinte distribución previa, $p(\beta) \propto \text{const}$.

No caso das funcións $f_1, \ldots, f_p$, asumiremos diferentes densidades en función do tipo de efecto que pretendamos modelar. No que segue, expresaremos o vector das avaliacións da función descoñecida f_j, $f_j = (f_j(v_{1j}), \ldots, f_j(v_{nj}))'$, como produto matricial dunha matriz de deseño, V_j, e o vector de parámetros descoñecidos γ_j, é dicir:

$$f_j = V_j \gamma_j.$$

Deste xeito, podemos expresar o preditor xeral, (2.2), en forma matricial, tal e como segue:

$$\eta = V_1 \gamma_1 + \cdots + V_q \gamma_q + X\beta, \tag{2.3}$$

onde X, denota a matriz de deseño usual dos efectos fixos.

A densidade de cada función f_j defínese elixindo de forma apropiada, as matrices de deseño, V_j, e a distribución previa dos vectores γ_j de parámetros descoñecidos. En xeral, adóitase considerar que a distribución previa de γ_j vén dada por:

$$p(\gamma_j \mid \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\gamma_j' K_j \gamma_j\right). \tag{2.4}$$

onde K_j é unha matriz de penalización que fai tender os parámetros cara o cero ou penaliza saltos demasiados bruscos entre parámetros veciños.

O parámetro da varianza, τ_j^2, equivale a inversa do parámetro de suavizado -nas aproximacións frecuentistas-, e controla o equilibrio entre a flexibilidade e a suavidade. No caso da inferencia EB, considérase que τ_j^2 é unha función constante descoñecida determinada mediante unha estimación REML.

A matriz de deseño, V_j, e a matriz de penalización K_j, caracterizan o tipo de modelo STAR ante o cal nos atopamos. Na páxina 554 de Fahrmeir et al. (2013) podemos atopar unha relación dos diferentes tipos de matrices de penalización e de deseño en función do método empregado para estimar os coeficientes. A continuación describiremos diferentes densidades a priori en función do tipo de covariable considerada.

2.2.1. Modelado dos efectos das covariables continuas e escalas temporais

Existen diversas alternativas para modelar o efecto das covariables continuas ou as tendencias temporais. Neste Traballo Fin de Máster centrarémonos no uso de P-splines, xa presentados na Sección 1.2 do Capítulo 1 do presente traballo.

Suporemos que as funcións f_j descoñecidas e suaves, asociadas a cada covariable, x_j, poden estimarse mediante splines polinómicos de grao l definidos nun conxunto de nodos equiespaciados: $x_j^{\min} = k_0 < k_1 < \cdots < k_{d-1} < k_d = x_j^{\max}$ no dominio de x_j.

Cada función pode expresarse como combinación lineal de $M_j = d + l$ B-splines, B_m, é dicir:

$$f_j(x_j) = \sum_{m=1}^{M_j} \gamma_{jm} B_m(x_j).$$

Na expresión anterior, $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jM_j})'$ corresponde ao vector dos coeficientes de regresión (que non coñecemos). A matriz de deseño $V_j \in M_{n \times M_j}$, está formada polas funcións da base avaliadas nas observacións x_{ij}, é dicir, $V_j(i,m) = B_m(x_{ij})$. Tal e como vimos no Capítulo 1, a escolla do número de nodos é esencial: se empregamos poucos nodos, o spline resultante non captará de forma axeitada a variabilidade dos datos, pola contra con moitos nodos as curvas estimadas tenden a sobreaxustar os datos, dando lugar a funcións irregulares e abruptas. Para solventar isto, Eilers e Marx (1996), propuxeron empregar un número bastante grande de nodos equiespaciados (entre 20 e 40) para asegurar a flexibilidade e definiron ademais penalizacións baseadas en diferenzas de primeira e segunda orde dos coeficientes dos B-splines (xa introducidas no Capítulo anterior):

$$P(\lambda_j) = \frac{1}{2}\lambda_j \sum_{m=k+1}^{M_j} \left(\Delta^k \gamma_{jm}\right)^2, k = 1, 2. \tag{2.5}$$

onde λ_j é un parámetro de suavizado e Δ^k é un operador diferenza de orde k. As diferenzas de primeira orde penalizan os saltos bruscos entre entre parámetros sucesivos, $\gamma_{jm} - \gamma_{j,m-1}$, mentres que as diferenzas de segunda orde penalizan, $2\gamma_{j,m-1} - \gamma_{j,m-2}$. Dende a perspectiva Baiesiana, empregaremos paseos aleatorios (*random walks*) de primeira ou segunda orde como densidades previas dos coeficientes de regresión. Estes paseos aleatorios de primeira e segunda orde, defínense como:

$$\gamma_{jm} = \gamma_{j,m-1} + u_{jm} \qquad \gamma_{jm} = 2\gamma_{j,m-1} - \gamma_{j,m-2} + u_{jm}. \tag{2.6}$$

con erros gaussianos $u_{jm} \sim N(0, \tau_j^2)$ e densidades previas difusas para os valores iniciales, $p(\gamma_{j1}) \propto const$ ou $p(\gamma_{j1})$ e $p(\gamma_{j2})$, respectivamente. Desta forma, a distribución conxunta dos parámetros de regresión γ_j, poden ser codificados como produto das densidades condicionais definidas en (2.6) e poden expresarse na forma xeral (2.4).

Por outro lado a matriz de penalización vén dada por, $K_j = D'D$ onde D é unha matriz de diferenzas de primeira ou segunda orde.

Para modelar tendencias temporais, no canto de empregar paseos aleatorios (2.6) é máis útil considerar densidades previas baseadas en procesos autoregresivos (ver Fharmeir e Lang, 2001) que tamén se poden expresar na forma xeral introducida en (2.4).

2.2.2. Modelado dos efectos espaciais

Supoñamos que o índice $s \in \{1, \ldots, S\}$ representa unha rexión ou localización dentro dun conxunto de S rexións. Por simplicidade, asumiremos que cada unha das rexións foron nomeadas de forma consecutiva. Tal e como comentamos no Capítulo 1, unha forma moi común de introducir os efectos espaciais correlacionados é supoñer que as localizacións veciñas teñen máis similitudes que outras calquera. Habitualmente, asúmese que dúas rexións, s e s' son veciñas se posúen unha fronteira común.

Tal e como explicamos no Capítulo 2, a forma máis simple e tamén a máis empregada para definir as densidades previas da función de avaliacións $f_{spat(s)=\beta_s}$ é:

$$\gamma_s \mid \gamma_r, s \neq s', \tau_j^2 \sim N\left(\frac{1}{|N(s)|}\sum_{s'\in\delta_s}\gamma_s', \frac{\tau_j^2}{|N(s)|}\right), \tag{2.7}$$

onde $|N(s)|$ denota ao número de veciños da rexión s. E $s' \in \delta_s$, simboliza que a rexión s' é veciña de s. Segundo a expresión (2.7), a media condicional de γ_s é unha media non ponderada dos valores que toman as s funcións das rexións veciñas. Esta densidade é unha xeneralización dos camiños aleatorios de primeira orde de dúas dimensións denominada campos aleatorios de Markov (MRF, Markov Random Field). Non obstante, existen densidades máis xerais baseadas en medias ponderadas descritas por Besag York e Mollié (Fharmeir, Kneib e Lang, 2004).

Neste caso a matriz de deseño, $V \in M_{n\times S}$ é unha matriz de incidencias formada por ceros e uns. Tal e como comentabamos no Capítulo 1, o termo correspondente á fila i e a columna s desta matriz vale un se a observación i está na rexión s, e cero noutro caso. A matriz de penalización, $k \in M_{S\times S}$, ten forma dunha matriz adxacente (*adjacency matrix)*, presentada no Capítulo 1.

Na actualidade, existen diversas alternativas, as MRF, por exemplo, en xeoestadística é popular o uso de campos aleatorios gaussianos estacionarios (GRF, Gaussian random fields, Ruppert, Wand e Carrol (2003)), que poden ser vistos como suavizadores bidimensionais baseados en funcións base, como por exemplo as radiais (introducidas no Capítulo 1). No caso de traballar con datos discretos, os MRF adoitan ser adecuados. Pero se dispoñemos, por exemplo, das localizacións exactas, adoita ser máis natural empregar estimadores de superficies. Pero, non sempre é así, en ocasións os GRF poden ofrecer mellores resultados que os MRF no estudo de datos discretos e viceversa. Polo tanto, non está moi claro cal das dúas opcións proporciona o mellor axuste. Neste TFM, empregaremos o método de MRF, pois dispoñemos de datos espaciais discretos.

2.2.3. Indicadores de grupos e efectos espaciais non estruturados

En moitas situacións observamos problemas de heteroxeneidades dentro de grupos debidas a outras covariables que non se observan. Supoñamos que $c \in \{1, \ldots, C\}$, indica o grupo ao que pertence unha observación particular. Na práctica, para captar esta heteroxeneidade, adoitase introducir efectos gaussianos identicamente distribuídos, $f_c = \beta_c$, con:

$$\beta_c \sim N(0, \tau^2), c = 1, \ldots, C. \tag{2.8}$$

A matriz de deseño $V \in M_{n \times C}$ é de novo unha matriz de incidencias formada por ceros e uns. Neste caso a matriz de penalización é a matriz indentidade ($K = I$).

Debemos de ter en conta que estamos considerando efectos aleatorios específicos para cada un dos clúster. Dende unha perspectiva clásica, (2.8), define *efectos aleatorios* i.i.d. Non obstante, dende o punto de vista Baiesiano, todos os parámetros descoñecidos son considerados aleatorios e, neste caso, a notación de 'efectos aleatorios" pode ser confusa. Neste traballo consideraremos (2.8), como unha aproximación para modelar unha función non suave.

A densidade introducida en (2.8), pode empregarse para modelar de xeito máis adecuado os efectos espaciais. En algúns casos, os efectos espaciais poden ter unha forte carga estrutural, e/ou tendencias locais. Por iso, en ocasións, é moi útil dividir os efectos espaciais, f_{spat}, en dous tipos, os efectos estruturados, f_{str}, que son efectos espaciais correlacionados e suaves; e os efectos non estruturados (non correlacionados e non estruturados), f_{unstr}. Mediante a estimación dos efectos estruturados e non estruturados podemos distinguir entre ambos tipos de factores. É dicir:

$$f_{spat} = f_{str} + f_{unstr}$$

Isto débese a que, en xeral, os efectos espaciais permítenos captar factores influentes na variable resposta que non son captados por outras variables do estudo.

Por exemplo, imaxinemos que queremos estudar os prezos dos alugamentos de pisos en Galicia. Se só temos en conta variables como o tamaño e estado da vivenda, o ano de construción, estamos obviando factores tan importantes como os espaciais, que permiten estudar,

diferencias do valor do solo en función da localización que non son captadas por outras variables do estudo. Pero ademais, incluíndo efectos espaciais non estruturados, poderemos captar ou analizar a existencia de tendenciais locais presentes nalgunha rexión particular.

Neste TFM, empregaremos MRF para estimar os efectos estruturados espaciais e para os efectos non estruturados suporemos que seguen a distribución presentada en (2.8).

2.2.4. Modelado de interacións

A continuación explicaremos como introducir no noso modelo de regresión, posibles interaccións entre covariables. Unha forma común de incorporalas baséase en modelos de coeficientes variables introducidos en Hastie e Tibshirani (1993) no contexto de suavizado spline. Nestes modelos asúmese que o efecto da covariable z_{ij} varía de forma suave no rango de valores da outra covariable, x_{ij}:

$$f_j\left(x_{ij}, z_{ij}\right) = g_j(x_{ij}) z_{ij}$$

Na maioría dos casos a variable de interacción, z_{ij}, é categórica. Mentres que o efecto modificador pode ser métrico, espacial ou un indicador de grupo desordenado. A diferencia disto, en Hastie e Tibsirani (1993) tan só se poden incluír efectos métricos.

Para estimar a función non linear g_j podemos empregar as densidades descritas na Sección 2.1.1. no caso de efectos métricos, 2.2.2. se son espaciais ou 2.2.3 se son grupos.

Os modelos con efectos espaciais modificadores, empréganse para modelar interaccións espazo temporais. Dende un punto de vista clásico, os modelos que incluen este tipo de efectos modificadores son denominados modelos de pendentes aleatorias). En notación matricial, tense que o vector das funcións de avaliación $f_j = diag(z_{1j}, \ldots, z_{nj}) V_j^* \gamma_j$, onde V_j^* é a matriz de deseño correspondente a densidade previa de g_j. Polo tanto, neste caso a matriz de deseño xeral: $V_j = diag(z_{ij}, \ldots, z_{nj}) V_j^*$.

No caso das interaccións entre covariables métricas, poden empregarse aproximacións non paramétricas baseades en axustes de superficies bidimensionais. Unha posibilidade, é supoñer que a superficie descoñecida, $f_j(x_{ij}, z_{ij})$, pode ser aproximada por un produto tensorial de dous P-splines unidimensionais.

$$f_j(x_{ij}, z_{ij}) = \sum_{m_1=1}^{M_j} \sum_{m_2=1}^{M_j} \gamma_{j,m_1,m_2} B_{j,m_1}(x_{ij}) B_{j,m_2}(z_{ij}).$$

De forma similar aos P-splines unidimensionais, a matriz de deseño, $V_j \in M_{n \times M_j^2}$ está composta de produtos de funcións base. Neste caso, as densidades previas de $\gamma_j = (\gamma_{j,1,1}, \ldots, \gamma_{j,M_j,M_j})'$ baséanse por exemplo en paseos aleatorios de primeira a orde, que facilmente se poden expresar na fórmula xeral introducida en (2.4), Fharmeir, Kneib e Lang (2004).

Entre as distintas posibilidades para realizar a estimación dos modelos STAR, centrarémonos na inferencia baseada na representación como modelos lineais xeneralizados mixtos (GLMM, Generalized Linear Mixed Models). Isto proporcionaranos a clave para estimar simultaneamente as funcións $f_j, j = 1, \ldots, p$, e os parámetros da varianza (ou inversa do suavizado), τ_j^2, dende unha perspectiva EB, (Sección 3.3). Antes de presentar estes cálculos, na seguinte Sección ofreceremos unha representación destes modelos como GLMM.

2.3. Representación como modelos mixtos

Para reescribir os modelos STAR como GLMM recorreremos, á formulación xeral dos modelos lineais mixtos.

Sexa $\gamma_j \in M_{d_j \times 1}$, e K_j a correspondente matriz de penalización de rango rk_j. En primeiro lugar, dividiremos os vectores dos coeficientes de regresión en dúas partes, unha parte penalizada e outra non penalizada.

$$\gamma_j = V_j^{unp} \gamma_j^{unp} + V_j^{pen} \gamma_j^{pen}. \tag{2.9}$$

Na expresión anterior, as columnas da matriz $V_j^{\text{unp}} \in M_{d_j \times d_j - rk_j}$ forman unha base do núcleo da matriz K_j. A matriz, $V_j^{pen} \in M_{d_j \times rk_j}$, determínase mediante a descomposición da matriz de penalización K_j de tal forma que:

$$V_j^{pen} = L_j \left(L_j' L_j\right)^{-1}, \text{ onde } K_j = L_j L_j'.$$

Non obstante, a descomposición anterior debe verificar ademais que: $L_j' V_j^{unp} = 0$ e $V_j^{unp} L_j' = 0$.

Na expresión, (2.9), o vector γ_j^{unp}, representa a parte de γ_j non penalizada por K_j, mentres que o vector γ_j^{pen} representa a desviación de γ_j do núcleo de K_j.

En xeral, a descomposición da matriz de penalización, $K_j = L_j L_j'$, calcúlase mediante a descomposición espectral, $K_j = \Gamma_j W_j \Gamma_j'$. A matriz diagonal, $W_j \in M_{rk_j \times rk_j}$, contén na súa diagonal os autovalores positivos $w_{jm}, m = 1, \ldots, rk_j$ de K_j de maior a menor, é dicir, $W_j = diag(w_{j1}, \ldots, w_{j,rk_j})$. Γ_j é unha $(d \times rk_j)$ matriz ortogonal que contén aos correspondentes autovectores.

Unha vez obtida a descomposición espectral, tomamos $L_j = \Gamma_j W_j^{\frac{1}{2}}$. Aínda que en realidade, nalgúns casos, existe unha descomposición espectral máis adecuada, por exemplo no caso dos P-splines, (presentados anteriormente), é mellor elixir $L_j = D'$, onde D é unha matriz de diferenzas de primeira ou segunda orde. Debemos ter en conta que, por exemplo, no caso da distribución previa presentada na Sección 3.1.3 (*de efectos aleatorios*) non é necesario descompoñer a matriz de penalización, posto que $K_j = I$. E neste caso, a parte non penalizada desaparece por completo.

A matriz V_j^{unp} é o vector identidade 1 para P-splines con penalizacións de camiños aleatorios de primeira orde e MRF. Para P-splines con penalizacións de camiños aleatorios de segunda orde, V_j^{unp} é unha matriz de dúas columnas. A primeira delas, é de novo, un vector identidade, e a segunda está composta polos nodos do spline (equidistantes).

Da ecuación (2.9) podemos deducir que:

$$\frac{1}{\tau_j^2} \gamma_j' K_j \gamma_j = \frac{1}{\tau_j^2} \left(\gamma_j^{pen}\right)' K_j \gamma_j^{pen}.$$

No caso da distribución previa máis xeral introducida en (2.4), para γ_j, séguese que:

$$p(\gamma_{jm}^{unp}) \propto const, m = 1, \ldots, d_j - rk_j$$

e

$$\gamma_j^{pen} \sim N(0, \tau_j^2 I). \tag{2.10}$$

Finalmente, definindo a matriz $\tilde{U} = V_j V_j^{unp}$ e $\tilde{V}_j = V_j V_j^{pen}$ e tendo en conta (2.9), podemos reescribir o preditor introducido en (2.3) como:

$$\eta = \sum_{i=1}^{p} V_j \gamma_j + X\beta = \sum_{j=1}^{p} \left(V_j V_j^{unp} \gamma_j^{unp} + V_j V_j^{pen} \gamma_j^{pen} \right) X\beta = \tilde{U}\gamma^{unp}\tilde{V}\gamma^{pen}.$$

Onde a matriz de deseño $\tilde{V}$ e os vectores de γ^{pen} están compostos das matrices V_j e dos vectores β_j^{pen}. É dicir, $\tilde{V} = (\tilde{V}_1, \ldots, \tilde{V}_p)$, e o vector $\gamma^{pen} = ((\gamma_1^{pen}), \ldots, (\gamma_p^{pen}))'$.

De forma similar a matriz $\tilde{U}$ e o vector γ^{unp}, veñen dados por $\tilde{U} = (\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_p, X)$ e $\gamma^{unp} = \left((\gamma_1^{unp})', \ldots, (\gamma_p^{unp})', \beta' \right)'$, respectivamente.

Empregando as matrices de deseño apropiadas, $\tilde{V}$ e $\tilde{U}$, podemos expresar o modelo como suma de efectos fixos, γ^{unp}, e aleatorios, γ^{pen}. Suporemos que $\gamma^{pen} \sim N(0, \Lambda)$, onde $\Lambda = \tau_j^2 I$. Desta forma poderemos empregar a metodoloxía dos modelos GLMM para estimar simultaneamente as funcións f_j e os parámetros da varianza τ_j^2, tal e como veremos na seguinte sección.

2.4. Inferencia baseada na metodoloxía dos modelos GLMM

A inferencia Baiesiana baséase na distribucións posteriores do modelo que dependen fundamentalmente da parametrización de cada modelo. Neste caso, a distribución posterior para a inferencia FB vén dada por:

$$p\left(\gamma_1, \ldots, \gamma_p, \tau_1, \ldots, \tau_p^2, \beta \mid y\right) \propto L(y, \gamma_1, \ldots, \gamma_p, \beta) \prod_{j=1}^{p} \left(p\left(\gamma_j \mid \tau_j^2\right) p\left(\tau_j^2\right) \right), \tag{2.11}$$

onde $L(\cdot)$ denota a verosimilitude que é o produto das contribucións das verosimilitudes individuais.

Para a inferencia EB, as varianzas τ_j^2 considéranse constantes e en consecuencia as distribucións previas, $p(\tau_j^2)$, desaparecen da expresion anterior. Polo tanto, en termos da representación dos modelos GLMM obtense:

$$p(\gamma^{unp}, \gamma^{pen} \mid y) \propto L(y, \gamma^{unp}, \gamma^{pen}) \prod_{j=1}^{p} \left(p\left(\gamma_j^{pen} \mid \tau_j^2 \right) \right), \tag{2.12}$$

onde $p(\gamma^{pen} \mid \tau_j^2)$ foi definida en (2.10).

Existen varios mecanismos para realizar inferencia nos modelos STAR, non obstante, nesta Sección basearémonos na posibilidade de representalos en forma dos modelos GLMM (Sección 3.2) dende o punto de vista EB. Para elo, empregaremos o método de mínimos cadrados con pesos de xeito iterativo (IWKS, iteratively weighted least squares) e (aproximacións) de máxima verosimilitude (REML) desenvolvidas para os GLMM. Realizaremos o proceso inferencial en dous pasos:

En primeiro lugar, obtemos unha primeira estimación de γ^{unp} e γ^{pen} como solucións do seguinte sistema lineal (dados os parámetros da varianza):

$$\begin{pmatrix} \tilde{U}'W\tilde{U} & \tilde{U}'W\tilde{V} \\ \tilde{V}'W\tilde{U} & \tilde{V}'W\tilde{V} + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \gamma^{unp} \\ \gamma^{pen} \end{pmatrix} = \begin{pmatrix} \tilde{U}'W\tilde{y} \\ \tilde{V}'W\tilde{y} \end{pmatrix} \tag{2.13}$$

O vector $\tilde{y} \in (nx1)$ e a matriz diagonal $W = diag(w_1, \ldots, w_n)$ son as observacións e os pesos empregados comunmente nos modelos lineais xerais, ver Capítulo 2.2.1 de Fharmeir e Tutz (2001).

O segundo paso consiste en estimar os parámetros da varianza $\widehat{\lambda}_j^2$ maximizando (a aproximación) o logaritmo da verosimilitude restrinxida:

$$l^*\left(\tau_1^2, \ldots, \tau - p^2\right) = -\frac{1}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\log\left(\left|\tilde{U}\Sigma^{-1}\tilde{U}\right|\right) - \frac{1}{2}\left(\tilde{y} - \tilde{U}\widehat{\gamma}^{unp}\right)' \sigma^{-1}\left(\tilde{y} - \widehat{\gamma}^{unp}\right) \tag{2.14}$$

con respecto a varianza dos parámetros $\tau^2 = \left(\tau_1^2, \ldots, \tau_p^2\right)'$. Na expresión anterior, $\Sigma = W^{-1} + \tilde{V}\Lambda\tilde{V}'$ é unha aproximación da matriz de covarianzas de $\tilde{y} \mid \widehat{\beta}^{pen}$.

Finalmente, repetiremos os dous pasos anteriores ata obter a converxencia. Maximizaremos (2.14), empregando por exemplo unha alternativa, numérica e eficiente, ao usual Fisher scoring iterations, como a descrita en 1997 por Harville; ver Fharmeir, Kneib e Lang (2004) [1].

Nota

Para realizar o cálculo de intervalos de confianza das estimacións $\widehat{f}_j$ debemos de partir da fórmula (2.13), Fahrmeir, Kneib e Lang (2004). Denotemos por H, a matriz do lado esquerdo da fórmula (2.13), a aproximación da matriz de covarianzas dos coeficientes de regresión $\widehat{\gamma}^{unp}$ e $\widehat{\gamma}^{pen}$ vén dada por H^{-1}. Dada $\widehat{f}_j = \tilde{U}_j\widehat{\gamma}_j^{unp} + \tilde{V}_j\widehat{\gamma}_j^{pen}$, obtense a matriz de convarianzas:

$$Cov(\widehat{f}_j) = \left(\tilde{U}_j\tilde{V}_j\right) Cov\left(\left(\widehat{\gamma}_j^{unp}\right)'\left(\widehat{\gamma}_j^{pen}\right)'\right)\left(\tilde{U}_j\tilde{V}_j\right)'.$$

Na expresión anterior $Cov\left(\left(\widehat{\gamma}_j^{unp}\right)'\left(\widehat{\gamma}_j^{pen}\right)'\right)$ obténse a partir dos correspondentes bloques de H^{-1}.

[1] Habitualmente, o logaritmo da verosimilitude restrinxida (2.14) adoita maximizarse segundo unha puntuación de Fisher, é dicir, $\widehat{\lambda}^2 = \tilde{\lambda} + F^*(\tilde{\tau}^2)^{-1}s^*\left(\tilde{\tau}^2\right)$, onde $\tilde{\lambda}^2$ denota aos parámetros da varianza da última interacción. O vector $s^*(\tau^2)$ vén dado por: $s_j^*(\tau^2) = -\frac{1}{2}tr\left(P\tilde{V}_j\tilde{V}_j{}'\right) + \frac{1}{2}\left(\tilde{y} - \tilde{U}\widehat{\gamma}^{unp}\right)'\Sigma^{-1}\left(\tilde{y} - \tilde{U}\widehat{\gamma}^{unp}\right), j = 1, \ldots, p$. Sendo,

$$P = \Sigma^{-1} - \Sigma^{-1}\tilde{U}\left(\tilde{U}'\Sigma^{-1}\tilde{U}\right)^{-1}\tilde{U}'\Sigma^{-1}. \tag{2.15}$$

Mentres que a información de Fisher, $F^*(\tau^2)$, vén dada por:

$$F_{jk}^*(\tau^2) = \frac{1}{2}tr\left(P\tilde{V}_j\tilde{V}_j'P\tilde{V}_j\tilde{V}_j'\right), j, k = 1, \ldots, p. \tag{2.16}$$

Non obstante, cando se dispoñen de moitas observacións as fórmulas (2.15) e (2.16) non son eficientes computacionalmente. Por exemplo, para $n = 3000$ observacións, tan só para o cálculo de Σ^{-1} fan falla $O(n^3)$. En Fharmeir, Kneib e Lang (2004), descríbese unha forma de evitar o cálculo desta inversa, que solucionará o problema computacional presentado anteriormente.

2.5. Modelos de regresión estruturada de risco

Unha das principais vantaxes que se introducen cos modelos STAR, ademais de que nos permiten introducir dunha maneira cómoda diferentes tipos de variables preditoras, é o modelado das respostas. Entre elas destacamos a posibilidade de realizar estudos de supervivencia. Nesta Sección presentaremos unha extensión dos modelos clásicos paramétricos de supervivencia de Cox (1972) que empregaremos no Capítulo 4 para analizar a supervivencia dos pacientes con síndrome coronario agudo na área hospitalaria de Santiago de Compostela. Estes modelos permitiranos estimar de forma flexible as covariables involucradas nos procesos de supervivencia, así como a estimación suave da taxa de risco basal; incluíndo, ao mesmo tempo covariables xeográficas.

Nos últimos anos os modelos de regresión estruturada de risco, Kneib (2005), están recibindo moita atención. Estes modelos, permítennos estudar, por exemplo, o tempo que transcorre entre a diagnose e a morte dun paciente. Pero tamén son aplicables noutros campos con estruturas de datos similares, por exemplo, en enxeñería (ruptura de máquinas) ou por exemplo, nas ciencias sociais (p.ex para estudar o tempo que transcorre entre unha boda e o divorcio). En xeral, son útiles naquelas situacións nas que se desexa estudar a influencia dalgunhas covariables na duración dun determinado evento.

O modelo clasicamente empregado na análise de supervivencia é o de Cox (1972). Nel as covariables determinan unha taxa de risco sobre a variable resposta:

$$\lambda(t \mid x_1, \ldots, x_p) = \lambda_0(t) \exp\left(x_1\gamma_1 + \cdots + x_p\gamma_p\right).$$

É dicir, o modelo de Cox expresa a taxa de risco como o produto dunha taxa base de risco inespecífica, $\lambda_0(t)$, que non depende das covariables; $x_1, \ldots, x_p$; e a exponencial dun preditor linear que non depende do tempo.

O modelo de Cox presentado, denomínase modelo de risco proporcional, porque o ratio das taxas de risco de dous individuos con vectores de covariables, u_1 e u_2, son proporcionais, é dicir, non dependen de t:

$$\frac{\lambda\left(t \mid u_1\right)}{\lambda\left(t \mid u_2\right)} = \exp\left(u_1 - u_2\right)' \gamma.$$

Non obstante, o modelo de Cox posúe limitacións importantes (Kneib, 2005), pois, en ocasións, supoñer un efecto linear sobre as variables preditoras é demasiado restritivo, ademais poden existir interaccións entre as covariables. E ao mesmo tempo, tamén pode ocorrer que a supervivencia dos pacientes estea correlacionada espacialmente, en función do lugar de residencia ou o lugar de tratamento.

Co obxectivo de solventar estes problemas presentaremos os modelos de regresión estruturados de risco. Para elo, reparametrizaremos a taxa base de risco, $g_0(t) = \log \lambda_0(t)$, e modelando de forma diferente os distintos tipos de covariables, extenderemos os modelos de Cox a uns modelos de taxas de risco estruturados da seguinte forma:

$$\lambda_i = \exp\left(\eta_i(t)\right), i = 1, \ldots, n, \tag{2.17}$$

empregando o seguinte preditor aditivo estruturado (Kneib e Fharmeir, 2007):

$$\eta_i(t) = g_0(t) + v_i'\gamma + \sum_{l=1}^{L} g_l(t)u_{il} + \sum_{l=1}^{L} f_j(x_{ij}) + f_{spat}(s_i) + b_s. \tag{2.18}$$

Na ecuación anterior (2.18), $g_l(t)$ denota aos efectos tempo dependentes das covariables u_l; $f_j(x_j)$ son os efectos non lineais das covariables continuas, x_j. O vector γ contén os efectos lineais usuais. Finalmente $f_{spat}(s)$ denota aos efectos espaciais estruturados da rexión s e b_s os efectos espaciais non estruturados. De novo a división dos efectos espaciais en dous tipos de efectos, sérvenos para detectar fortes tendencias espaciais e posibles variacións locais.

Ademais, estendendo o preditor introducido en (2.18), pódense incorporar interaccións entre dúas covariables continuas ou pendentes aleatorias (Fharmeir e Kneib, 2007).

De forma similar que na Sección anterior, podemos obter unha formulación xeral de (2.18) como modelos mixtos e deste xeito estimar os coeficientes. En Fharmeir e Kneib (2007), poden consultarse os detalles desta aproximación.

As variables γ considéranse efectos fixos e suporemos que $p(\gamma) \propto const.$ As funcións descoñecidas g_l e f_j, son modeladas empregando P-splines (Eliers e Marx, 1996) de forma análoga que na Sección anterior. Finalmente para realizar a estimación dos efectos espaciais estruturados empregaremos MRF.

Tal e como veremos no estudo da supervivencia do síndrome coronario agudo na área sanitaria de Santiago de Compostela, nos estudos de supervivencia é moi importante dispor dun mecanismo que nos permita medir a capacidade diagnóstica dun modelo. A continuación, tomando como referencia básica o artigo de Heagerty e Zheng (2005), introduciremos os conceptos de *sensibilidade incidente* e *especificidade dinámica* que nos permitirán definir curvas ROC dependentes do tempo (Receiver Operating Charaterisctic, Heagerty et al., 2005). No capítulo 4, empregaremos estes conceptos para avaliar ou medir a capacidade de discriminación do modelo de supervivencia que introduciremos.

2.5.1. Capacidade de discriminación do modelo

Co obxectivo de estudar a capacidade de discriminación dun modelo de supervivencia como o presentado en (2.18), introduciremos novos conceptos que nos permitirán definir curvas ROC tempo dependentes (Time-Dependent Receiver Operating Charasteristic, ver: Heagerty et al., 2000 e Heagerty e Zheng, 2005) a partir das cales calcularemos unha medida global para estimar a concordancia do modelo.

Curvas ROC tempo dependentes

A curva ROC, é unha ferramenta estatística que nos permite representar a sensibilidade [2] e a especificidade [3] dun marcador continuo, M, para discriminar por exemplo individuos sás ($D = 0$) dos enfermos ($D = 1$). Non obstante, en moitas enfermidades os resultados dependen do tempo, así en vez de empregar unha variable dicotómica, D, como indicadora dunha enfermidade sería máis correcto considerar unha función, $D(t)$, que dependa do tempo. E da mesma forma considerar curvas ROC que varíen como función do tempo. Un exemplo común, é o estado vital dun paciente que dun instante de tempo a outro pode cambiar.

Heagerty et al., (2000) propuxeron calcular curvas ROC baseadas nunha definición *cummulative/dynamic* da sensibilidade e a especificidade variante con tempo. Outra definición, chamada *incident/dynamic* da sensibilidade e a especificidade variable con tempo foi definida

[2] A sensibilidade é a probabilidade de que o test determine un *verdadeiro positivo*, é dicir, que indique que un individuo está enfermo cando efectivamente está enfermo.

[3] A especificidade é a probabilidad de que o test determine que un individuo está sá cando realmente o está, *verdadeiro negativo*.

en Heagerty e Zheng (2005) que será a que empregaremos neste traballo. Fundamentalmente, esta definición baséase, en considerar casos e controis. Ademais, empregando estas novas definicións de sensibilidade e especificidade poderemos definir a curva ROC (I/D) e desta maneira obter o $AUC(t)$ (Area Under Curve) a partir do cal calcularemos unha medida global para estimar a concordancia do modelo.

Notación

Introduciremos a seguinte notación:

Sexa T_i o tempo de supervivencia do individuo i e suporemos que só se observa o mínimo de T_i e C_i, sendo C_i un tempo de censura independente do paciente i. Definiremos o tempo de seguimento como $X_i = min(T_i, C_i)$, e $\delta_i = 1(T_i \leq C_i)$ denota o indicador de censura.

O tempo de supervivencia, T_i, tamén se pode representar como un proceso de conteo, $N_i^*(t) = 1(T_i \leq t)$ ou o correspondente incremento, $dN_i^*(t) = N_i^*(t) - N_i^*(t-)$.

Sexa $R_i(t) = 1(X_i \geq t)$ un indicador de risco. Ademais suporemos que cada individuo i posúe un conxunto de covariables invariantes, $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$.

Extensión dos conceptos de sensibilidade e especificidade

Nos test definidos en escalas continuas, as curvas ROC, son un método estándar para medir a capacidade de discriminación. Sexa Y_i a variable binaria indicadora de enfermidade, $\widehat{p}_i$ unha predición e c un criterio que nos permite clasificar as predicións como positivas ($\widehat{p}_i > c$) ou negativas ($\widehat{p}_i \leq c$). O obxectivo desta Sección é estender os conceptos de sensibilidade, $P(\widehat{p}_i > c \mid Y_i = 1)$, e especificidade, $P(\widehat{p}_i \leq c \mid Y_i = 0)$, para variables binarias tempo dependentes, como por exemplo o estado vital dun paciente que vai variando a medida que transcorren os días. Cando non se dispón dun valor de c previamente fixado, represéntase os valores da sensibilidade ("verdadeiros positivos") fronte a (1 − especificidade), "falsos positivos", para todos os posíbeis valores de $c \in (-\infty, \infty)$.

Heargety et al. (2005), propoñen unha modificación destas definicións de sensibilidade e especificidade que nos permitirá construír curvas ROC que dependan do tempo, $ROC(t)$. Para elo, nun determinado tempo t, dividiremos os individuos en dous grupos excluentes, os **casos** que serán aqueles pacientes que en t teñen risco de padecer o evento e **controis** que son individuos xa falecidos en t. Dado un valor de corte c, definiremos as versións incidente e

dinámica (*incident/dinamyc*) da sensibilidade e da especificidade, tal e como segue, Heagerty e Zheng (2005):

$$\text{sensibilidade}^{\mathbb{I}}(c,t) : P\left(M_i > c \mid T_i = t\right) = P\left(M_i > c \mid dN_i^*(t) = 1\right),$$

$$\text{especificidade}^{\mathbb{D}}(c,t) : P\left(M_i \leq c \mid T_i > t\right) = P\left(M_i \leq c \mid N_i^*(t) = 0\right).$$

A sensibilidade representa a fracción de individuos falecidos en t cuxo marcador é maior que c. Por outra banda, a especificidade representa a fracción de suxeitos cun marcador menor ou igual a c entre os que sobreviven máis alá dun tempo t. A vantaxe que nos ofrecen estas definicións é que neste caso, un individuo i, pode ser considerado un control nun determinado momento $t < T_i$, pero máis tarde pasar a ser considerado un caso cando $t = T_i$.

Curvas ROC tempo dependentes

Empregando a definición de sensibilidade incidente e especificidade dinámica, podemos calcular novas curvas ROC. Neste traballo fin de máster basearémonos nas curvas ROC (I/D). Estas curvas defínense como unha función $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$, onde p denota as taxas dinámicas dos falsos positivos (1-especificidade), e $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$ denota as correspondentes taxas incidentes dos verdadeiros positivos. Especificamente, consideremos c^p de forma que:

$$P(M_i > c_p \mid T_i > t) = 1 - \text{especificidade}^{\mathbb{D}}(c^p, t) = p.$$

A verdadeira taxa positiva, $ROC_t^{\mathbb{I}/\mathbb{D}}(p)$ é a sensibilidade que se obtén empregando este umbral, é dicir, $ROC_t^{\mathbb{I}/\mathbb{D}}(p) = \text{sensibilidade}^{\mathbb{I}}(c^p, t) = P(M_i > c^p \mid T_i = t)$. Empregando as funcións das taxas dos verdadeiros e falsos positivos, $TP_t^{\mathbb{I}}(c) = \text{sensibilidade}^{\mathbb{I}}(c^p, t)$, e $TP_t^{\mathbb{D}}(c) = 1 - \text{especificidade}^{\mathbb{D}}(c^p, t)$, permítennos escribir as curvas ROC como composición de $TP_t^{\mathbb{I}}(c)$, e a inversa da función, $\left[TP_t^{\mathbb{D}}\right]^{-1}(p) = c^p$, Heagerty e Zheng (2005):

$$ROC_t^{\mathbb{I}/\mathbb{D}}(p) = TP_t^{\mathbb{I}}\left\{\left[FP_t^{\mathbb{D}}\right]^{-1}(p)\right\}, \text{ para } p \in [0,1].$$

AUC tempo dependente

As definicións anteriores baseadas en definicións incidente/dinámicas das curvas ROC, permítenos definir o AUC (Area Under Curve), que consideraremos como unha medida global de concordancia para cada tempo t:

$$AUC(t) = \int_0^1 ROC(t)^{\mathbb{I}/\mathbb{D}}(p)dp.$$

O AUC é unha medida da capacidade diagnóstica da curva ROC. Máis especificamente, mide a probabilidade de que o test diagnóstico dun enfermo seleccionado ao azar (caso) sexa maior que o doutro paciente sá seleccionado ao azar (control). Esta área toma valores entre 0.5 e 1. Valores próximos a un indicarán que a capacidade preditiva do modelo é boa, mentres que valores pretos ao 0.5, significan que o modelo non é moi informativo, Heagerty e Zheng (2005).

2.5.2. Índice C de concordancia

Na sección anterior, presentamos os métodos ROC como ferramentas para determinar a capacidade dun marcador e desta forma distinguir casos e controis nun tempo, t. Pero, ás veces, non se dispón dun t previamente identificado, nestes casos sería de gran utilidade dispoñer dunha medida global de precisión. A continuación describiremos como a partir das curvas ROC tempo dependentes podemos construír un resumo estándar de concordancia.

A medida global de resumo que adoptaremos vén dada pola seguinte expresión, que expresa a probabilidade de que un individuo falecido antes que outro teña un marcador máis grande:

$$C = P\left[M_j > M_k \mid T_j < T_k\right].$$

Supoñamos que as observacións (M_j, T_j) e (M_k, T_k) son independentes e que T_j é continuo de forma que $P(T_k = T_j) = 0$. Neste contexto, $P\left[T_j < T_k\right] = \frac{1}{2}$. Nas contas sucesivas denotaremos por $P(x)$ a probabilidade ou a densidade segundo conveña en cada contexto.

C defínese como unha ponderación de pesos medios da área baixo a curva ROC tempo dependente. Poden consultarse os detalles en Heagerty e Zheng (2005):

$$
\begin{aligned}
& P\left[M_j > M_k \mid T_j < T_k\right] \\
& = P\left[\{M_j > M_k\} \cap \{T_j < T_k\}\right] \times 2 \\
& = \int_t P\left[\{M_j > M_k\} \cap \{T_j = t\} \cap \{t < T_k\}\right] \times 2dt \\
& = \int_t P\left[\{M_j > M_k\} \cap \{T_j = t\} \cap \{t < T_k\}\right] \times 2dt \\
& = \int_t P\left[\{M_j > M_k\} \mid \{T_j = t\} \cap \{t < T_k\}\right] \times 2 \times P\left[\{T_j = t\} \cup \{t < T_k\}\right] dt \\
& = \int_t AUC(t) \times 2 \times P\left[T_j = t\right] \times P\left[t < T_k\right] dt \\
& = \int_t AUC(t) \times w(t) dt = E_T\left[AUC(T) \times 2 \times S(T)\right],
\end{aligned}
$$

onde $w(t) = 2f(t)S(t)$. Baseándonos na definición I/D da sensibilidade e a especificidade, $AUC(t) = P\left(M_j > M_k \mid T_j = t, T_k > t\right)$.

O índice C que acabamos de introducir sérvenos como unha medida global para avaliar a capacidade preditiva dos modelos de supervivencia. Un valor de 0.5 indica que o modelo non ten capacidade preditiva, mentres que se acada o valor 1, indicará que se realizou unha separación perfecta.

2.6. Implementación de modelos STAR: Bayes X

Nos Capítulos 3 e 4, realizaremos a análise de dúas bases de datos biomédicas empregando a metodoloxía presentada. No Capítulo 3, veremos un exemplo dun modelo STAR con resposta de Poisson mentres que no Capítulo 4, presentaremos unha análise de supervivencia. A estimación dos modelos que presentaremos foi estimada empregando a tecnoloxía de modelos mixtos e `remlreg objects` do software BayesX.

Bayes X (Brezger, Kneib, e Lang, 2005) é un programa de dominio público desenvolvido na década pasada no Departamento de Estadística da Universidade de Munich. Trátase

dun software que nos permite estimar os modelos de regresión aditiva estruturada. Ademais posúe ferramentas que nos permiten manipular as bases de datos e mapas xeográficos así como visualizar os resultados obtidos. Na páxina web, http://www.statistik.lmu.de/ bayesx/bayesx.html, podemos descargalo de forma totalmente gratuíta xunto con varios manuais de referencia.

Actualmente en R existen varios paquetes nos que están implementados os modelos STAR entre os que se atopan os paquetes *BayesX* e *R2BayesX*, que son os máis próximos ao software orixinal BayesX. Non obstante, na actualidade estes paquetes non permiten acceder automaticamente ao propio programa BayesX, aínda que os autores están a traballar nesta mellora (Brezger, Kneib, e Lang, 2005).

Neste Traballo Fin de Máster empregaremos o programa BayesX para realizar as estimacións dos modelos STAR, así como os paquetes de R anteriores como interface gráficos para visualizar os efectos das covariables e a información xeográfica.

Capítulo 3

Patróns espaciais na taxa de abstinencia do alcohol

Nas últimas décadas, son moitos os estudos nos que se considera a área xeográfica como un factor decisivo a ter en conta nos estudos clínicos. A análise do impacto destes efectos xeográficos é especialmente importante para capturar posibles heteroxeneidades espaciais. Neste Capítulo pretenderemos investigar as tendencias espaciais na taxa de abstinencia ao alcohol (AWS) en Galicia. Ademais, para poder ter en conta outros posibles factores de influencia, empregaremos os modelos STAR cunha resposta Poisson. Desta forma poderemos modelar de forma flexible os efectos non lineais xunto cos espaciais.

3.1. Introdución

O Síndrome de abstinencia alcohólica, AWS (do inglés, Alcohol Withdrawal Syndrome) é a expresión clínica da interrupción brusca ou diminución da inxesta de alcohol por outras razóns de saúde que sofre un paciente cunha dependencia física ao mesmo. En xeral, manifféstase despois de entre 6 e 24 horas de abstinencia, se esta vén dada por unha enfermidade ou lesión; ou voluntaria, por exemplo, tras unha abstinencia forzada no curso dunha desintoxicación programada. (Monte Secades, e Rabuñal Rey, 2011) [1].

[1]Parte de este Capítulo ao igual que o vindeiro foron presentados na Conferencia Internacional de Estatística Espacial que tivo lugar en Avignon (Francia) do 9 ao 12 de Xu-

A definición do síndrome de abstinencia alcohólica universalmente aceptada na literatura é a aportada polo manual DSM-IV. Nela exíxense varias condicións:

- Interrupción ou redución dun consumo forte e prolongado de alcohol.
- Aparición posterior, nunhas horas do día, de dous ou máis síntomas: hiperactividade autonómica (sudación, taquicardia, hipertensión), tremor de mans, insomnio, alucinacións táctiles, auditivas ou visuais, náuseas, vómitos, axitación psicomotriz e ansiedade.
- Os síntomas do criterio anterior producen alteracións clinicamente significativas na esfera cognitiva, social ou ocupacional.
- Os síntomas non son debidos a outra enfermidade médica ou psiquiátrica definidas.

A intoxicación por alcohol e pola abstinencia ao mesmo, representa un perigo considerable. En xeral, os pacientes perden o control do consumo de alcohol e non poden controlar a cantidade de alcohol que necesitan inxerir continuamente. Ademais, a redución da dose de alcohol pode conducir a síntomas de abstinencia desagradables que o paciente palía inxerindo novas doses. Habitualmente, continúan co consumo de alcohol a pesar de coñecer as graves consecuencias que ocasionan tanto na súa saúde (Lukasik e Sommerfeld, 2014) como na vida persoal (perda de amizades, ruptura matrimonial, perda de relación cos familiares) e profesionais (problemas no traballo ou incluso a perda do mesmo o que ocasiona a falta de medios de subsistencia...)

A abstinencia ao alcohol é un problema frecuente no medio hospitalario. Ademais a repercusión sobre a evolución clínica dos pacientes é moi importante, chegándose a triplicar, por exemplo, a mortalidade no postoperatorio daquelas persoas que sofren este síndrome. Ademais, recóllese que 66 % dos españois adultos maiores de 15 anos consumiron alcohol durante o ano pasado. Ademais o 5 % da poboación adulta teñen o risco de sufrir graves enfermidades relacionadas co consumo do alcohol (Enquisa Nacional de Saúde, 2011-2012.) España atópase entre os líderes en Europa en termos de consumo de alcohol, cunha taxa anual de case 10 litros de consumo de alcohol puro por persoa (Organización Mundial da Saúde, 2014). En Galicia, esta taxa de consumo anual per cápita é un 40 % máis alta que a media nacional e ademais existe unha alta prevalencia de abuso do alcohol, polo que os

ño (http://www.spatialstatisticsconference.com) e publicados en Environmental Science Procedia (http://www.journals.elsevier.com/procedia-environmental-sciences/).

ingresos hospitalarios en Galicia por AWS non son precisamente casos aislados (Mateos et al., 2002).

A pesar do comentado anteriormente, o AWS é unha enfermidade de dependencia ao alcohol (Asociación Americana de Psiquiatría, 2000) que nos últimos tempos recibiu pouca atención. Os escasos estudos clínicos existentes na literatura, e de guías de práctica clínica sobre o seu tratamento, fai que exista unha gran variabilidade no seu manexo, non só entre diferentes países senón tamén entre os distintos centros hospitalarios ademais de entre os diferentes clínicos (Monte Secades e Rabuñal Rey, 2011).

Debido a escaseza de datos sobre a epidemioloxía do AWS, en 2011 Gonzalez-Quintela et al. publicaron un artigo no que se investigaba as tendencias espazo-temporais da tasa de abstinencia ao alcohol en Galicia e a súa posible relación con varios factores demográficos entre 1996 e 2006. Neste traballo, non se realizou unha regresión espacial, senón que se empregaron os modelos GAM (Hastie e Tibshirani, 1990) empregando unha resposta Poisson para modelar as taxas de AWS en cada municipio galego de forma separada.

Neste traballo fin de máster, reanalizaremos estes datos para investigar as tendencias espaciais das taxas de AWS, pero neste caso empregaremos os modelos STAR (presentados no Capítulo 2) empregando de novo unha resposta de Poisson. Desta forma, poderemos modelar de forma flexible tanto os efectos espaciais como os non lineais das covariables.

3.2. Descripción da base de datos

Este estudo foi realizado en Galicia, e nel inclúense todas as altas hospitalarias (dende xaneiro de 1996 a decembro de 2006) diagnosticadas de AWS (con ICD-9-CM[2] códigos, 291.8, 291.0 e 291.3.)

Para levar a cabo á análise incluiremos, ademais, algunhas variables socio-demográficas, agregadas por cada municipio galego e restrinxidas a poboación maior de quince anos, recollidas polo Instituto Nacional de Estadística[3]:

[2]En España, todas as altas hospitalarias e diagnoses (incluso defuncións) rexístranse segundo unha clasificación internacional de enfermidades (International Classification of Diseases, 9th Revision, Clinical Modification, ICD-9-CM).

[3]Para o cálculo de todas as variables socio-demográficas, tivéronse en conta os datos publicados para o ano 2001 na páxina web do Instituto Nacional de Estadística (www.ine.es)

- A **taxa de paro** (paro), calculada como o porcentaxe de persoas, en idade de traballar, que están buscando emprego. En media, o porcentaxe medio de parados entre os pacientes ingresados por este síndrome foi do 12 %. Debemos ter en conta que este dato é anterior ao comezo da crise polo que sería moi interesante ampliar este estudo ata a actualidade.

- A **taxa de actividade profesional** (actividade) calculada como a porcentaxe dos maiores de 15 anos activos economicamente, (sexan ou non traballadores). En media, esta variable foi do 43 %, acadándose o máximo no 52.77 %.

- **Nivel de estudios medio** (edu). Este nivel foi calculado como a puntuación media indivual de estudos en cada un dos concellos. En media os pacientes ingresados tiñan a educación primaria incompleta (Ver Táboa 3.1).

- O **nivel socioeconómico** (socio), foi calculado como a media da puntuación socioeconómica do sustentador principal de cada fogar galego en cada un dos municipios. Entre os ingresados por AWS, obtívose que o nivel socioeconómico medio foi 0.33. Polo tanto en media, os pacientes ingresados eran desempregados. (Ver Táboa 3.2).

Ademais incluiremos no estudo efectos espaciais co obxectivo de comprobar as impresións clínicas sobre a influencia das tendencias xeográficas na taxa de AWS.

O feito de estudar o número de episodios de AWS nunha comunidade como Galicia, ten especial importancia como comentaremos a continuación. No punto medio do período estudado, Galicia tiña unha poboación de 2780000 persoas repartidas en 315 concellos. Unha das principais características demográficas de Galicia é que a poboación está moi dispersa e distribúese de forma irregular, sendo as zonas de costa as máis poboadas. Ademais, cada concello se divide en parroquias, e cada unha delas en aldeas, en xeral con menos de 50 persoas. Aproximadamente un terzo da poboación vive nas zonas urbanas, mentres que o resto viven nas zonas semiurbanas ou rurais, precisamente, Galicia é unha das rexións con maior índice de ruralidade en España (Prieto-Lara e Ocaña-Riola, 2010).

En primeiro lugar debemos diferenciar entre a sociedade urbana galega, cuxo consumo e características podemos equiparalas ao resto de España (sobre todo o Norte) e a Galiza rural que presenta unha personalidade socioeconómica e cultural propia. Galicia ten a súa propia identidade cultural e a súa propia lingua (o galego), ademais do español. Debemos ter

en conta que Galiza é unha sociedade vitícola. De maneira tradicional o consumo de alcohol (principalmente o viño) concíbese como un alimento da vida cotiá (Mateos et al., 2002).

En Galicia, o abuso do alcohol constitúe un importante problema para a saúde pública. Nunha enquisa recente, o 95 % da poboación maior de 15 anos consumiran alcohol ao menos unha vez nas súas vidas, o 79 % consumiran alcohol nos 12 meses anteriores, o 56 % consumira alcohol nos 30 días anteriores, e o 24 % consume alcohol diariamente (Segundo os datos do Ministerio de Saúde español, 2007). É mais, segundo González-Quintela et al. (2011) en Galicia, a prevalencia dos consumidores habituais de alcohol é a mais alta de todo o país. Aproximadamente o 5-10 % da poboación galega adulta pódese considerar abusivos do alcohol, en función do nivel de risco considerado (máis de 80 ou máis 40 gramos ao día, respectivamente).

Nivel de Estudos	**Puntuación**
Sen estudos (analfabetos)	0
Educación elemental (<5 anos)	1
Educación primaria incompleta	2
Eduación primaria completa	2.5
Eduación secundaria	3
Formación profesional de primeiro grao	3.5
Educación universitaria (<4 anos)	3.5
Educación universitaria ($\geq$ 4 anos)	4
Tese doutoral (PhD)	4.5

Táboa 3.1: Puntuación do nivel de estudos segundo o Instituto Nacional de Estadística.

Nivel Socioeconómico	Puntuación
Desempregados buscando o seu primeiro emprego ou persoas inactivas	0
Persoas desempregadas	0.5
Persoas xubiladas ou institucionalizadas	1
Traballadores non cualificados (agricultura, servizo, industria, ...)	1
Pequeno negocio en agricultura, (sin empregados)	1.5
Membros de cooperativas agrarias	1.5
Propietarios de negocios en agricultura (con empregados)	2
Membros de cooperativas non agrarias; administrativos; xerentes;	2
Obreros cualificados non agrarios; militares	2
Propietarios de granxas,	2.5
Pequenos negocios (non agrarios e sen empregados)	2.5
Funcionarios públicos ou técnicos que traballan como empregados;	2.5
Propietarios de negocios non agrarios (con empregados)	3
Autónomos técnicos e profesionais	3
Directores xerais	3
Administradores gobernamentais	3

Táboa 3.2: Puntuación do nivel socioeconómico segundo o Instituto Nacional de Estadística.

3.3. Metodoloxía estadística

Para estudar a taxa de AWS empregaremos os modelos STAR, empregando unha resposta Poisson incluíndo estruturas espaciais. Adaptando a este exemplo a fórmula xeral dun modelo STAR introducida en (2.1) da Sección 2.2 do Capítulo 2 :

$$\eta = \text{offset}\left(\log(poboación)\right) + f_1(paro) + f_2(socio) + f_3(actividade) + f_4(edu) + f_{spat}(s) + b_s,$$

onde η, denota a taxa de AWS, $f_i, i = 1, \ldots, 4$; son funcións descoñecidas e suaves empregadas para modelar as covariables continuas. $f_{spat}(s)$, representa os efectos espaciais correlacionados da rexión s. Finalmente, b_s denota os efectos espaciais incorrelados e non

estruturados. Estimando de forma separada estes dous tipos de efectos espaciais, podemos estudar se existen marcadas tendencias espaciais (efectos espaciais estruturados) ou se tamén hai tendencias locais (non estruturadas).

Debemos ter en conta que necesitamos axustar o modelo tendo en conta a poboación de cada municipio incorporando como *offset* o logaritmo da poboación de cada un deles.

Para levar a cabo a inferencia empregamos as técnicas empíricas de Bayes introducidas no Capítulo 2. Na inferencia EB, tal e como comentamos, a varianza e os parámetros de suavizado considéranse constantes descoñecidas e estimarémolas empregando REML. Para modelar as covariables continuas, empregamos P-splines con 20 nodos equidistantes e densidades previas de camiños aleatorios de segunda orde (second order random walk prior). Para modelar os efectos espaciais estrutrados empregaremos cadeas de Markov aleatorias (MRF, Markov Random Fields, ver Capítulo 3). Finalmente suporemos que b_s ten a seguinte distribución previa $b_s \in N(0, \tau^2)$, sendo τ^2 o parámetro da varianza.

3.4. Resultados

Na Táboa 3.3 mostramos os resultados obtidos tras axustar os modelos tendo en conta ou non os efectos espaciais. Segundo varios criterios estatísticos, (Akaike information criterion (AIC), Bayesian information criterion (BIC), e generalized cross-validation, GCV), podemos ver como efectivamente, o modelo que inclúe os efectos espaciais é mellor.

Modelos	2 log (verosimilitude)	Graos de liberdade	AIC	BIC	GCV
Sen efectos espaciais	-40029.6	45.8481	-39937.9	-39766.1	7.36371
Con efectos espaciais	-41458.1	163.492	-41131.1	-40518.7	3.50782

Táboa 3.3: Akaike information criterion (AIC), Bayesian information criterion (BIC), e generalized cross-validation (GCV) incluíndo ou non no modelo os efectos espaciais.

Nas Figuras 3.1 e 3.2, representamos os efectos das covariables continuas incluídas no modelo. Tal e como se pode ver nas figuras anteriores, estas estimacións son semellantes en ambos casos. Non obstante, ao incluír os efectos espaciais, os efectos suavízanse e son máis doados de interpretar.

Na Figura 3.3 representamos os efectos espaciais. A principal vantaxe de incluír os efectos espaciais no noso modelo é que nos permite corroborar que efectivamente a taxa de AWS non se distribúe de forma uniforme en todo o territorio galego. En vermello represéntanse os municipios con maiores taxas medias de AWS, en gris taxas medias e en verde aparecen os municipios de menores taxas. Cabe destacar que nas grandes cidades se observaron taxas de AWS máis baixas. Mentres que os concellos coas taxas máis elevadas correspóndense na maioría dos casos con zonas rurais.

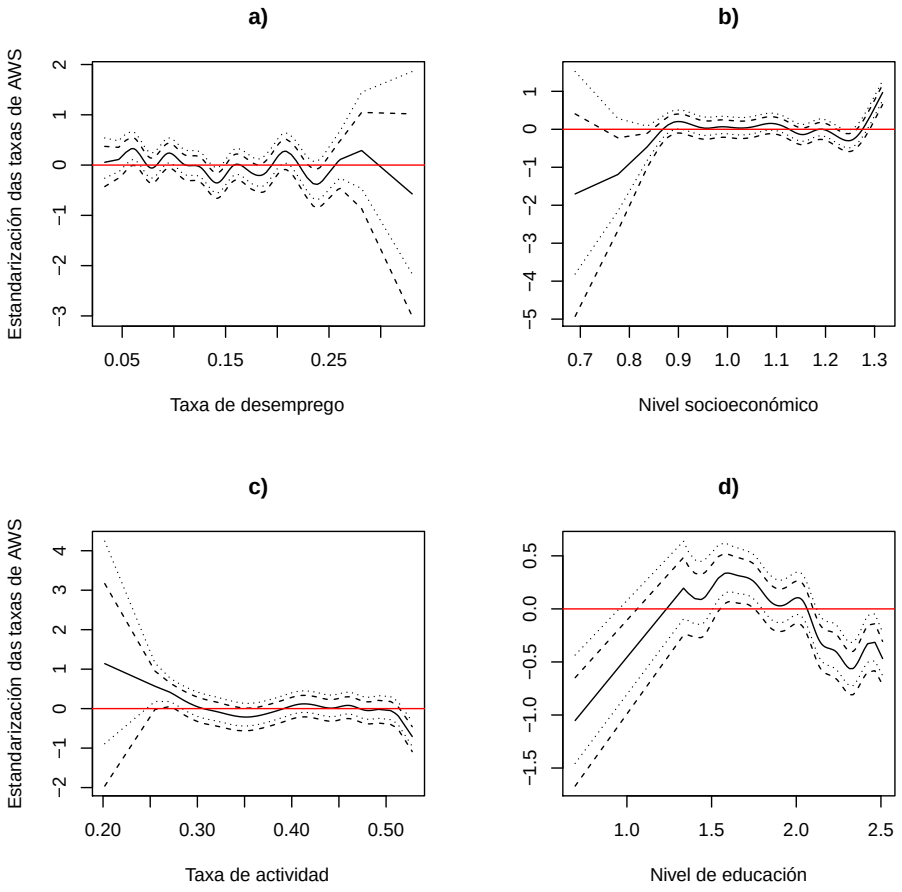


Figura 3.1: Estimación dos efectos das variables socioeconómicas: nivel socioeconómico (a), taxa de desemprego (b), taxa de actividade (c), e o nivel de educación (d) con respecto a taxa de abstinencia ao alcohol (AWS), sen incluír os efectos espaciais no modelo de regresión. Represéntanse ademais as bandas de estimación puntuais ao 95 % de confianza.

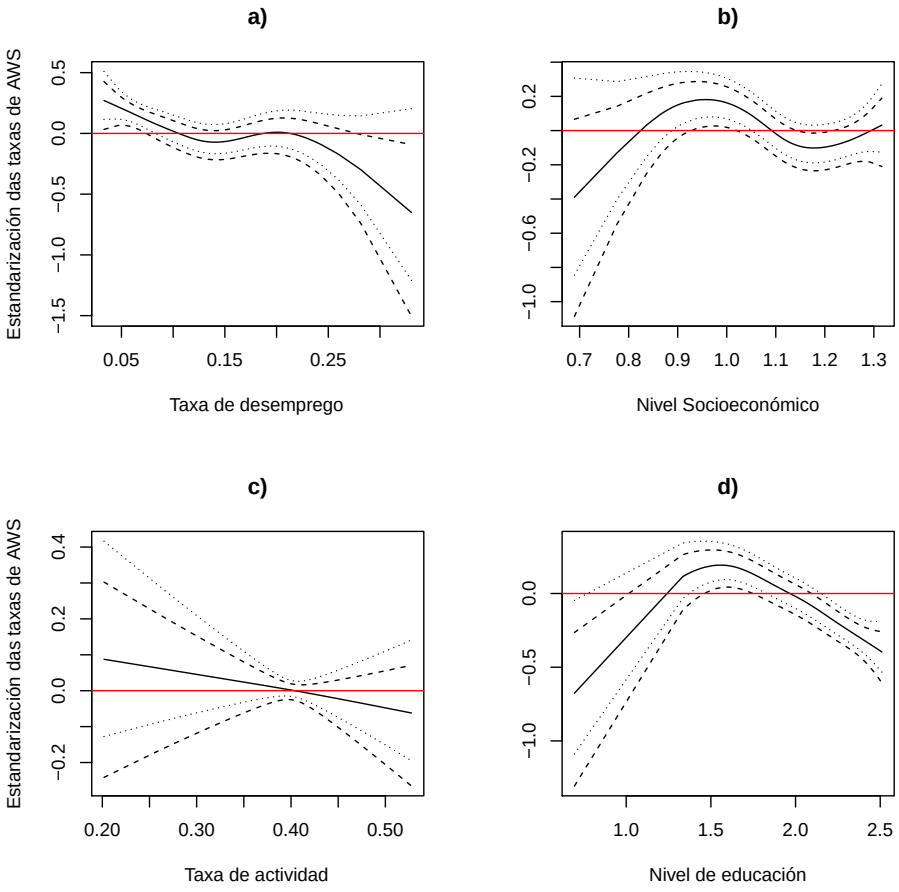


Figura 3.2: Estimación dos efectos das variables socioeconómicas: nivel socioeconómico (a), taxa de desemprego (b), taxa de actividade (c), o nivel de educación (d) con respecto a taxa de abstinencia ao alcohol (AWS), incluíndo os efectos espaciais no modelo de regresión. Represéntanse ademais as bandas de estimación puntuais ao 95 % de confianza.

Figura 3.3: Efectos espacias estruturados (a), e as probabilidades a posteriori ao 95 % de confianza. Neste caso os efectos non estruturados non son significativos. En vermello represéntanse os municipios con maiores taxas medias de AWS, en gris taxas medias e en verde aparecen os municipios de menores taxas.

3.5. Discusión

Neste estudo queda demostrado que a taxa de casos severos de AWS distribúese de xeito independente co nivel de educación da poboación de definición. Esta relación non é linear, naqueles municipios con puntuacións medias dos niveis de educación (menos de 5 anos), de forma que diminúen as taxas de AWS cando aumenta o nivel de educación. Estas relacións débense a que a distribución das taxas de AWS é desigual ao longo do territorio galego. En conxunto, estes resultados poden axudar a establecer prioridades para desenvolver medidas preventivas nalgunhas zonas. Cando se realiza a análise da variabilidade espacial, non se evidencia que o efecto do nivel educativo nas taxas de AWS continúe existindo incluso despois de que se teñan en conta outros factores de risco potenciais, como a taxa de actividade, a taxa de desemprego ou o nivel socioeconómico. Xa, para rematar, recalcaremos de novo que efectivamente cando se introduce a área xeográfica como un factor determinante na saúde,

a relación entre o nivel de educación e a taxa de AWS, é máis suave, de xeito similar ao acontecido co resto de variables.

Finalmente no Capítulo 4, co obxectivo de ilustrar a utilidade e a flexibilidade dos modelos STAR, analizaremos unha base de datos do síndrome coronario agudo na área hospitalaria de Santiago de Compostela

Capítulo 4

Supervivencia do síndrome coronario agudo na área sanitaria de Santiago de Compostela

4.1. Introdución

O síndrome coronario agudo (ACS, Acute Coronary Syndrome) comprende un conxunto de entidades producidas pola erosión ou ruptura dunha placa de ateroma, que determina a formación dun trombo intracoronario, causando unha anxina inestable, un infarto de miocardio (IAM) ou morte súbita, segundo a cantidade e duración do trombo, a existencia de circulación colateral e a presenza de vasoespasmo no momento da ruptura. (O' Connor et al., 2010). Ao longo deste capítulo investigaremos a posible existencia de desigualdades xeográficas que poidan afectar á supervivencia dos paciente con ACS. [1]

Na maioría dos estudos clínicos non se inclúe a área xeográfica como un factor determinante na saúde a pesares de que diversas investigacións demostran que, nalgúns casos, existen desigualdades territoriais que poden aumentar a mortalidade e o risco de padecer

[1] Parte de este Capítulo foi presentado na Conferencia de Estatística Espacial que tivo lugar en Avignon (Francia) do 9 ao 12 de Xuño de 2015 www.spatialstatisticsconference.com) e tamén parte dos mesmo publicado en Environmental Science Procedia (www.journals.elsevier.com/procedia-environmental-sciences).

algunhas enfermidades. Isto débese principalmente a influencia de factores socioeconómicos ou posibles diferencias nas condicións ambientais. Neste sentido, os modelos clásicos de supervivencia como o de Cox (Cox proportional hazards model, Cox (1972)) presentan limitacións para estudar este tipo de efectos espaciais, como xa comentamos na Sección 2.5 do Capítulo 2.

Neste Capítulo, empregaremos os modelos de supervivencia estruturados xeoaditivos (Kneib e Fahrmeir, 2007), introducidos na Sección 2.5. do Capítulo 2 os cales nos permitirán estimar de forma flexible os procesos de supervivencia do síndrome coronario agudo incluíndo, ao mesmo tempo, as covariables xeográficas.

Finalmente, retomando os conceptos de *sensibilidade incidente* e *especificidade dinámica* introducidos na Sección 2.5.1 do Capítulo 2, representaremos as curvas ROC dependentes do tempo (Receiver Operating Charaterisctic, Heagerty et al., 2005) e calcularemos a capacidade de discriminación do modelo de supervivencia introducido.

4.2. Descrición da base de datos

Para levar a cabo o estudo, incluímos todos os pacientes que foron ingresados no Hospital Clínico Universitario de Santiago de Compostela entre Xaneiro de 2003 e Decembro de 2010 cun diagnóstico síndrome coronario agudo con período de seguimento ata novembro de 2011.

Como comentamos o estudo realizouse na área sanitaria de Santiago de Compostela, a cal posúe unha extensión de 4574 km^2 dividida en 46 municipios, e unha poboación de 451000 habitantes en 2010, (Figura 4.1).

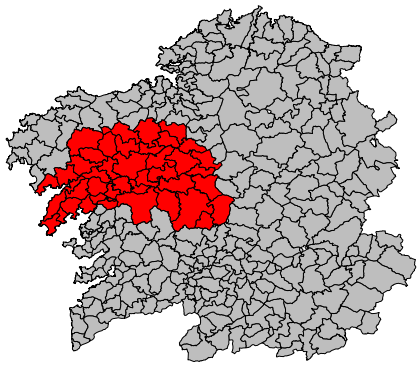


Figura 4.1: En vermello represéntase a área sanitaria de Santiago de Compostela sobre o mapa de Galicia.

Na base de datos incluíronse 4594 pacientes, dos cales 243 individuos posuían datos perdidos ou non se coñecía a súa residencia e 471 residían fora de Galicia. Polo tanto, a base de datos finalmente empregada para realizar o estudo constaba de 3880 pacientes. A idade media dos pacientes era 67 anos, a maioría homes (71 %).

Unha das complicacións que presenta a análise de supervivencia é o feito de que algúns datos obsérvanse de forma incompleta. O exemplo máis común son os datos censurados pola dereita, que aparecen cando non ocorre o evento antes de finalizar o período de seguimento. A vantaxe que nos ofrecen estes modelos, xa introducidos no Capítulo 2 é que nos permiten incorporar datos censurados. Neste estudo, o 24 % dos pacientes morreron antes de rematar o tempo do estudo polo que o 76 % dos pacientes están censurados pola dereita. A mediana do tempo de seguimento foi 1129 días (aproximadamente 3 anos). Para os pacientes que sobreviviron ao longo do período de seguimento (datos censurados) foi aproximadamente de tres anos e medio (1301 días), mentres que a mediana do tempo de seguimento para os individuos non censurados foi de 466 días (un ano e tres meses).

4.3. Formulación do modelo

A variable resposta do modelo é a morte por calquera causa que tivo lugar antes de agosto de 2011. Como covariables, incluíremos aquelas que forman parte do GRACE score [2]. Este score é considerado unha medida de risco do ACS (Granger et al., 2003). Contén tanto **covariables continuas**: *idade*, frecuencia cardíaca (*fc*), presión sanguínea sistólica, (*tas*), *creatinina* e os niveis de *troponina*; como **covariables categóricas**, como a variación na elevación do segmento ST, (*cambioST*, variable dicotómica (Si/Non)), intervención percutánea coronaria no hospital, (*icp*, variable dicotómica) e o tipo de insuficiencia cardíaca (*killip*). Killip e Kimball (1967), describiron a evolución de pacientes con infarto agudo de miocardio en función da presencia ou ausencia de determinados síntomas que indican disfunción ventricular, diferenciando catro clases. A clase 1, se o infarto non foi complicado, a clase 2, se se produciu unha insuficiencia cardíaca moderada, e a clase 3 cando esta insuficiencia se considera grave e finalmente a clase 4, considera aqueles pacientes que padeceron shock cardioxénico. Ademais incorporamos como variable o municipio de residencia dos pacientes.

Retomando o modelo de regresión aditiva estruturada de risco presentado en (2.18) na Sección 2.5 Capítulo 2, construiremos o seguinte modelo para investigar os patróns de supervivencia no ACS:

$$\begin{aligned}\eta_i(t) =& g_0(t) + \gamma_1 killip1 + \gamma_2 killip2 + \gamma_3 \text{cambioST} + \gamma_4 \text{icp} + f_1(idade) + f_2(fc) + \\ & f_3(creatinina) + f_4(\log(troponina)) + f_5(icp) + f_6(tas) + f_{spat}(s_i) + b_{s_i},\end{aligned} \tag{4.1}$$

onde a resposta η_i é unha variable binaria que vale 1 se o individuo i está falecido en t e 0 noutro caso. Por outro lado, $g_0(t)$, é o logaritmo da taxa de risco basal (centrada), $f_1, \ldots, f_6$ son funcións suaves descoñecidas das covariables continuas. Tanto $g_0(t)$ como as funcións $f_j, j = 1, \ldots, 6$ son modeladas empregando P-splines cúbicos con penalizacións de segunda orde usando 20 nodos.

En (4.1) a variable, $killip1$ toma o valor 1 se $killip = 2$ e 0 noutro caso. E $killip2$, vale 1 cando $killip = 3$ ou $killip = 4$ e cero nos demais casos. Tomaremos como categoría de referencia, $killip = 1$.

[2] O GRACE score (Global Registry of Acute Coronary Events) foi desenrolado por Granger et al. (2003), co obxectivo de determinar posibles factores que axuden a predicir a morte dos pacientes diagnosticados de ACS.

Finalmente, o sumando $f_{spat(s_i)} + b_{s_i}$ refire aos efectos espaciais sendo s_i o municipio onde reside o paciente. $f_{spat}(s_i)$, representa os efectos espaciais estruturados e b_{s_i} denota os efectos espaciais non estruturados. Estimando de forma separada estes dous tipos de efectos espaciais, podemos estudar se existen marcadas tendencias espaciais (efectos espaciais estruturados) ou se tamén existen tendencias locais (non estruturadas).

A dificultade do modelo anterior, (4.1), radica en decidir que variables deben ser considerar como efectos non paramétricos e cales como tempo dependentes ou efectos paramétricos. En realidade, non existe unha regra xeral que nos permita tomar esta decisión, senón que debemos basearnos en consideracións teóricas sobre o mecanismo de xeración dos datos (Fahrmeir et al., 2007). Un bo comezo pode ser incluír todas as covariables continuas como efectos non paramétricos e logo ir reducindo a complexidade do modelo. No caso das covariables categóricas, distinguir entre efectos variables co tempo e efectos paramétricos é máis complicado. Para manter a máxima simplicidade do modelo, na medida do posible tan só consideraremos interaccións entre variables ou efectos tempo dependentes se existen fortes indicacións teóricas para supoñer ese comportamento. Ademais, podemos usar o criterio de información de Akaike (AIC) ou o criterio de información baiesiano (BIC) para axudarnos a construír o modelo máis axeitado. Neste caso o AIC do modelo resultante é 15323, e o BIC 15580.7. Non obstante, en xeral, son máis importantes as consideracións teóricas das variables.

Neste exemplo, para formular o modelo (4.1) baseámonos precisamente nas impresións clínicas. Para realizar a estimación do modelo anterior, empregamos o software BayesX. Como interface gráfico utilizamos o software de acceso libre R. De novo, para levar a cabo a inferencia empregaremos técnicas empíricas bayesianas que nos permiten estimar a regresión e os parámetros de suavizado considerando a regresión estruturada de risco como un modelo mixto, de xeito similar ao que realizamos no Capítulo 2 cos modelos STAR xerais. Véxase Fahrmeir et al. (2007).

4.4. Resultados

Na Táboa 4.1, móstranse os resultados obtidos para as covariables categóricas. Pode verse que as covariables categóricas teñen efectos significativos na supervivencia, excepto a variación na elevación do segmento ST durante o infarto.

Variable	γ_i	DT	HR	IC 95 % (HR)	p-valor
*killip1 (Infarto moderado)	0.582	0.127	1.789	$(1.512, 2.112)$	<0.001
*killip2 (Infarto grave ou moi grave)	0.802	0.157	2.230	$(1.743, 2.864)$	<0.001
icp (intervención percutánea coronaria)	-0.517	0.073	0.596	$(0.545, 0.689)$	<0.001
cambioST (variación do segmento ST)	0.109	0.081	1.11	$(0.961, 1.305)$	0.178

Táboa 4.1: Estimación das covariables categóricas do modelo (4.1). e DT, á desviación típica. HR, denota á taxa de risco (Hazard Rate). IC, denota ao intervalo de confianza ao 95 % dos HR. *Tomouse como categoría de referencia $killip = 1$, (Infarto leve).

Este tipo de modelos permítenos estimar conxuntamente a taxa de risco basal e os efectos das covariables. Na Figura 4.2 representamos a estimación do logaritmo da taxa de risco (log-baseline). Obsérvase que nos primeiros días despois de sufrir o infarto, o risco de morte é moi alto, sen ter en conta o resto de covariables. Non obstante, pasados uns días do momento inicial, o risco de morte diminúe. A partir de aí obsérvase unha tendencia crecente suave que aumenta conforme van pasando os anos.

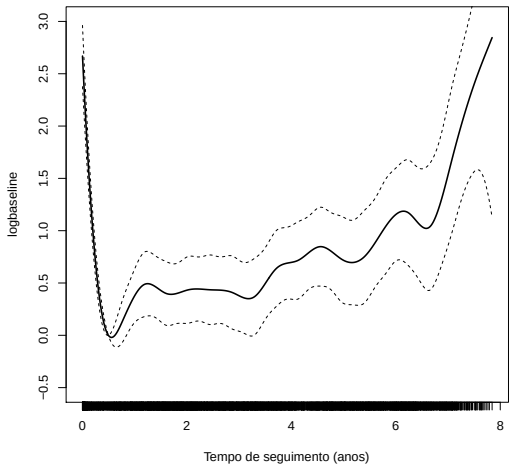


Figura 4.2: Logaritmo da taxa de risco basal (log-baseline) para os pacientes con ACS xunto coa bandas de confianzas puntuais ao 95 %.

Nas Figuras 4.3 e 4.4, representamos os efectos non paramétricos das covariables continuas. Na Táboa 4.2, incluímos os graos de liberdade resultado do axuste destas variables.

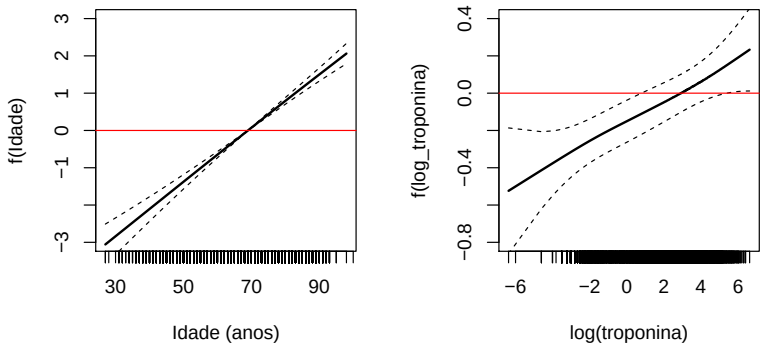


Figura 4.3: Representación do efecto da idade na supervivencia do ACS (esquerda) e do logaritmo dos niveis troponina (dereita) xunto coas bandas de confianza puntuais ao 95 %. A medida que aumenta a idade e o valor da troponina a supervivencia dos pacientes diagnostica de ACS é moito menor.

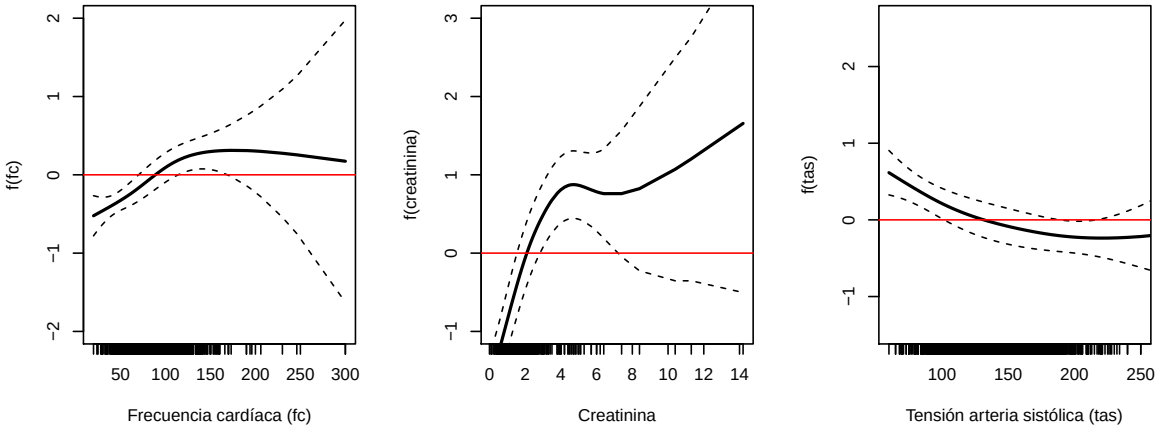


Figura 4.4: Representación dos efectos non paramétricos das covariables que miden a frecuencia cardíaca (*fc*), a presión sanguínea sistólica, *tas*,e a *creatinina*, xunto coas bandas de confianza puntuais ao 95 %.

Efectos continuos	**Graos de liberdade**
Idade	2
Troponina	1.43
fc	2.93
Creatinina	3.59
Tensión arterial sistólica	2.85
Taxa de risco basal,$g_0(t)$	11.28

Táboa 4.2: Graos de liberdade dos efectos non paramétricos estimados.

Na Figura 4.5 representamos a distribución xeográfica do risco de morte por ACS na área sanitaria de Santiago de Compostela. Os colores vermellos indican maior risco de morte por ACS, os colores laranxas e amarelos indican riscos medios altos ou medios (respectivamente). Por outra banda, nas rexións representadas en verde, o risco é menor e polo tanto a supervivencia é maior que no resto das rexións. Tal e como podemos observar a distribución xeográfica do risco de morte de ACS non é uniforme na área sanitaria de Santiago

de Compostela, senón que dependendo do lugar de residencia do paciente, o risco de morte varía. Na Figura 4.6, representamos a significación destes efectos, neste caso os efectos non estruturados non resultan ser significativos.

Na Figura 4.5, podemos ver como na parte este do mapa o risco de morte é menor mentras que na parte suroeste semella moito maior. Este patrón confírmase ao observar os mapas da significación dos efectos espaciais estruturados (Figura 4.6). Ribeira (ao 95 % de significación), Porto do Son, Pobra do Caramiñal, e Outes (80 % de significación) son zonas preocupantes. Pola contra en Touro, Silleda (95 %) e en lugares como Vila de Cruces, Arzúa, Boimorto, e o Pino (80 %) a supervivencia é maior que no resto da área analizada. Por outra banda, os efectos non estruturados non resultaron significativos.

Unha posible explicación desta maior mortalidade por ACS pode deberse ao feito de que nas rexións máis afastadas do hospital de referencia o tempo no que se tarda en realizar unha anxeoplastia é moito maior o que pode ocasionar graves consecuencias na saúde do paciente.

A situación máis crítica dos pacientes con ACS vívese nas primeiras seis horas cando se produce o momento agudo da dor, isto explica a maior mortalidade nas primeiras horas que observabamos na Figura 4.2. A rápida detección dos síntomas é vital nestas primeiras horas. O problema acontece nos pequenos hospitais locais, pois non teñen medios para poder desobstruír as arterias, e cando deciden trasladar aos pacientes ao hospital xeral máis próximo, en ocasións é tarde, e o paciente falece ou aínda que recupere o corazón queda danado, o que explica unha maior mortalidade tamén a longo prazo. Isto é precisamente o que observamos na comarca do Barbanza, onde os pacientes van antes ao hospital de Ribeira que ao de Santiago de Compostela.

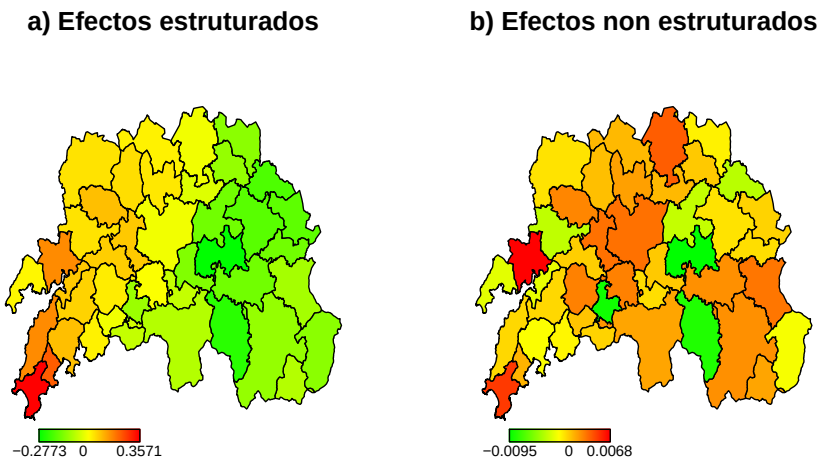


Figura 4.5: Estimación dos efectos espaciais estruturados (a) e non estruturados (b). Os municipios vermellos seguidos dos laranxas indican maior risco de morte dos pacientes diagnosticados de ACS, pola contra os habitantes das zonas representadas de cor verde, sobreviven máis tempo. As rexións coloreadas de amarelo presentan riscos medios.

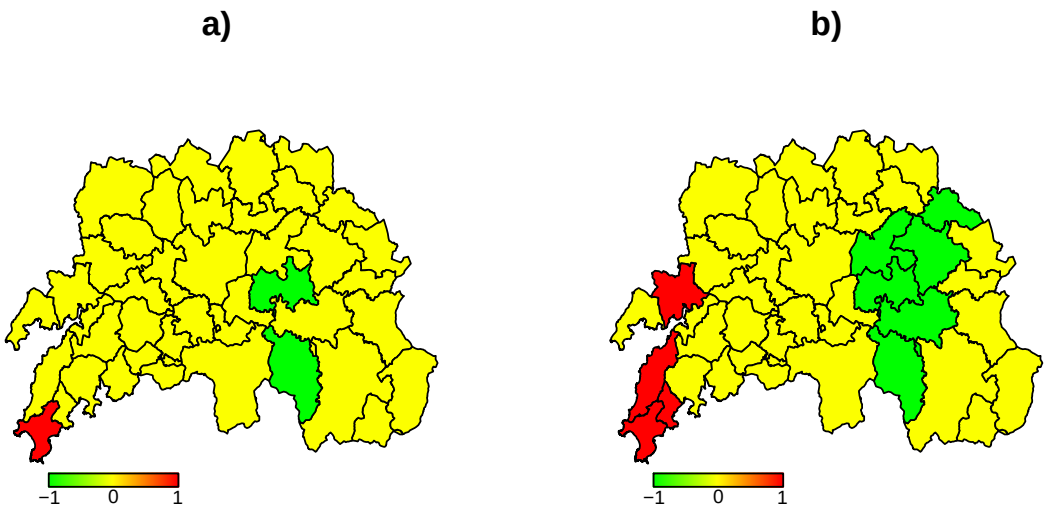


Figura 4.6: Significación dos efectos espaciais estruturados ao 95 % (a) e 80 %(b).

Dende unha perspectiva médica, é de grande interese investigar a evolución desta enfermidade ao longo do tempo e desta forma determinar se existen diferentes patróns xeográficos ao longo do período analizado. Elixiremos como referencia o bienio do 2003-2004, e estudaremos as diferenzas existentes con respecto aos outros bienios: 2005-2006; 2007-2008; 2009-2010. Tal e como podemos observar na Figura 4.7, a supervivencia do ACS parece que está cambiando nos últimos anos (2009-2010).

Nesta evolución, cabe destacar a a situación da comarca do Barbanza podemos observar que entre o 2009 e o 2011, o risco de morte diminúe respecto do 2003. Como comentamos, no momento inicial do estudo a mortalidade na comarca do Barbanza é moito maior que no resto posto que os pacientes ademais de estar a máis distancia do hospital de referencia, antes de ser derivados para este, eran atendidos no hospital de Ribeira, onde non dispoñen dos medios necesarios para tratar aos pacientes con ACS, e polo tanto cada vez era maior o tempo que

tardaban en derivalos ao hospital de Santiago de Compostela. Horas de vital importancia para poder sobrevivir sen secuelas. Non obstante, podemos observar como a medida que pasaron os anos, esta situación mellora. Isto, pode deberse a posta en marcha do Programa *Progaliam* que propugna que todo paciente con dolor precordial debe ser enviado o antes posible ao hospital de referencia. Entrementres, na zona este do mapa prodúcese un aumento relativo da mortalidade, acusado pola mellora na Comarca do Barbanza. Non obstante, esta variación non é significativa, excepto no período 2003-2004 (Figura 4.8).

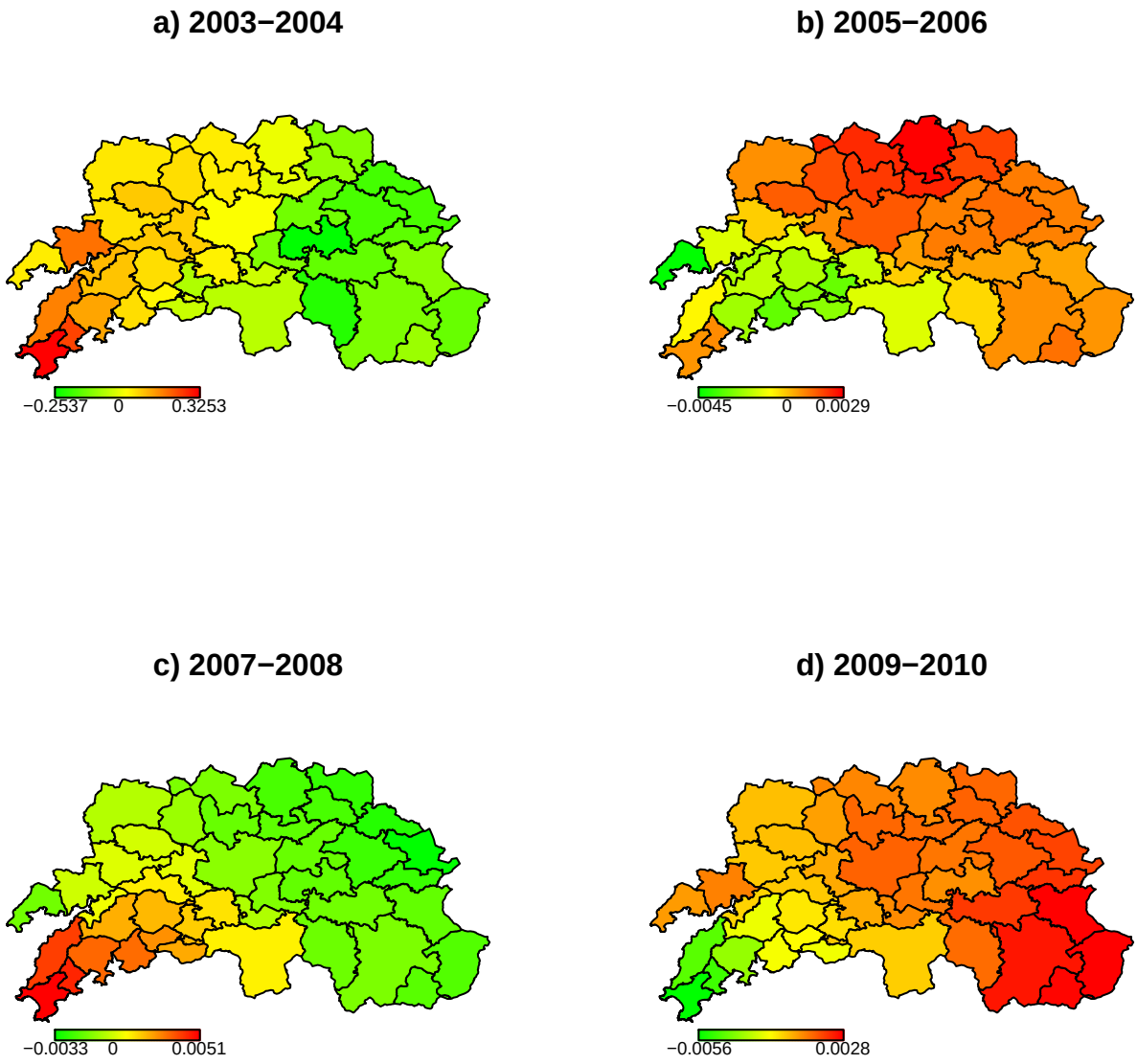


Figura 4.7: Comparación da supervivencia de ACS na área sanitario de Santiago de Compostela respecto do bienio de referencia 2003-2004.

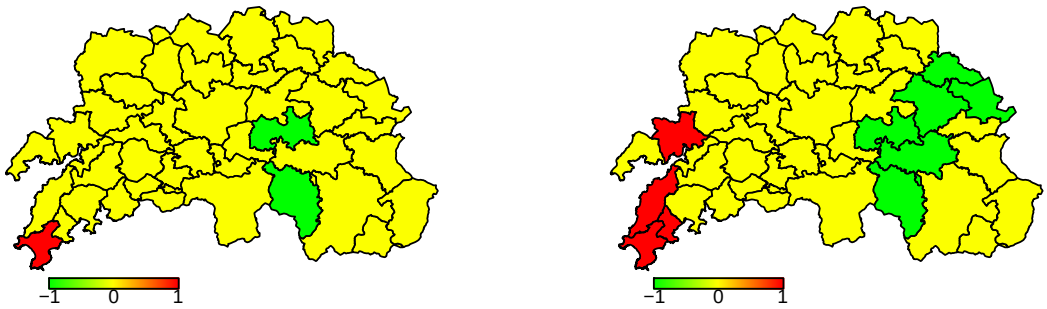


Figura 4.8: Significación ao 95 % dos efectos espaciais no 2003-2004 (esquerda) e ao 80 % (dereita). En vermello represéntanse as significacións positivas e en verde as negativas.

4.5. Capacidade de discriminación do modelo

Finalmente, estudaremos a capacidade de discriminación do modelo de supervivencia presentado en (4.1), baseándonos nas curvas ROC tempo dependentes (Time-Dependent Receiver Operating Charasteristic, ver: Heagerty et al., 2000 e Heagerty e Zheng, 2005) introducidas na Sección 2.5.1 do Capítulo 2.

4.5.1. Curvas ROC tempo dependentes

Como vimos na Sección 2.5.1 do Capítulo 2, empregando a definición de sensibilidade incidente e especificidade dinámica propostas por Heagerty e Zheng (2005), podemos calcular novas curvas ROC (I/D) e desta maneira obter o $AUC(t)$ (Area Under Curve) a partir do

cal calcularemos unha medida global para estimar a concordancia do modelo. Esta área toma valores entre 0.5 e 1. Valores próximos a un indicarán que a capacidade preditiva do modelo é boa, mentres que valores pretos ao 0.5, significan que o modelo non é moi informativo, Heagerty e Zheng (2005). Para realizar estes cálculos empregaremos o paquete de R, *risksetROC*.

Se calculamos o valor do AUC para cada día t de seguimento, podemos ver como este valor se mantén elevado para todo t. (Figura 4.9). Polo que estamos ante un modelo con bastante precisión á hora de realizar predicións.

Na Figura 4.10 representamos a curva $ROC(t)$ en $t = 365$ días para diferentes anos do estudo. Neste caso aprécianse diferenzas para cada un dos bienios considerados.

Na Figura 4.11 representamos $AUC(t)$ separando os datos por diferentes períodos, podemos observar como a capacidade preditiva do modelo varía, de feito mellora co paso dos anos.

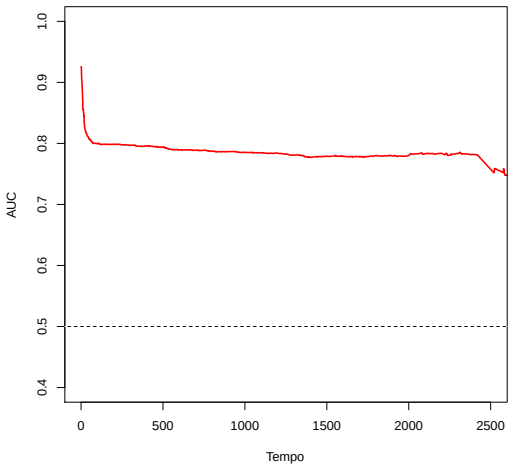


Figura 4.9: Representación dos valores de $AUC(t), t = 1, \ldots, 2500$.

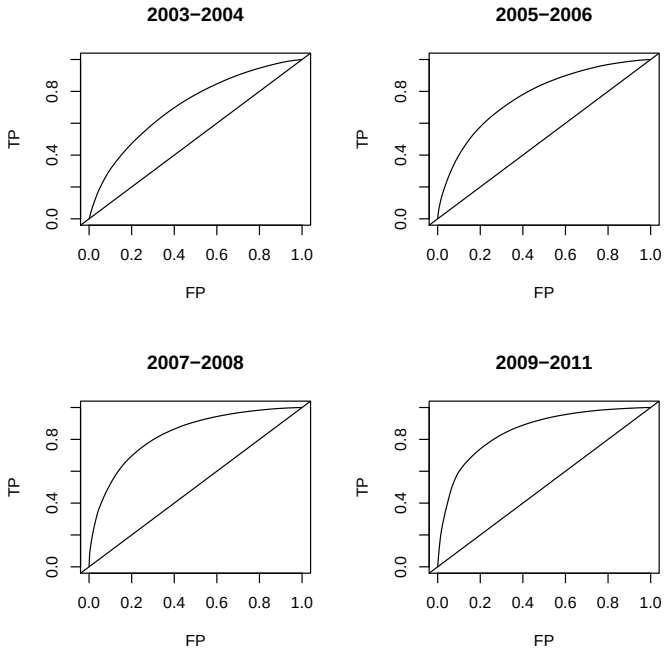


Figura 4.10: Estimación da curva ROC(t) en $t = 365$ días en diferentes períodos do estudo.

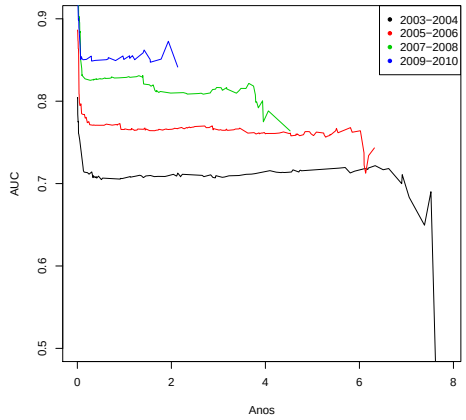


Figura 4.11: Valores de $AUC(t)$ para dos bienios estudados.

4.5.2. Índice C de concordancia

Na sección 2.5.1 do Capítulo 2, presentamos os métodos ROC como ferramentas para determinar a capacidade dun marcador e desta forma distinguir casos e controis nun tempo, t. Pero, ás veces, non se dispón dun t previamente identificado, nestes casos sería de gran utilidade dispoñer dunha medida global de precisión. Neste exemplo, empregaremos o índice C de concordancia definido na Sección 2.5.2 como medida global de resumo. Como xa expusemos, valores próximos a 0.5 indican que o modelo non ten capacidade preditiva, mentres que valores pretos ao 1, indicarán que se realizou unha separación moi boa.

No modelo considerado o valor do índice C é aproximadamente 0.8. Ademais, calculouse o valor do índice C en cada un dos municipios da área sanitaria de Santiago de Compostela. Os concellos aparecen coloreados en vermello, laranxa, amarelo e verde, de maior a menor valor do índice C, respectivamente. Tal e como se amosa no mapa da Figura 4.12, excepto en Val do Dubra ($C = 0.54$), Santiso ($C = 0.63$) e Pontecesuras ($C = 0.52$), coloreados en verde pistacho -todos eles concellos moi pequenos, e polo tanto con inestabilidades-, no resto de municipios da área sanitaria de Santiago de Compostela o índice C acada valores altos, e moi altos naqueles concellos como Rodeiro, Silleda, Boimorto, Frades e Dodro, representados en vermello.

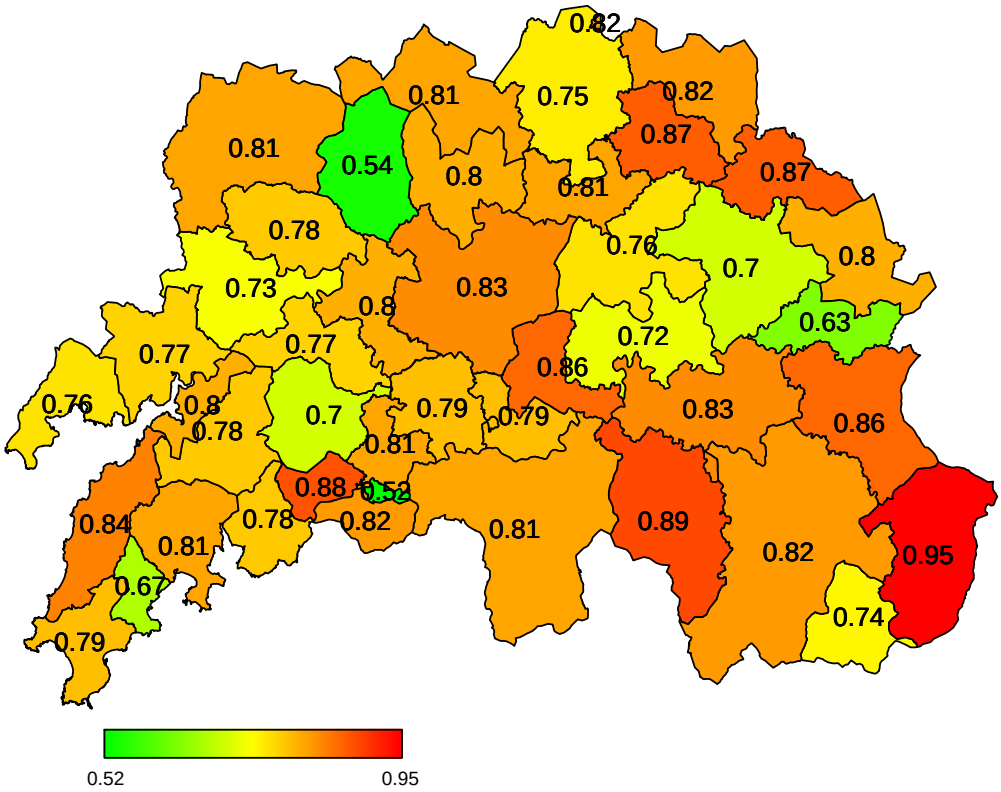


Figura 4.12: Cálculo do índice c en cada municipio da área sanitaria de Santiago de Compostela. O índice c, *c (for concordance) index*, é unha medida común para avaliar a capacidade preditiva dos modelos de supervivencia, incluso cando a variable resposta posúe datos censurados. Os municipios aparecen coloreados en vermello, laranxa, amarelo e verde de maior a menor valor do índice C respectivamente.

Ademais tamén calculamos o índice c, do modelo considerado por períodos:

	2003-2004	**2005-2006**	**2007-2008**	**2009-2011**
Índice c	0.72	0.77	0.82	0.83

Cuadro 4.3: Valor do índice c nos diferentes bienios do estudo.

Segundo os resultados obtidos, o índice C confirma, novamente, que a capacidade preditiva do modelo aumenta a medida que pasan os anos. A mellora da historia clínica dos pacientes grazas á adopción do sistema de historia clínica electrónica, IANUS, e o avance das tecnoloxías médicas, son posibles causas que permiten realizar mellores métodos diagnósticos e desta forma aumentar a supervivencia dos pacientes.

Conclusións

Neste capítulo empregamos os modelos xeoaditivos de supervivencia para estudar as desigualdades xeográficas nos pacientes diagnosticados de ACS. Tal e como vimos, estes modelos permítennos estudar dunha forma flexible os factores que afectan na supervivencia destes pacientes.

A modo de conclusión, poderíamos afirmar que a análise das desigualdades xeográficas é de vital importancia na práctica clínica que permiten poñer en marcha campañas para procurar a equidade do sistema sanitario. No caso do síndrome coronario agudo, existen diferencias xeográficas na supervivencia destes pacientes, e ademais nos últimos anos este patrón está cambiando. Xa para rematar, comprobamos que a capacidade discriminatoria do modelo presentado é boa, e mellora nos últimos anos.

Capítulo 5

Comentarios finais

Este traballo supón unha introdución aos modelos de regresión aditiva estruturada, facendo especial fincapé na súa utilidade, á hora de modelar de forma flexible e unificada os efectos das covariables continuas e os efectos espaciais. Ademais de permitirnos incorporar unha ampla familia de variables resposta no modelo: familia exponencial, respostas categóricas, tempos de supervivencia incluíndo datos censurados.

A versatilidade deste tipo de modelos mostrouse a través de dúas grandes aplicacións biomédicas. En primeiro lugar, investigáronse as tendencias espaciais na taxa de abstinencia ao alcohol empregando un modelo STAR cunha resposta de Poisson. A análise mostra as desigualdades xeográficas existentes na distribución destas taxas en Galicia, así como a súa relación con varios factores sociodemográficos incluídos no estudo.

Unha potencialidade importante dos modelos STAR é que permite considerar aos modelos de supervivencia con datos censurados, como un caso particular. No presente traballo estudouse de forma flexible os procesos de supervivencia do síndrome coronario agudo na área sanitaria de Santiago de Compostela, incluíndo, ademais de covariables clínicas do paciente, covariables espaciais e temporais. Obxectivouse que, efectivamente, existen diferenzas xeográficas na supervivencia destes pacientes, e ademais nos últimos anos este patrón está cambiando. Ademais, comprobouse que a capacidade de discriminación do modelo era moi boa mediante o uso de curvas ROC tempo-dependentes e o índice C de concordancia.

As vantaxes prácticas que se introducen ca utilización dos modelos STAR na investigación

biomédica son múltiples. Non obstante estes modelos tamén presentan limitacións. Entre as que destacamos:

1. A regresión clásica adoitase formalizar como a media condicionada da resposta en función dos valores das variables explicativas, e tamén nos modelos STAR. Pero, centrarse tan só en estimar medias pode ocasionar erros cando modelamos datos de estruturas complexas. En moitas aplicacións biomédicas, por exemplo, non só interesa explicar o efecto das covariables en función da media da resposta senón que necesitamos coñecer a distribución completa. Este tema está a ser resolto resolto cos modelos GAM para localización, escala e forma (GAMLSS, Rigby e Stasinopoulos, 2005), onde se asumen como respostas distribucións paramétricas complexas. Sin embargo, neste ámbito GAMLSS non están implementados posibles efectos espaciais na resposta.

2. En ocasións, tamén se precisa modelar respostas multivariantes. Na maioría das investigacións publicadas sobre regresión para respostas multivariantes asúmese unha distribución específica para a variable resposta sen motivo aparente, e son escasas as contribucións empregando preditores non paramétricos.

   A regresión distribucional multivariante (Klein et al, 2014) permítenos solventar este problema. Esta regresión distribucional supón unha xeneralización dos GAMLSS multivariantes xa que os parámetros da distribución da resposta non sempre están relacionados coa localización, a escala ou a forma, senón que dependen de funcións máis complexas (Klein et al., 2014). Ademais, esta regresión permitiranos introducir de forma sinxela, efectos de tipo espacial e/ou temporal. O estudo metodolóxico deste tipo de modelos será precisamente unha liña de investigación futura, que se converterá na tese de doutoramento da autora deste Traballo Fin de Máster.

3. Nos últimos tempos os datos funcionais están a ser considerados en multitude de aplicación biomédicas, non obstante os efectos de covariables funcionais aínda nos están incorporados na regresión distribucional, e nos modelos STAR en particular.

Esta liña de investigación é nova e foi motivada polo seguinte problema médico. Como resultado de varios proxectos previos en colaboración coa Unidad de Epidemioloxía Clínica do Hospital de Santiago, disponse dunha ampla mostra da poboación xeral adulta, cunha extensa fenotipación e documentación individual, cunha almacenaxe reglada de mostras biolóxicas e recollida da inxesta dietética durante unha semana, que nos permitirá aproximar

varios retos actuais no campo das enfermidades crónicas relacionadas co envellecemento da poboación. Contamos co valor engadido de que se monitorizou a glucosa intersticial durante unha semana, mediante dispositivos de monitorización continua. Deste xeito, dispoñemos dos perfís de glucosa de cada individuo como dato funcional. Na tese de doutoramento da autora deste traballo, proporanse modelos de regresión distribucional que aplicaremos a base de datos presentada, incorporando o estudo de datos funcionais. A estatística funcional (Ramsay e Silverman, 1997) será útil para incorporar os perfís de glucosa nos modelos de regresión de interese.

Recentemente, McLean et al. (2012) propuxeron os FGAM (Functional GAM), que supoñen unha extensión dos GAM aditivos ao campo funcional. Pero, según o noso coñecemento, ata o momento as contribucións á regresión STAR con datos funcionais son escasas, non existindo aínda no ámbito da regresión distribucional. Na futura tese de doutoramento da autora preténdese cubrir este baleiro, tratando de propoñer novos modelos de regresión distribucional multivariante con datos funcionais, que suporán unha ferramenta estatística de grande aplicabilidade biomédica.

Bibliografía

[1] Biomarkers Definition Working Group. (2001). *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. Vol. 69, pp: 89-95.*

[2] Breslow, N. E. and Clayton, D. G. (1993). *Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. Vol. 88 (421), pp: 9-25.*

[3] Brezger, A., Kneib, T., e Lang, S. (2005). *BayesX: Analyzing Bayesian structured additive regression models. Journal of Statistical Software. Vol. 14 (11).*

[4] Cadarso-Suarez, C., Meira-Machado, L., Kneib, T. e Gude, F. (2010). *Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data. Statistical Modelling. Vol. 10, pp: 291-314.*

[5] Cox, D.R. (1992). *Regression models and life tables (with discussion). Journal of the Royal Statistical Society. Series B. Vol. 34, pp: 187-220.*

[6] Cuarta edición do Manual DSM-IV diagnóstico e estatístico dos transtornos mentais. Dispoñible en http://www.mdp.edu.ar/psicologia/cendoc/archivos/Dsm-IV.Castellano.1995.pdf. Consultado en 1/05/2015.

[7] Dierckx, P. (1993). *Curve and surface fitting with splines. Oxford: Oxford University Press.*

[8] Eliers, P. H. C., e Marx, B. D. (1996). *Flexible smoothing using B-Splines and penalties (with comments and rejoinder). Stadistic Science. Vol. 11, pp: 89-121.*

[9] Enquisa Nacional de Saúde. España 2011/12. *Serie Informes monográficos n^o 1. Consumo de alcohol. Ministerio de Sanidad, Servicios Sociales e Igualdad. Madrid 2013.*

[10] Fahrmeir, L. e Lang, S. (2001). *Bayesian inference for generalized addictive mixed models based on Markov random field priors. Journal of the Royal Statistical Society: Series C. Vol. 50, pp: 715-745.*

[11] Fahrmeir, L., Kneib, T., e Brezger, A. (2005). *Bayesian semiparametric regression based on MCMC techniques: A tutorial.* Consultado o 12/05/2015 en http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.6326&rep=rep1&type=pdf.

[12] Fahrmeir, L., Kneib, T., Lang S. e Marx, B. (2013). *Regression. Models, methods and Applications. Heidelberg, Berlin: Springer.*

[13] Fahrmeir, L., Kneib, T., e Lang, S. (2004). *Penalized structured additive regression for space-time data: a bayesian perspective. Statistica Sinica. Vol. 14, pp: 715-745.*

[14] Fahrmeir, L. e Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models. New York: Springer-Verlag.*

[15] Gonzalez-Quintela, A., Fernandez-Conde, S., Alves, M.T., Campos, J., López-Raton, M., Puerta, R., Monte, R. e Gude, F. (2011). *Temporal and spatial patterns in the rate of alcohol withdrawal syndrome in a defined community. El Sevier. Vol. 45, pp: 105-111.*

[16] Granger C.B., Goldberg R.J., Dabbous O., Pieper K.S., Eagle K.A., Cannon C.P., Van de Werf F., Avezum Á., Goodman S.G., Flather M.D. e Fox K.A.A. (2003). *Global Registry of Acute Coronary Events Investigators. Predictors of hospital mortality in the Global Registry of Acute Coronary Events. Arch Intern Med. Vol. 163, pp: 2345-53.*

[17] Green, P.J. e Silverman, B.W. (1993). *Non parametric Regression and Generalized Linear Models. Chapman and Hall/CRC.*

[18] Harrell, F., Lee K. e Mark D. (1996). *Tutorial in bioestadistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stadistics in medicine. Vol. 15, pp: 361-387.*

[19] Hastie, T.J. e Tibshirani, R.J. (1993). *Varying-coefficient models. Journal Royal Statistical Society: Series B. Vol. 55, pp: 757-796.*

[20] Hastie, T.J. e Tibshirani, R.J. (1990). *Generalized Additive Models. London: Chapman-Hall.*

[21] Heagerty, P.J., Lumley T. e Pepe M.S. (2000). *Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics. Vol. 56, pp: 337-344.*

[22] Heagerty, P.J. e Zheng, Y. (2005). *Survival Model Predictive Accuracy and ROC Curves. Biometrics. Vol. 61, pp: 92-105.*

[23] Hennerfeind, A., Brezger, A. e Fharmeir, L. (2005). *Geoaddictive Survival Models.* Dispoñible na dirección web: http://epub.ub.uni-muenchen.de/1783/1/paper-414.pdf. (Consultado o 26/03/2015).

[24] Killip T., Kimball, J.T. (1967). *Treatment of myocardial infarction in a coronary care unit. A two year experience with 250 patients. American Journal of Cardiology. Vol. 20, pp: 457-64.*

[25] Klein, N. Kneib T., Klasen S., Lang S. (2014). *Bayesian Structured Additive Distributional Regression for Multivariate Responses. Journal of the Royal Statistical Society: Series C, DOI:10.1111/rssc.12090.*

[26] Kneib T., Fharmeir, L. (2007). *Mixed model approach for geoadditive hazard regression. Scandinavian Journal of Statistics. Vol. 34, pp: 207-228.*

[27] Kneib, T. (2005). *Mixed model based inference in structured additive regression.* University of München. Dispoñible en http://edoc.ub.uni-muenchen.de/5011/1/Kneib-Thomas.pdf.

[28] Mateos, R., Páramo, M., Carrera, I., e Rodríguez-López, A. (2002). *Alcohol consumption in a southern European region (Galicia, Spain). Subst. Use Misuse. Vol. 37, pp: 1957-1976.*

[29] McCullagh, P. e Nelder, J.A. (1989). *Generalized Linear Models. New York/Boca Raton: Chapman-Hall.*

[30] McLean M.W., Hooler G., Staicu A.M., Scheipl F., Ruppert D. (2012). *Functional Generalized Additive Models. Journal of Computational and Graphical Statistics. Vol. 23 (1), pp: 249-269.*

[31] Monte Secades, R. e Rabuñal Rey, R. (2011). *Guía de práctica clínica: Tratamiento del síndrome de abstinencia alcohólica, 2ª edición. Galicia Clínica. Vol. 72 (2), pp: 51-53.*

[32] O'Connor R.E., Brady W., Brooks S.C., Diercks D. et al. (2010). *Part 10: acute coronary syndromes: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Circulation. Vol. 122 (suppl 3), pp: S787-S817.*

[33] Organización Mundial da Saúde (2004). *European health for all database (HFABD). Global status report on alcohol 2004.* Dispoñible en: http://www.who.int/substance-abuse/publications/statusreportalcoholeuro/en/index.html. Consultado o 2/05/2015.

[34] Prieto-Lara, E. and Ocaña-Riola, R. (2010). *Updating Rurality Index for Small Areas in Spain. Social Indicators Research. Vol. 95 (2), pp: 267-280.*

[35] Rice, J.A. e Wu, C.O. (2001). *Non parametric mixed effects models for unequally sampled noisy curves. Biometrics. Vol. 57, pp: 253-259.*

[36] Rigby R.A., Stasinopoulos, D.M. (2005). *Generalized additive models for location, scale and shape (with discussion). Applied Statistics, Vol. 54, pp: 507-554.*

[37] Rue, H. e Held, L. (2005). *Gaussian Markov Random Fields. Chapman & Hall/CRC: Boca Raton, FL.*

[38] Ruppert, D. Wand, M.P. e Carroll, R.J. (2003). *Semiparametric Regression. University Press. Cambridge.*

[39] Wand, M.P. e Jones, M.C. (1995). *Kernel Smoothing. Boca Raton, FL. Chapman and Hall.*

[40] Wood, S.N. (2006). *Generalized Additive Models: An introduction with R. Boca Raton, FL. Chapman-Hall.*

[41] Wood, S. N. (2003). *Thin Plate regression splines. Journal of the Royal Stadistical Society B. Vol 65, pp: 95-114.*

[42] Wood, S.N. e Tibshirani, R.J. (1990). *Generalized Additive Models. London, UK: Chapman-Hall.*