Trabajo Fin de Máster

Modelos de regresión de resposta multivariante e a súa aplicación a datos biomédicos

Martín Lema Pailos

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Modelos de regresión de resposta multivariante e a súa aplicación a datos biomédicos

Título en español: Modelos de regresión de respuesta multivariante y su aplicacin a datos biomédicos

English title: Multivariate response regression models and their application to biomedical data

Modalidad: Modalidad B

Autor/a: Martín Lema Pailos , Universidade da Coruña

Director/a: César Andrés Sánchez Sellero,, Universidade de Santiago de Compostela.

Tutor/a: María Pata, Biostatech Advice Training & Innovation in Biostatistics, S.L.

Breve resumen del trabajo:

En el análisis estadístico de datos biomédicos, es muy habitual encontrarse ante una gran cantidad de información correlacionada. En los últimos años han surgido varias técnicas de modelización conjunta (Joint Modelling, JM), muy diferentes en función del tipo de variables respuesta con las que se está trabajando, y que permiten no solo estudiar a la vez qué factores pueden estar afectando a dos o más variables de interés, sino también la relación que pueda existir entre ellas. El objetivo de este trabajo es el estudio y aplicación de técnicas JM a bases de datos reales en Biomedicina..

Recomendaciones:

Otras observaciones:

Don César Andrés Sánchez Sellero,, de la Universidade de Santiago de Compostela.y doña María Pata, de Biostatech Advice Training & Innovation in Biostatistics, S.L. , y don/doña , de , informan que el Trabajo Fin de Máster titulado

Modelos de regresión de resposta multivariante e a súa aplicación a datos biomédicos

fue realizado bajo su dirección por don Martín Lema Pailos para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 13 de xuño de 2022.

El director:
Don César Andrés Sánchez Sellero,

La tutora:
Doña María Pata

El autor:
Don Martín Lema Pailos

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumo vii

Prefacio ix

1. Base de datos 1
1.1. Descripción do estudo . . . 1
1.1.1. Descripción das variables . . . 1
1.2. Algunhas propiedades das variables . . . 3

2. Modelos lineais xeralizados 5
2.1. Modelos lineais . . . 5
2.2. Modelos lineais xeralizados . . . 6

3. Modelos aditivos xeralizados 11
3.1. Suavizado univariante . . . 11
3.1.1. Suavizadores . . . 12
3.2. Modelos aditivos xeralizados . . . 20
3.3. Aplicación a datos biomédicos . . . 22
3.3.1. Buscamos a mellor distribución para a variable resposta . . . 22
3.3.2. Introducimos as variables explicativas idade e sexo . . . 23
3.3.3. Incorporamos as variables sociodemográficas . . . 25
3.3.4. Sumamos agora o efecto da diabetes . . . 28
3.3.5. Variables antropomórficas . . . 29
3.3.6. Incluimos as variables sobre a glucosa e a hemoglobina . . . 31
3.3.7. Resumo dos modelos e discusión final . . . 31

4. Modelos aditivos xeralizados de localización, escala e forma 43
4.1. GAMLSS . . . 43
4.1.1. Os seus inicios: modelar o parámetro de escala . . . 43
4.1.2. Modelo aditivos xeralizados de localización, escala e forma (GAMLSS). . . . 43
4.2. Aplicación a datos biomédicos . . . 45

5. Modelos Joint Modelling 57
5.1. Modelos de regresión de cópulas bivariadas . . . 57
5.1.1. Cópulas . . . 57
5.2. Formulación do modelo . . . 62
5.3. Inferencia . . . 64
5.3.1. Inferencia bayesiana . . . 64
5.3.2. Inferencia por máxima verosimilitude penalizada . . . 64
5.4. Aplicación a datos biomédicos . . . 65
5.4.1. Distribucións marxinais . . . 65

5.4.2. Selección da función cópula 66
5.4.3. Axuste do modelo CGAMLSS 67
5.4.4. Resultados 69

Bibliografía **73**

Resumo

Resumo en galego

Neste traballo aplicamos os CGAMLSS (*bivariate copula generalised additive models for location, scale and shape*) coa finalidade de explicar a presión arterial sistólica e diastólica a través dos datos proporcionados polo SERGAS da poboación galega. Realizamos un repaso dos modelos lineais e dos modelos lineais xeralizados (GLM) e empregamos bases de datos adecuadas que nos permitan ver as limitacións deste tipo de modelos. Para paliar estas carencias introduciuse o concepto de función suavizadora univariante xunto co concepto de bases de Splines. Este novos conceptos permitírannos pasar aos modelos aditivos xeralizados (GAM) e modelos aditivos xeralizados de localización, escala e forma (GAMLSS). Vimos tamén os graos de liberdade efectivos destes modelos así como os algoritmos de estimación dos seus coeficientes. Con este novos modelos fixéronse axustes para a base de datos presentada para tratar de explicar a presión arterial sistólica e diastólica. Por último introduciremos o concepto de cópula e algún exemplos das mesmas que nos permitirá presentar os modelos finais do traballo, os CGAMLSS, xunto con algún comentario para realizar inferencia dos mesmos. Remataremos tratando de explicar a relación entre as dúas presións estudadas antes empregando este tipo de modelos.

English abstract

In this work we apply the CGAMLSS (textit bivariate copula generalized additive models for location, scale and shape) in order to explain systolic and diastolic blood pressure through data provided by SERGAS of the Galician population. We review linear models and generalized linear models (GLMs) and use appropriate databases that allow us to see the limitations of such models. To alleviate these shortcomings the concept of univariate softening function was introduced along with the basic concept of Splines. These new concepts will allow us to move to generalized additive models (GAM) and generalized additive models of location, scale, and shape (GAMLSS). We also looked at the effective degrees of freedom of these models as well as the algorithms for estimating their coefficients. With this new model, adjustments were made to the database presented to try to explain systolic and diastolic blood pressure. Finally, we will introduce the concept of copulation and some examples of them that will allow us to present the final models of the work, the CGAMLSS, along with some comments to make an inference from them. We will end by trying to explain the relationship between the two pressures studied before using this type of model.

Prefacio

Tradicionalmente, os modelos de regresión representan a dependencia existente entre unha variable resposta en función dun conxunto de variables predictoras (que habitualmente denominamos como covariables, variables regresoras ou variables independentes). Este campo da estatística converteuse nun dos máis importantes e conta cun gran número de técnicas e modelos ampliamente empregados en diferentes ámbitos. Permítenos construír modelos para explicar as posibles relacións entre a variable resposta e as distintas covariables. O obxectivo final deste modelos é predicir ou estimar o valor dunha variable tendo en conta o valor dun conxunto de covariables coñecidas.

As técnicas de regresión empréganse ampliamente en múltiples disciplinas (entre as que se inclúen a bioloxía, as ciencias ambientais, as ciencias socias ou economía) pero podemos destacar que se trata dunha ferramenta moi útil para estudos biomédicos, como por exemplo, para estudar factores de risco, para explorar patróns de pronóstico ou derivar predicción de pacientes entre outras moitas posibilidades. Como proba disto están as numerosas publicacións que podemos atopar facilmente se introducimos as palabras clave "*modelo de regresión*" en calquera buscador de artigos científicos.

O primeiro modelo de regresión foi proposto a finais do século XIX. Dende ese momento converteuse nunha área de estudo moi activa e en constante cambio tanto a nivel teórico como a nivel empírico. Todos os avances que se fixeron neste campo tratan de aumentar a flexibilidade dos modelos, tanto para o caso das variables resposta como nos efectos das covariables.

Supoñamos que temos un conxunto de observacións $\{y_i,\, i = 1, \ldots, n\}$ dunha variable resposta xunto co valor das covariables $\{x_{1i}, \ldots, x_{mi}\}$ que conforman entre todas a información dispoñible sobre o individuo i. O modelo de regresión máis simple existente é aquel que supón que o efecto das covariables é lineal, dando lugar así ao seguinte modelo:

$$y_i = \beta_0 + \beta_1\, x_{1i} + \cdots + \beta_m\, x_{mi} + \varepsilon_i$$

onde β_j, $j = 0, \ldots, m$ son os coeficientes de regresión do modelo (descoñecido) que deben ser estimados e $\varepsilon_i \sim N(0, 1)$. Neste tipo de modelos á hora de facer inferencia consideramos que a variable resposta segue unha distribución normal. Vemos así claramente unha necesidade de ampliar os modelos a un caso máis xeral no que a variable resposta non teña porque ser normal e poida ser binaria, categórica ou outra distribución continua distinta da normal. Por este motivo xurdiron o que se coñece como Modelos Lineais Xeralizados (GLM que se poden consultar en ([21]) e ([20])).

Os modelos de regresión GLM permiten diferentes respostas ademais da normal e un grado de non linealidade na estrutura do modelo. Neste tipo de modelos o valor esperado (condicional), $\mathbb{E}[Y_i|X_i] = \mu_i$, está vinculado a un predictor lineal do seguinte xeito:

$$\eta_{i=g(\mu_i)} = \beta_0 + \beta_1\, x_{1i} + \cdots + \beta_m\, x_{mi} \tag{1}$$

onde g é unha función enlace (ou función *link*) que asegura que as restricións do espazo de parámetros se manteñen. Nos modelos GLM asúmese que a variable resposta pertence a unha distribución da familia exponencial (por exemplo a Poisson, binomial, Gamma ou a distribución normal, entre outros). Podemos observar que os modelos lineais non son máis que un caso particular dos modelos lineais xeralizados.

Ata agora flexibilizamos as posibilidades da variable resposta pero todos estes modelos seguen tendo unha gran limitiación: os efectos das covariables consideradas son unicamente lineais, o cal non se vai

axustar ben para moitos dos modelos que pretenderemos axustar en distintos ámbitos como pode ser a medicina. Por este motivo, dende comezos da década dos 90 aparecen os Modelos Aditivos Xeralizados (GAM, consultar ([12])) para solucionar estes problemas. Nos modelos de regresión GAM o predictor lineal da ecuación ([1]) reemplázase por un predictor aditivo como indicamos:

$$\eta_i = g(\mu_i) = \beta_0 + f_{_{1(x_{1i})+\cdots+f_m(x_{mi})}} (2)$$

onde β_0 é o intercepto global e f_j, $j = 1, \ldots, m$ son as función de suavizado descoñecidas que deben ser estimadas. Deste xeito os modelos GAM permiten unha maior variedade de efectos das covariables que sexan continuas e tamén permiten introducir os efectos vistos con anterioridade para o caso de variables binarias, categóricas. Tamén permite introducir interaccións entre as distintas variables.

De todas formas, a pesar de que os GAM son moito máis flexibles que os GLM, aínda teñen algunhas limitacións pois unicamente nos permiten modelar o valor da media da variable resposta e non o resto de parámetros (como por exemplo o parámetro de dispersión no caso dunha Gamma). Para traballar tamén modelando outros parámetros da distribución xorden o que coñecemos como Modelos de Regresión de Localización, Escala e Forma (GAMLSS). Nestes modelos suponse que a variable resposta, Y_i, segue unha distribución paramétrica que non ten por que estar dentro da familia exponencial. Ademais, cada parámetro da distribución da resposta, esta explicado mediante predictores aditivos. Temos así que $Y_i \sim EF(\mu_1, \phi_i, \nu_i, \tau_i)$ onde os parámetros da distribución da resposta poden ser expresados en función das variables explicativas. Seguindo ([29]) podemos definir os modelos GAMLSS como segue:

$$\begin{aligned}
\eta^{\mu} &= g_1(\mu) = \beta_0^{\mu} + f_1^{\mu}(x_{1i}) + \cdots + f_m^{\mu}(x_{mi}), \\
\eta^{\phi} &= g_2(\phi) = \beta_0^{\phi} + f_1^{\phi}(x_{1i}) + \cdots + f_m^{\phi}(x_{mi}), \\
\eta^{\nu} &= g_3(\nu) = \beta_0^{\nu} + f_1^{\nu}(x_{1i}) + \cdots + f_m^{\nu}(x_{mi}), \\
\eta^{\tau} &= g_2(\tau) = \beta_0^{\tau} + f_1^{\tau}(x_{1i}) + \cdots + f_m^{\tau}(x_{mi})
\end{aligned} \tag{3}$$

onde a primeira ecuación de ([3]) se refire ao parámetro de localizacion (μ) de y_i, a segunda ao parámetro de escala (ϕ) e as dúas últimas aos parámetros de forma (ν e τ).

A pesar de toda esta flexibilidade que podemos ter con estes últimos modelos, sempre son modelos univariantes no sentido de que unicamente se considera unha variable resposta. Habitualmente en aplicacións biomédicas e preciso modelar de forma conxunta dúas ou máis variables resposta para determinar a relización existente entre elas. Porén, na maioría dos casos de modelos de regresión con resposta multivariada é necesario asumir unha distribución para as variables resposta sen ningunha razón aparente producindo modelos pouco flexibles. Por este motivo, recentemente os métodos de regresión para respostas bivariadas empregando funcións cópula estánse desenvolvendo moi rapidamente. A maior vantaxe que nos permiten estes modelos é que as distribucións marxinais das variables resposta poden ser diferentes.

Deste xeito, modelar a dependencia entre variables resposta empregadno funcións cópula volveuse moi popular nos últimos anos como unha forma de modelización multivariable en distintos ámbitos. Esta metodoloxía volveuse moi útil especialemente para o campo da mediciña, como por exemplo se verá neste traballo para modelar a relación existente entre a presión arterial sistólica e diastólica. Pódense ver estes modelos dende un punto de vista frecuentista ([19]) ou dende un punto de vista bayesiano ([17])).

Este traballo estará dividido en cinco capítulos. No primeiro deles describiremos a base de datos que se empregará para axustar todos os modelos que se irán presentando posteriormente indicando as técnicas empregadas na mostraxe de forma detallada. Os datos foron recollidos polo Servizo Galego de Saúde (SERGAS) en 2004 e constan de máis de 2500 observacións xunto cun total de máis de 20 variables aínda que nós imos seleccionar 17 para os nosos modelos. Describiremos tamén o significado de cada unha das variables que imos considerar destacando a presión arterial sistólica (PAS) e a presión arterial diastólica (PAD) que serán as variables resposta de todos os modelos axustados.

Unha vez presentada a base de datos, no capítulo dous presentaremos de forma breve e introductoria os primeiros modelos de regresión existentes, os modelos lineais xunto coa súa correspondente

extensión, os modelos lineais xeralizados. Explicaremos os algoritmos de aproximación empregados para a estimación dos seus coeficientes que nos serán de gran utilidade para comprender os modelos máis complexos que se presentarán posteriormente. Para rematar este capítulo veremos algunhas das limitacións destes modelos que xustificarán a necesidade de buscar outros modelos.

No capítulo tres pasaremos a falar dos modelos aditivos e os modelos aditivos xeralizados, que non serán máis que unha modificación dos modelos lineais e lineais xeralizados nos que ahora permitiremos que as variables explicativas teñan efectos non paramétricos sobre a variable resposta. Para definir estes modelos comenzamos describrindo o caso do suavizado univariante e introducindo novos conceptos como os splines (e definiremos distintos tipos de funcións de suavizado) e as bases de splines (entre as que destacaremos os B-Splines e os P-splines). Veremos como os distintos elementos da regresión empregando splines (os nodos, o parámetro de suavizado e a base empregada) afecta ao axuste realizado. Tamén explicaremos como obter os graos de liberdade para estes modelos e o algoritmo P-IRLS que se emprega para a aproximación dos seus coeficientes (non é máis que unha adaptación do algoritmo IRLS que se emprega nos modelos GLM). Unha das claves destes novos modelos e a aparición do parámetro de suavizado (λ) para regular a rugosidade das nosas funcións de suavizado, veremos distintos criterios de selección automática como poden ser o Criterio de Validación Cruzada Xeralizado (GCV), o UBRE ou a Máxima Verosimilitude Restrinxida (RMLE). Unha vez visto o caso univariante facilmente se poden extender todos os conceptos para o caso multivariante dando lugar aos modelos aditivos e modelos aditivos xeralizados (GAM), tanto na súa presentación e estrutura como os seus graos de liberdade, algoritmos de aproximación e criterios de selección do parámetro de suavizado. Para rematar o capítulo axustaranse distintos modelos á base de datos presentada anteriormente e discutiranse os resultados e conclusións obtidas.

No capítulo catro presentaránse os modelos aditivos xeralizados de localización, escala e forma (GAMLSS). Mentres que os modelos GAM unicamente nos permiten modelar o parámetro de localización (μ) da distribución que segue a nosa variable resposta, con estes novos modelos poderemos modelar tamén o parámetro de escala (ϕ) e os parámetros de forma (ν e τ) permitindo así un maior número de distribucións para a variable resposta e unha maior flexibiliade (pois antes todos estes parámetros se consideraban fixos). Comezaremos polos seus inicos nos cales unicamente se considerada o parámetro de escala (ϕ) ademais de o xa visto de localización para posteriormente extender isto aos catro parámetros mencionados. Presentaremos tamén brevemente o axuste do modelo entre o que destacaremos dous algoritmos de aproximación dos coeficientes (o algoritmo CG e o algoritmo RS) que nos permitirán maximizar a función de verosimilitude penalizada. Unha vez presentados e explicados os modelos procedemos de novo a ver a súa aplicación a datos biomédicos axustando distintos modelos xunto coas súas correspondentes interpretacións, discusións e conclusións.

Para rematar, no capítulo cinco presentaremos os modelos que buscamos inicialmente no noso traballo, os Joint Modelling. No noso caso centrarémonos nos CGAMLSS (*bivariate copula generalised additive models for location, scale and shape*), que non son máis que modelos que pretenden modelar a distribución conxunta dun par de variables respota (y_1,y_2) dadas un conxunto de covariables e unha función cópula que especifique a estrutura de dependencia entre as dúas variables resposta. Para isto comezaremos presentando a definición de función cópula bivariada xunto con algúns exemplos destacados deste tipo de funcións cun único parámetro. Veremos graficamente como se estrutura a relación de dependencia entre as variables para cada unha delas e como varía esta relación en función do valor do parámetro da cópula. Unha vez visto isto formularemos o modelo que posteriormente axustaremos e veremos como facer inferencia sobre el, tanto dende un punto de vista bayesiano como dende o punto de vista frecuentista. Para rematar axustaremos un modelo para os nosos datos xunto coas representacións gráficas correspondentes e as conclusións que se poidan sacar de dito axuste.

Capítulo 1

Base de datos

Neste primeiro capítulo introduciremos a base de datos que imos empregar ao longo de todo o traballo para o axuste de modelos GAM, modelos GAMLSS e CGAMLSS. Comezaremos presentando a base de datos xunto cunha descripción do estudo realizado para obtela (extraendo dita descripción de ([24])) e as variables que se inclúen nela. Para rematar faremos un pequeno estudo dalgunhas das variables da base de datos para así ver as súas propiedades.

1.1. Descripción do estudo

Esta base provén do Servizo Galego de Saúde que cubre a máis do 95 % da poboación de Galicia. O estudo co que se obtiveron estes datos foi levado a cabo entre marzo e xuño de 2004 e considerouse unha mostra aleatoria representativa da poboación adulta galega (maiores de 18 anos).

A mostra seleccionouse mediante un procedemento de mostraxe por conglomerados en dous pasos. Inicialmente seleccionouse poboación dependente de centros de atención primaria seleccionados aleatoriamente en cada unha das catro provincias galegas (neste tipo de mostraxe, a poboación de cada unha das provincias considerouse como unha poboación independente). Dentro de cada centro seleccionouse aleatoriamente suxeitos maiores de 18 anos.

Tivéronse en conta factores culturais e socioeconómicos para a selección dos conglomerados, con centros de saúde estratificados polo municipio, a demografía e a tipoloxía. Na selección individual, tívose en conta o sexo e a idade do individuo. Quedaron excluídas deste estudo as persoas xestantes. O protocolo do estudo foi aprobado polo Comité Ético de Investigación Clínica de Galicia. Todos os participantes firmarmon o seu consentimento informado.

Todos os suxeitos foron contactados por correo para concretar unha cita para realizar o estudo. Por cada falta de resposta ou negativa, escolleuse un novo suxeito de forma aleatoria. A información foi recollida a través dunha entrevista persoal no centro de saúde correspondente mediante un cuestionario estruturado sobre distintos aspectos como a saúde persoal e familiar e o estado sociodemográfico do individuo. Posteriormente realizouse un exame para medir a presión arterial do individuo xunto coas súas características antropométricas.

1.1.1. Descripción das variables

Indicamos a continuación cales serán as variables que consideraremos no noso traballo de entre todas as que se recolleron neste estudo:

- **Sexo:** unicamente se consideraron dúas categorías: varón ou muller.
- **Idade:** recollida como unha variable continua e non discreta.

- **Nivel educativo:** clasificado nun dos seguinte grupos: analfabetos ou sen educación formal aínda que saiban ler e escribir (grupo 1), educación completada aos 10-11 anos (grupo 2), educación completada aos 13-14 anos (grupo 3), educación completada aos 16-19 anos ou estuos non universitarios posteriores (grupo 4) e estudios universitarios (grupo 5).
- **Lugar de residencia:** dentro das cales diferenciamos dúas variables: unha que nos indica se o individuo vive en **costa** ou en **interior** e outra que nos diferenza entre **rural** e **urbano**. Temos tamén unha terceira variable categórica con catro clases combinando as posibilidades das dúas variables comentadas.
- **Medidas antropométricas:** dentro destas variables encontramos a **talla**, o **peso**, a **cadeira** e a **cintura**. Todas as medicións foron obtidas por médicos e enfermeiros. O peso corporal medíuse con precisión 0.1 kg co suxeito descalzo e vestido con roupa lixeira. A altura medíuse descalzo cunha precisión de 0.1 cm usando un dispositivo portátil montado ma parede.
- **IMC:** o índice de masa corporal calculado como variable continua.
- **Glucosa:** recolle a medida da glucosa en sangue a través dunha analítica que mide a glucosa en sangue en xaxún. Mídese en mg/dl.
- **Glucosa ás 2 horas:** recolle a medida da glucosa (un tipo de azucre) en sangue a través dunha analítica que mide a glucosa pasadas dúas horas da última comida realizada. Mídese en mg/dl.
- **Presión arterial** dentro da cal diferenciamos a presión arterial sistólica (**PAS**) e a presión arterial diastólica (**PAD**). Foi medida tres veces cun esfigmomanómetro en posición sentada cun intervalo de 5 minutos. A primeira delas desbotouse e realizouse unha media entre as dúas últimas medidicións. Con esta media considérase que queda determinada a PA do paciente.
- **HBA1C:** trátase dos resultados da proba da hemoglobina glicosada que é un exame de sangue para a diabetes de tipo 2 e a prediabetes. O resultado desta proba, e consecuentemente da nosa variable, mídese en porcentaxes e comprende os seguintes intervalos: un nivel de HBA1C normal é menor ao 5.7 %, a prediabetes atópase entre 5.7 % e 6.4 % e a diabetesd e tipo 2 para valor por encima de 6.5 %.
- **Diabetes:** temos dúas variables categóricas que recollen esta información, unha con 5 categorías (*diab_55*) na cal nos atopamos con diabetes normal, diabético descoñecido (o individuo ten diabetes pero non o sabía), diabético coñecido, e dous tipos de prediabéticos, os intolerantes á glucosa (IGT) e os que teñen alteración da glucosa en xaxún (IFG). Temos tamén outra variable (*diab_44*) na que se xunta todas as persoas diabéticas nunha soa categoría, independentemente de que o soubesen ou non.
- **Fumador:** variable dicotómica que indica se o individuo é fumador ou non.
- **Insulina:** recolle a cantidade de insulina en sangue de cada individuo medida en ?UI/ml.
- **Colesterol:** mide o colesterol en mmol/l.
- **Triglicéridos:** mide os triglicéridos en mmol/l.
- **Alcohol:** recolle o consumo de alcohol de cada individuo.
- **Hipertiroidismo:** variable dicotómica que indica se o individuo padece ou non hipertiroidismos.

Antes de comezar co estudo cabe destacar que existen variables faltantes para algún dos individuos do estudo, polo tanto coa fin de que todos os modelos de regresión axustados empregen a mesma poboación (independentemente das covariables consideradas en cada modelo) imos eliminar aqueles individuos que non teñan a totalidade das variables consideradas. Polo tanto pasaremos de ter unha mostra de 2870 individuos a ter agora unha mostra de 2528. individuos

1.2. Algunhas propiedades das variables

As variables resposta que imos considerar nos modelos axustados posteriormente serán a presión arterial sistólica (PAS) e a presión arterial diastólica (PAD). Sabemos que algúns deses modelos supoñen que a variable resposta segue unha distribución normal polo que imos facer unha representación destas dúas variables mediante un histograma para facernos unha idea de canto se axustan a normalidade cada unha delas. Podemos ver estas representacións na Figura 1.1.

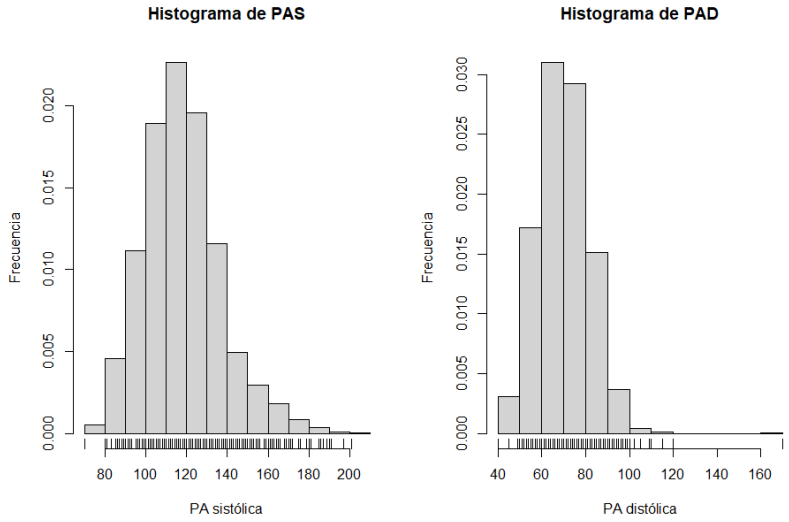


Figura 1.1: Representación mediante un histograma das variables presión arterial sistólica (PAS) e presión arterial diastólica (PAD) da base de datos do Servizo Galego de Saúde (SERGAS) extraídos entre marzo e xuño do ano 2004.

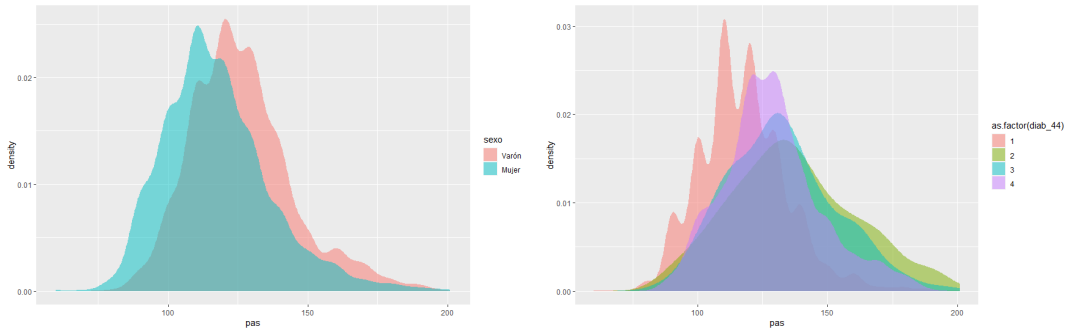
Vemos claramente como ningunha das dúas variables parece seguir unha distribución normal, polo que posteriormente cando axustemos aos nosos modelos veremos se os residuos seguen unha distribución normal e no caso de non ser así ver que distribución se axusta mellor aos nosos datos. Podemos comprobar que efectivamente non seguen unha distribución normal mediante un test de Shapiro Wilk, que en ambos casos nos devolve un p-valor menor que 2.2e-16, que é máis baixo que todos os niveis de significación habituais e polo tanto existen evidencias estatísticamente significativas para rexeitar a hipótese nula de normalidade.

Tras ver o comportamento das que serán as nosas variables resposta, imos facer un resumo do resto de variables. Para iso empregamos a función `summary` de R que nos permite ver que os nosos datos están compostos por 1172 homes e 1356 mulleres. Tamén podemos ver o número de individuos pertencentes a cada grupo de cada variable categórica. Observamos que en todas as variables hai unha distribución bastante equilibrada dos datos entre os grupos ca única diferenza das variables relacionadas coa diabetes, na cal a meirande parte dos integrantes pertencen ao grupo de non diabéticos (como era de esperar, pois a porcentaxe de poboación que ten diabetes non é para nada maioritaria).

Unha vez vistas as variables categóricas pasamos a mirar as variables continuas, aquelas que teñan valores atipicamente altos ou atipicamente baixos imos os individuos correspondentes para que nos axustes dos efectos suavizados dos modelos GAM e GAMLSS obteñamos funcións con menos variabilidade nos extremos e ademais non podemos extrapolar o comportamento da función de efecto de suavizado para valores tan altos e tan baixos cunha cantidade de datos tan pequena nesas zonas. Tras eliminar estes datos (un total de 8), quedámonos cun tamaño mostral de 2520 individuos.

Unha vez arranxado isto, procedemos a ver como inflúen algunha das nosas variable categóricas nas variables resposta a través de gráficos que representen a distribución da variable *pas* separandoo

por grupos das variables categóricas como podemos ver na Figura 1.2.



(a) Sexo (b) Diab_44

Figura 1.2: Representación da función de densidade da presión arterial sistólica separada por sexos na esquerda e polas tipos de diabéticos na esquerda.

Vemos así como parece que estas variables parece que será necesario introducilas nos modelos que vaimos axustando. Aínda que non o mostremos, sucede algo similar para o caso da presión arterial diastólica.

Como comentamos antes que parece que os nosos datos non teñen porque seguir unha distribución normal, imos representar un histograma xunto con distintas funcións de distribucións coñecidas para ver cal se aproxima máis ás nosas variables resposta. Podemos ver os resultados obtidos na Figura 1.3.

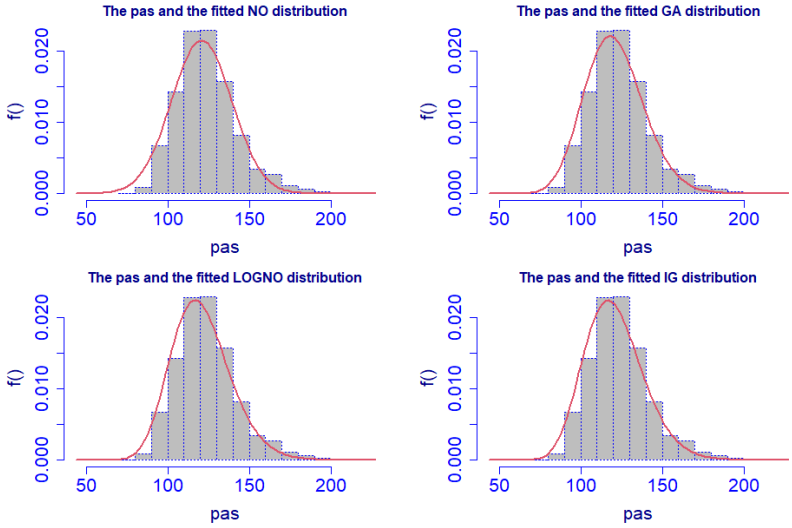


Figura 1.3: Representación mediante un histograma da variable presión arterial sistólica (PAS) xunto coa estimación da densidade considerando distintas distribucións entre as que se encontra, comezando por arriba e pola esquerda a Normal, a Gamma, a Log-Normal e Gamma Inversa.

Vemos como parece que a Log-Normal e a Gamma se axustan mellor ao histograma que a normal. O mesmo acontece se represetamos a presión arterial diastólica.

Capítulo 2

Modelos lineais xeralizados

Neste capítulo introduciremos os modelos lineais e os modelos lineais xeralizados (GLM) para comprobar as súas limitacións e xustificar así a introdución dos modelos aditivos e modelos aditivos xeralizados (GAM). Para redactar este capítulo seguíuse (34). Comezaremos falando da orixe dos modelos lineais, nos cales trataremos de axustar un modelo para o valor esperado dunha variable resposta que segue unha distribución normal en función do efecto lineal das covariables. Posteriormente pasaremos a xeralizar estes modelos para o caso de variables resposta que se sigan unha distribución dentro da familia exponencial. Trataremos tamén os algoritmos de aproximación empregados para os parámetros destes modelos que posteriormente nos servirán de base para comprender os algoritmos de estimación dos modelos GAM, GAMLSS e Joint Modelling.

2.1. Modelos lineais

Un dos principais obxectivos da estatística é cuantificar a influencia dun conxunto de variables p chamadas covariables, que denotaremos por X_1, X_2, ..., X_p, sobre unha medida dunha variable de interés, que chamaremos variable resposta e denotaremos con Y.

Para analizar a relación entre as covariables e a variable resposta empregamos o que se coñece como modelos de regresión, dentro dos cales se atopan os modelos aditivos xeralizados e os joint modelling que imos estudar, pero tamén outros modelos máis sinxelos e fáciles de interpretar que nos axudarán a comprender os conceptos que explicaremos ao longo do traballo.

A forma clásica de analizar esta relación é mediante un modelo lineal. É dicir, supoñemos que a variable resposta ,Y, é gaussiana e que as covariables,X_1, X_2, ..., X_p, afectan linearmente sobre Y. Deste xeito, para obter un modelo de regresión lineal múltiple, abonda con considerar unha combinación lineal das variables explicativas do seguinte xeito

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon. \tag{2.1}$$

Sendo, como xa se dixo, Y a variable resposta, X_1, X_2, ..., X_p unha colección de variables explicativas e $\beta_1, \beta_2, \ldots, \beta_p$ os parámetros que acompañan a ditas variables e que denominaremos coeficientes de regresión.

Recordemos que como en todos os modelos de regresión, podemos considerar un deseño fixo ou un deseño aleatorio para as variables explicativas. Imos expresar o modelo segundo o primeiro caso para posteriormente ver como se realizará a estimación de parámetros. Deste xeito, baixo deseño fixo, podemos expresar o modelo (2.1) como:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i \tag{2.2}$$

sendo Y_i a variable resposta do i-ésimo individuo, $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ as variables explicativas do mesmo e ε_i o erro asociado a dito individuo. Polo tanto, podemos pasar agora e expresar o modelo (2.2) de

forma matricial como segue:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{2.3}$$

Podemos expresar isto como

$$Y = X\,\beta + \varepsilon, \tag{2.4}$$

onde Y é o vector resposa, X a matriz de deseño (onde cada fila representa a un individuo), β é o vector de coeficientes e ε o vector de erros que verifica que $\varepsilon \sim N(0, \sigma^2\, I_n)$.

Considerando o modelo da forma (2.4), procederemos a estimar β mediante un procedemento que se coñece como mínimos cadrados. Deste xeito, escolleremos como estimador aquel $\hat{\beta}$ que verifique que

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \arg\min \sum_{i=1}^{n} (Y_i - x_i\,\beta)^2. \tag{2.5}$$

Podemos expresar ese problema de minimización en notación matricial de forma equivalente:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y - X\,\beta)'\,(Y - X\,\beta).$$

Derivando facilmente podemos obter que a súa solución é

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Observación 2.1. Notemos que para que o estimador este ben definido é necesario que a matrix$zX'X$ sexa invertible e polo tanto necesitamos ter un número de observacións maior ou igual que o número de covariables que imos considerar ($n \geq p$).

2.2. Modelos lineais xeralizados

Unha das suposicións que facemos nos modelos lineais é que a nosa resposta é gaussiana. Deste xeito, unha extensión dos modelos lineais que nos permitirá considerar outro tipo de variables resposta será o que coñecemos como modelos lineais xeralizados, introducidos por primeira vez por Nelder e Wedderburn en 1972 (consultar ([21])). Para construír este tipo de modelos aplicámoslle unha transformación h para que a predición realizada teña o dominio adecuado dando lugar ao seguinte modelo:

$$E\,(y/x_1, x_2, \ldots, x_p) = h(\eta) = h(\beta_0 + \beta_1\,x_1 + \beta_2\,x_2 + \cdots + \beta_p\,x_p). \tag{2.6}$$

De forma equivalente podemos escribir que

$$g(E\,(y/x_1, x_2, \ldots, x_p)) = \beta_0 + \beta_1\,x_1 + \beta_2\,x_2 + \cdots + \beta_p\,x_p. \tag{2.7}$$

onde $g = h^{-1}$ é o que se coñece como función link (esta función será diferente dependendo da distribución que siga a variable resposta coa que estemos traballando). Neste tipo de modelos supoñemos que Y pertence a unha familia exponencial, a función link g cumpre que $g(\mu) = \eta$ e ademais $Var(Y) = \phi \cdot V(\mu)$ onde V é unha función de μ e ϕ o parámetro de dispersión.

O caso gaussiano contémplase dentro dos GLM sen máis que considerar como función link a identidade. Como exemplo podemos comentar que se a variable resposta segue unha distribución Gamma entón consideramos como función link a función logarítmica.

A diferenza do que acontece nos modelos lineais, neste caso a estimación de β realízase maximizando a función de log-verosimilitude $\ell(\beta)$ de forma iterativa a través do algoritmo de Fisher Scoring (IRLS) que procede como indicamos a continuación:

1. Establécense uns valores iniciais $\hat{\mu}_i = y_i + \delta$ e $\hat{\eta}_i = g(\hat{\mu}_i)$, sendo δ un valor próximo ou igual a cero de forma que faga viable o cálculo de $\hat{\eta}_i$.

2. Calculamos os pseudovalores
$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\, g'(\hat{\mu}_i)$$
e os pesos
$$\omega_i = \frac{1}{g'(\hat{\mu}_i)^2\, V(\hat{\mu}_i)},$$
sendo $V(\mu) = Var_\mu(Y)$.

3. Obtemos β como o estimador por mínimos cadrados penalizados
$$\hat{\beta} = arg\, min \left[\sum_{i=1}^{n} \omega_i\, (z_i - x_i'\, \beta)^2 + \sum_j \lambda_j\, \beta_j'\, S_j\, \beta_j \right].$$
Podemos escribilo de forma matricial como
$$\hat{\beta} = arg\, min\, \|\sqrt{W}\, (z - X\, \beta)\|^2. \tag{2.8}$$

4. Actualizamos $\hat{\eta}_i = x_i'\, \hat{\beta}$ e $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

5. Repetir os pasos 2, 3 e 4 ata converxencia.

Se queremos saber os graos de liberdade dun GLM, non temos máis que obter o valor que fai mínimo (2.8), obtendo así
$$\hat{\beta} = (X^T\, W\, X)^{-1}\, X^T\, W.$$
Daquela os graos de liberdade serán os mesmos que a traza da matriz de proxección do noso modelo. É dicir,
$$df = tr\left[X\, \hat{\beta}\right] = tr\left[X\, (X^T\, W\, X)^{-1}\, X^T\, W\right],$$
que non vai ser máis que o número de parámetros do modelo.

Tras o modelo linear, engadimos flexibilidade ao incluír distribucións distintas á normal na nosa variable resposta Y cos GLM. Aínda así, continuamos supoñendo que os efectos das covariables continuas é sempre lineal o cal non ten porqué ser adecuado para moitos casos reais, especialemente no ámbito biomédico.

Unha forma de flexibilizar o noso modelo, ademais de introducir diferentes distribucións para a variable resposta, é asumir que o efecto da covariable continua X pode ter unha forma lineal descoñecida, f:
$$\eta = \beta_0 + f(x).$$
Dentro do contexto dos modelos GLM, existen alternativas paramétricas para estimar f, como supoñer que o efecto de X é un efecto non lineal pero paramétrico:
$$\eta = \beta_0 + f(x) = \beta_0 + f_\beta(x).$$
Unha das posibiliades é considerar a regresión polinómica, que vén dada por :
$$\eta = \beta_0 + \beta_1\, x + \beta_2\, x^2 + \cdots + \beta_p\, x^p.$$
Deste xeito canto maior sexa o grao do polinomio, maior será a flexibilidade no axuste do modelo.

O problema principal da regresión polinómica é que os seus axustes non son suaves (debido a que esiximos que a función sexa continuamente diferenciable), se tratamos de buscar unha regesión que produza un axuste suave e ao mesmo tempo que detecte o comportamento local (o que vemos que non

sucede na gráfica da Figura 2.1) podemos empregar o que se considera como funcións splines e que darán lugar aos modelos aditivos xeralizados que veremos no seguinte capítulo.

Para comprobar isto que acabamos de comentar imos empregar a base de datos `mcycle` da libraría `MASS` xunto coa función `glm` de R para realizar un axuste lineal, cuadrático e cúbico que trate de explicar a variable resposta *accel* en función da variable explicativa *times*. Podemos ver os resultados obtidos na gráfica da Figura 2.1.

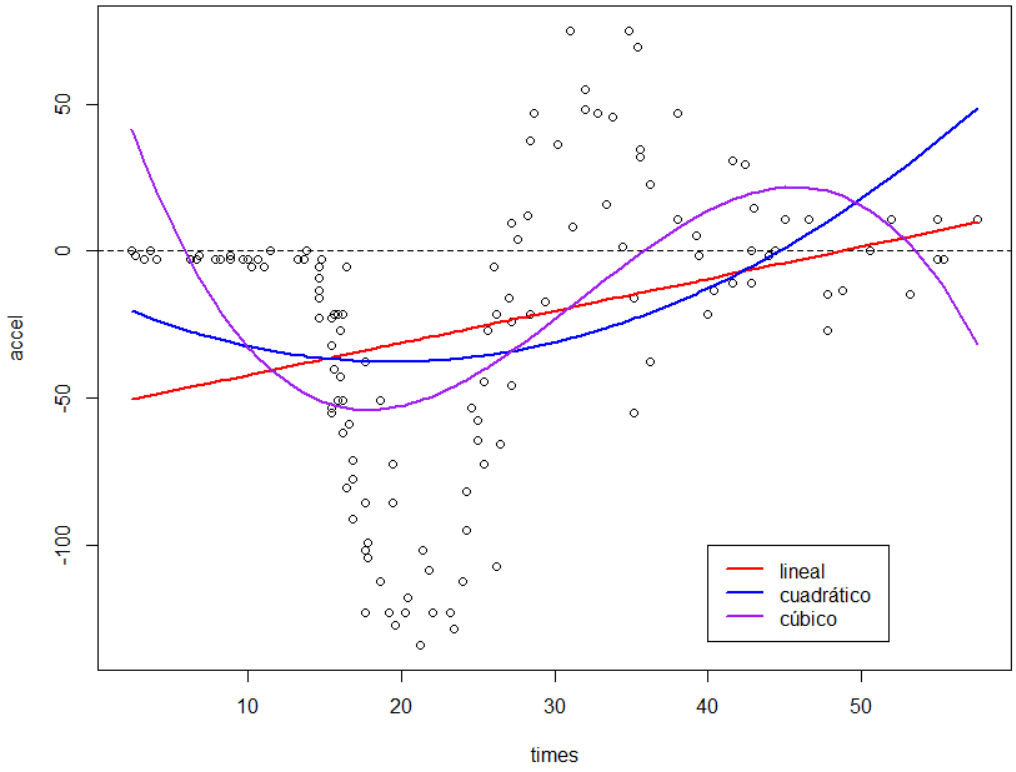


Figura 2.1: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste lineal (en vermello), cuadrático (en azul) e cúbico (en violeta) dun modelo que trate de explicar a función de regresión de *accel* en función de *times*

Vemos que efectivamente, a pesar de considerar ata un axuste de grao tres o noso modelo non explica de forma axeitada os datos. Poderíamos considerar que aumentando o grao do axuste, este melloraría notablemente pero podemos ver que non é así na gráfica da Figura 2.2

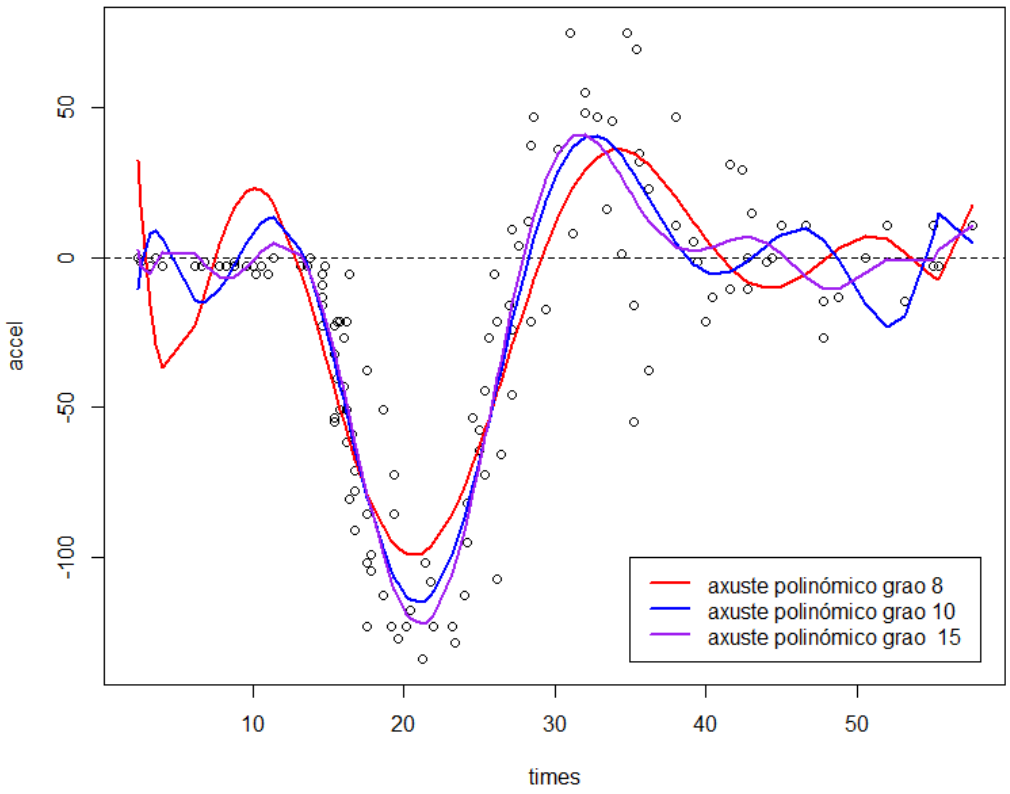


Figura 2.2: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste polinómico de grao 8 (en vermello), de grao 10 (en azul) e de grao 15 (en violeta) dun modelo que trate de explicar a función de regresión de *accel* en función de *times*

Vemos que aínda que considerando un grao bastante grande nos imos aproximando ao verdadeiro comportamento do modelo que desexamos axustar, o feito de esixir que todas as derivadas sexan continuas (como acontece no caso da regresión polinómica que estamos considerando) provoca que as nosas curvas de axuste teñan demasiada rugosidade, algo que non é desexable nos axustes de modelos e que ademais non se corresponde co comportamento real da curva que tratamos de aproximar.

Polo tanto, en vista aos resulados obtidos nas Figura 2.1 e 2.2 parece claro que ante determinadas situacións, como acontece na base de datos `mcycle`, é preciso considerar modelos non lineais (que non sexa regresión polinómica) para tratar de explicar o efecto dunha variable explicativa sobre unha variable resposta. Xustificamos así a introdución dos modelos aditivos, que precisamente traballarán con efectos non lineais das covariables, algo que parece axeitado para esta base de datos.

Capítulo 3

Modelos aditivos xeralizados

Coas Figura 2.1 e 2.2 parece clara a necesidade da existencia de modelos que nos permitan introducir variables explicativas que teñan un efecto non lineal sobre a variable resposta. Estes novos modelos serán o que se coñecen como modelos aditivos. Nesta sección introduciremos os modelos aditivos, os modelos aditivos xeralizados (GAM), seguindo (34). Para iso comezaremos presentando un modelo aditivo xeralizado no que poderemos atopar o que se coñece como funcións de suavizado. Para presentar este tipo de funcións e ver como traballar con elas introduciremos o caso particular dos suavizadores unidimensionais para presentar conceptos como a definición de splines, que é unha base de splines e o papel fundamental que xogan nos modelos aditivos. Tras isto, pasaremos a falar dos algoritmos de estimación dos parámetros dos modelos, que non serán máis que unha extensión dos algoritmos para o caso dos GLM. Posteriormente pasaremos ao caso multivariante considerando máis dunha variable explicativa e despois ampliaremos este tipo de modelos a variables resposta distintas da normal, pero sempre dentro dunha familia exponencial, como xa acontecera no caso dos modelos lineais. Veremos tamén a aplicación destes modelos a datos biomédicos empregando a base de datos recollida no Capítulo 1.

Un modelo aditivo xeralizado é un modelo lineal xeralizado ao que se lle inclúe o efecto de funcións de suavizado de covariables (isto quere dicir que o efecto que teñen as variables explicativas sobre a variable resposta xa non ten por que ser lineal). Podemos presentar este tipo de modelos mediante a seguinte estrutura:

$$g(\mu_i) = A_i\,\theta + f_1\left(x_{1i}\right) + f_2\left(x_{2i}\right) + f_3\left(x_{3i}, x_{4i}\right) \tag{3.1}$$

onde $\mu_i \equiv \mathbb{E}(Y_I)$ e $Y_i \sim EF(\mu_i, \phi)$. Y_i é a variable resposta, $EF(\mu_i, \phi)$ denota a familia exponencial de distribución coa media μ_i e parámetro de dispersión ϕ. A_i é a fila da matriz do modelo paramétrico, θ o correspondente vector de coeficientes e f_j a función de suavizado da covariable x_k.

3.1. Suavizado univariante

Para introducir a forma de representar e estimar as compoñentes do modelo (3.1) imos comezar supoñendo un modelo que conteña unicamante unha función dunha covariable como mostramos a continuación

$$y_i = f(x_i) + \varepsilon_i \tag{3.2}$$

onde y_i é a variable resposta, x_i a covariable, f a función de suavizado e ε_i son variables aleatorias independentes dunha $N(0, \sigma^2)$.

Cando tratamos de axustar unha curva f de xeito que a estimación pase preto dos puntos da nosa mostra e ao mesmo tempo esta curva sexa suave xurden o que coñecemos como splines. Deste xeito, neste novo contexto non imos buscar unicamente un modelo que se axuste adecuadamente aos nosos datos, se non que tamén pediremos que este modelo sexa suave (é dicir, que a curva resultante teña pouca rugosidade e non realice unha sobreestimación). Tratando de resolver esta cuestión aparece o

que coñecemos como problema de mínimos cadrados penalizados, que ten como función obxectivo a que mostramos a continuación:

$$S(g) = \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \lambda \int (g''(x))^2 \, dx \tag{3.3}$$

onde λ é o que se coñece como parámetro de penalización. Podemos ver que a Ecuación (3.3) ten unha forma moi similar ao caso da ecuación de mínimos cadrados (2.5) pero cun termo a maiores que permite manter un equilibrio entre unha curva que realice un bo axuste e ao mesmo tempo sexa unha curva suave.

3.1.1. Suavizadores

A solución de minimizar a función da ecuación (3.3) son o que se coñece como funcións splines e que presentaremos a continuación.

Definición 3.1. Unha función $f : [a, b] \to \mathbb{R}$ é un spline polinómio de grao l se verifica que:

1. $f(x)$ é $(l-1)$ veces continuamente diferenciable.
2. $f(x)$ é un polinomio de grao l para $x \in [k_j, k_{j+1})$ con $j = 1, \ldots, m-1$.

Deste xeito o espazo das funcións splines pode considerarse como un espazo vectorial de dimensión $(m+l-1)$ onde m é o número de nodos no que temos dividido o noso dominio da función de suavizado e l o grao da función spline que estamos considerando. Deste xeito podemos representar cada spline nunha base de $d = (m + l - 1)$ funcións, que denominaremos bases de splines, como mostramos a continuación:

$$f(x) = \sum_{j=1}^{d} \beta_j \, B_j(x) \tag{3.4}$$

Vendo a representación da función spline dada en (3.4), podemos preguntarnos a necesidade de empregar unha base para represetar os suavizadores do noso modelo. Estaremos interesados en representar as nosas funcións de suavizado nunha base para poder así expresar os modelos aditivos xeralizados como un GLM sobreparametrizado en función do número de nodos (isto será debido a que seleccionaremos un número de nodos arbitrario que compensaremos co parámetro de suavizado λ) e poder empregar os métodos de estimación vistos para o caso dos GLMs sen máis que realizar pequenas modificacións neles.

Bases de splines

Chegados a este punto, debemos preguntarnos cales so as bases de splines máis empregadas e as súas principais características. Sabemos que o conxunto de splines é un espazo vectorial de dimensión o número de graos de liberdade do spline e polo tanto admite diversas bases.

Unha idea inicial pode ser considerar que f é un polinomio e deste xeito se creemos que f é un polinomio de orde 3 unha base para esta función sería $B_1(x) = 1$, $B_2(x) = x$, $B_3(x) = x^2$ e $B_4(x) = x^3$.Este tipo de bases resultan problemáticas cando non estamos interesados en ver o comportamento de f unicamente nus puntos determinados, se non que nos interesa o seu comportamento en todo o dominio, pois temos que f ten que ser continuamente diferenciable o que provoca que teña grandes oscilacións, especialmente na fronteira do dominio.

Visto que esta idea inicial non parece axeitada, presentaremos a continuación unha base (flexible) local que é a formada polos *Basic*-splines (ou simplemente B-Splines, para máis detalles, ver (4)). O principal motivo do uso xeralizado destas bases é computacional: teñen unhas propiedades numéricas moi boas debido ao feito de que os B-Splines teñen un soperte mínimo e a superposición entre os

distintos elementos da base tamén é mínima, o que provoca que o coeficiente asociado a un B-Spline esté relacionado co menor número de coeficientes asociados a outros B-Splines posibles.

As bases de B-splines están definidas de forma local e unicamente toman valores positivos no intervalo entre $l+2$ nodos, sendo l o grao da base de B-Splines escollido. Podemos definilos de forma recursiva, comezando polo B-spline de orde $l=0$ como mostramos a continuación:

$$B_j^0(\nu) = I(k_j \leq \nu < k_{j+1}) = \begin{cases} 1 & k_j \leq \nu < k_{j+1},\ j = 1, \ldots, d-1 \\ 0 & \text{noutro caso,} \end{cases}$$

onde $I(\cdot)$ denota a función indicadora. Os B-Splines de maior orde podemos definilos como segue:

$$B_j^l(\nu) = \frac{\nu - k_{j-l}}{k_j - k_{j-l}} B_{j-1}^{l-1}(\nu) + \frac{k_{j+1} - \nu}{k_{j+1} - k_{j+1-l}} B_j^{l-1}(\nu).$$

Podemos ver distintas bases de B-Splines de distintos graos na Figura 3.1 considerando nodos equiespazados.

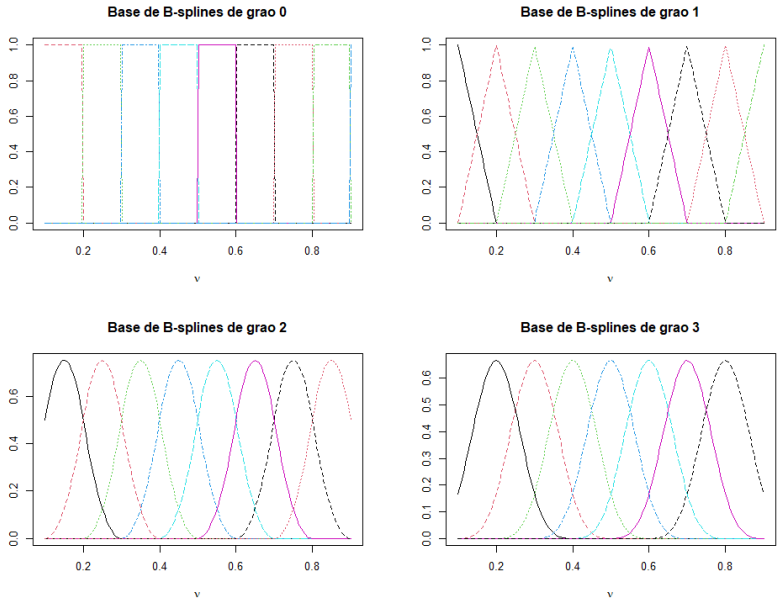


Figura 3.1: Funcións das bases de B-Splines para graos, de esquerda a dereita comezando por arriba, $l=0$, $l=1$, $l=2$ e $l=3$ respectivamente, tomando nodos equidistantes.

En xeral, un B-Spline de grao l ten as seguintes características ([6])

- Consiste en $l+1$ pezas de polinomios de l nodos internos.
- Como cada B-Spline está composto por un conxunto de polinomios, é sinxelo calcular a súa derivada.
- Ademais, para todas as bases de funcións a $(l-1)$ derivada é continua nos puntos de unión.
- Toman valores positivos no intervalos dos $l+2$ nodos adxacentes e cero no resto de dominio.
- Para cada punto $\nu \in [a,b]$, temos que $\sum_{j=1}^{d} B_j(\nu) = 1$.

- Cada función da base solápase con exactamente $2\,l$ función da base adxacente (agás nos extremos).
- Para cada punto ν, temos que $l+1$ B-Splines son distintos de cero.

Antes de continuar introducindo novos conceptos precisos para o axuste de modelos GAM, imos comprobar que efectivamente resolven os problemas de malos axustes da regresión polinómica que vimos no capítulo anterior. Para iso retomaremos o mesmo exemplo e compararemos un axuste lineal e cuadrático xunto co axuste realizado mediante unha función spline. Podemos ver estes axustes na Figura 3.2.

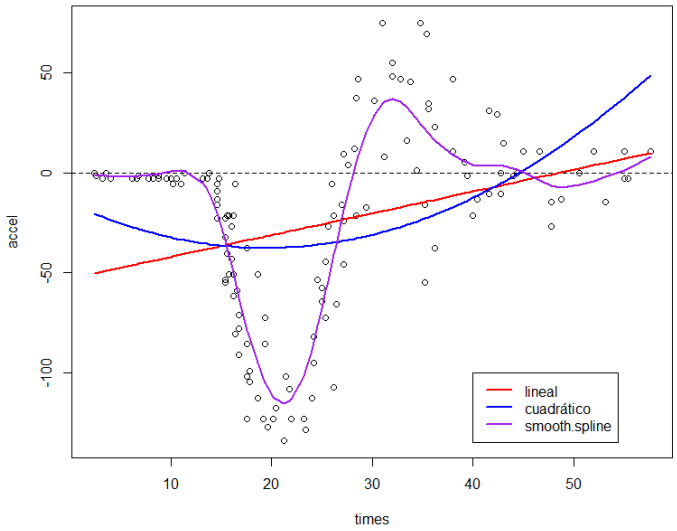


Figura 3.2: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste polinómico lineal(en vermello), cuadrático (en azul) e dunha función de suavizado (en violeta) dun modelo que trate de explicar a función de regresión de *accel* en función de *times*

Podemos ver así na gráfica da Figura 3.2 unha xustificación para o uso de splines no canto de modelos polinómicos pois para valores baixos da variable explicativa x, podemos ver como o modelo polinómico estima un pico da función de regresión que non existe na realidade, mentres que a estimación do spline si recolle o comportamento real da función nesta situación. Recordemos ademais que para comezar a ver un axuste relativamente bo no caso da regresión polinómica, tiñamos que considerar un polinomio de grao 15. Pola contra, os graos de liberdade efectivos (que posteriormente explicaremos) no caso da función de suavizado son de aproximadamente 12. É dicir, conseguimos un axuste que explica mellor o comportamento dos datos, con menor rugosidade e ademais teremos un modelo máis sinxelo, pois temos menos graos de liberdade.

Chegados a este punto podemos preguntarnos como podemos controlar os graos de liberdade do modelo que queremos axustar. Como xa comentamos, unha forma de facelo será seleccionando o número de nodos no dominio da función f de (3.2). Comprobemos de forma práctica o efecto que ten a selección do número de nodos á hora de realizar o axuste do noso modelo. Para iso continuaremos empregando a base de datos considerada ata agora e realizaremos distintos axustes de modelos aditivos considerando diferentes cantidades de nodo. Podemos comprobar efectivamente na Figura 3.3 que a maior número de nodos, o axuste que se realiza mellora con respecto ao anterior. Probouse tamén a considerar unha cantidade de nodos maior que dez, pero a partir deste valor todos os axustes realizados eran practicamente equivalentes.

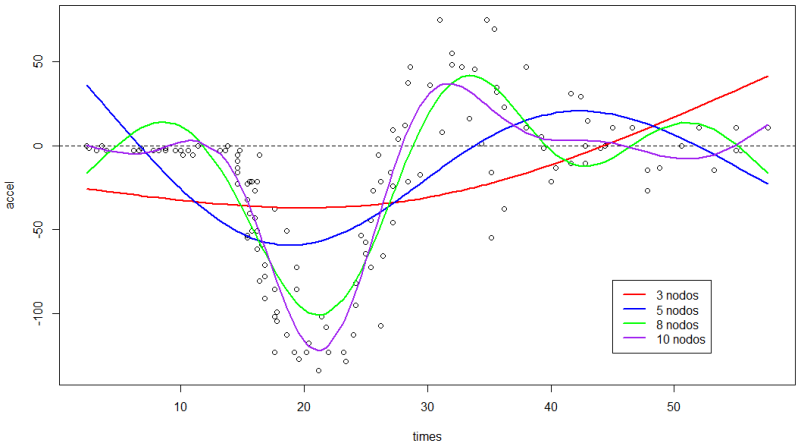


Figura 3.3: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste dun modelo aditivo considerando: 3 nodos (en vermello), 5 nodos (en azul), 8 nodos (en verde) e 10 nodos (en violeta) dun modelo que trate de explicar a función de regresión de *accel* en función de *times*

Porén, esta forma de regular o suavizado da estimación da curva non será moi eficiente posto que non saberemos a posición que deben ter os nodos (non sempre teñen porque ser equiespazados) nin o número adecuado para unha boa estimación (unha posibilidade é ir tanteando en función do axuste conseguido, pero non resulta útil máis alá de entender o efecto que ten o número de nodos no axuste realizado).

Por este motivo fixaremos un número de nodos o suficientemente alto (nas funcións empregadas en R para o axuste de modelos GAM a selección do número de nodos realízase de forma automática, aínda que se pode modificar manualmente se se desexa) como para permitir que os grados de liberdade do modelo estén representados por estes nodos e posteriormente controlar o suavizado do modelo mediante o parámetro de penalización λ do problema de mínimos cadrados penalizados. Deste xeito, no canto de minimizar a función

$$\|y - X\,\beta\|^2$$

podemos obter as estimacións do modelo minimizando a función

$$\|y - X\,\beta\|^2 + \lambda \sum_{j=2}^{k-1} \{f(x_{j-1}*) - 2\,f(x_j*) + f(x_{j+1}*)\}^2$$

onde estamos supoñendo que a medida de rugosidade se representa como unha suma de segundas diferenzas ao cadrado da función nos nodos. Deste xeito, cando f sexa moi rugosa a penalización tomará valores altos e cando sexa pouco rugosa a penalización tomará valores baixos.

Daquela, o parámetro de suavizado, λ, controla dalgunha forma a compensación entre o suavizado da función f e a fidelidade da estimación aos datos. A principal vantaxe desta penalización é que o suavizado non depende do número e da posición dos nodos se non que o fai dun parámetro de suavizado como podemos ver na gráfica da Figura 3.4.

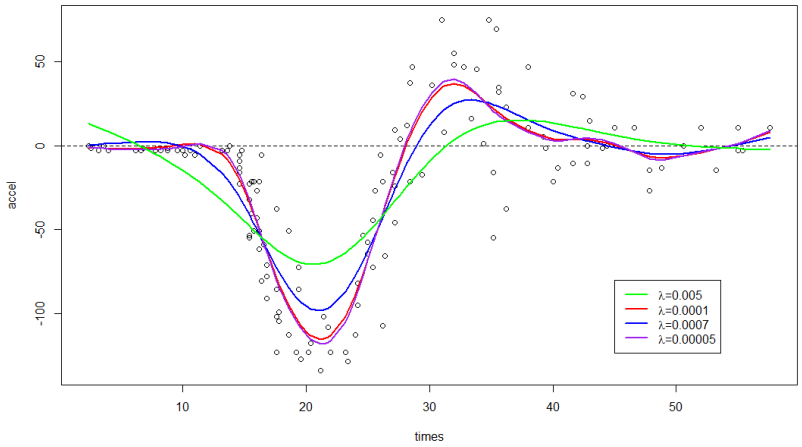


Figura 3.4: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste dunha función spline que trate de explicar a función de regresión de *accel* en función de *times* tomando os distintos valores do parámetro de suavizado λ: 0.005 (en verde),0.0001 (en vermello), 0.0007 (en azul) e 0.000005 (en violeta).

Vemos así como a curva vermella representa o axuste dado para un λ calculado de forma óptima empregando o comando `smooth.spline`. Mentres que a curva azul e verde representa un λ de maior valor que provoca unha curva demasiado suave (é dicir, un sobresuavizado especialmente no caso de $\lambda = 0.005$) e a curva morada representa un axuste infrasuavizado ao considerar un λ demasiado pequeno.

Unha vez vista esta base, podemos presentar algúns suavizadores penalizados existentes que se poden expresar seguindo a fórmula (3.4) (consultar (33)):

- **Smoothing splines.** (ver (31) e (13) para máis detalles).

- **Cubic regression splines.** (ver (33) para máis detalles) Existen moitas bases equivalentes para representar os splines cúbicos. Unha das posibilidades é parametrizar o spline en termos do seu valor nos nodos.

  Consideremos que queremos definir unha función spline cúbica, $f(x)$, con k nodos $x_1, \ldots, x_k$. Sexa $\beta_j = f(x_j)$ e $\delta_j = f''(x_j)$. Entón o spline pode ser representado como

  $$f(x) = a_j^- \, \beta_j + a_j^+(x) \, \beta_{j+1} + c_j^- \, \delta_j + c_j^+(x) \, \delta_{j+1} \quad \text{se} \quad x_j \leq x \leq x_{j+1} \tag{3.5}$$

  onde as bases de funcións veñen dadas por

  $$a_j^-(x) = \frac{x_{j+1} - x}{h_j} \qquad c_j^-(x) = \frac{[(x_{j+1} - x)^3 / h_j - h_j \, (x_{j+1} - x)]}{6}$$
  $$a_j^+(x) = \frac{x - x_j}{h_j} \qquad c_j^+(x) = \frac{[(x - x_j)^3 / h_j - h_j \, (x - x_j)]}{6}$$

  A imposición de que o spline debe ter a segunda derivada continua, en x_j e deber ter segunda derivada nula en x_1 e x_k

- **Thin plate regression splines.** (ver (32) para máis detalles.) Empréganse cando queremos realizar suavizados multivariantes. Para realizar este tipo de suavizados temos que buscar unha

forma de medir a rugosididade no campo multivariante. Para o caso bivariante podemos definir a seguinte medida:

$$J(g) = \int\int \left[\left(\frac{\partial^2 g}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x_1 \, \partial x_2} \right)^2 + \left(\frac{\partial^2 g}{\partial x_2^2} \right)^2 \right] dx_1 \, dx_2.$$

Neste contexto escolleremos a función g que minimice o seguinte valor:

$$S(g) = \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \lambda \, J(g) \tag{3.6}$$

sendo λ un número real que xoga o papel do parámetro de suavización. A solución a este problema de minimización é o que coñecemos como *thin plate splines* e que procedemos a definir.

Consideremos un conxunto de puntos de $\mathbb{R}^2$ non alineados $t_1, \ldots, t_n$, as funcións

$$\begin{aligned} \eta(r) &= \frac{1}{16\,\pi} \, r^2 \log \, r^2 \quad \text{se} \quad r > 0, \\ \eta(0) &= 0 \end{aligned}$$

e as seguintes funcións lineais básicas,

$$\begin{aligned} \phi_1(x_1, x_2) &= 1 \\ \phi_2(x_1, x_2) &= x_1 \\ \phi_3(x_1, x_2) &= x_2. \end{aligned}$$

Deste xeito, calquer función lineal pode se expresada como combinación lineal das tres funcións anteriores. Teremos polo tanto a seguinte definición.

Definición 3.2. Unha función g é un **thin plate spline** sobre os puntos $t_1, t_2, \ldots, t_n$ se presenta a seguinte forma

$$g(t) = \sum_{i=1}^{n} \delta_i \eta(\|t - t_i\|) + \sum_{j=1}^{3} a_j \, \phi_j(t)$$

sendo as constantes δ_i e a_j determinados coeficientes.

- **P(enalized) Splines** (ver (7) para máis detalles). Os P-splines son suavizadores de rango baixo que empregan unha base de B-splines (xeralmente definidos en nodos equiespazados) cunha penalización aplicada directamente aos parámetros β_i para controlar a rugosidade da función. Deste xeito, cando traballamos coa función obxectivo que queremos minimizar, dada por

$$\|\sqrt{W} \, (z - X \, \beta)\|^2 + \lambda \, \beta^T \, S \, \beta$$

construiremos unha matriz de penalización S que penalice as diferenzas (ao cadrado) da orde desexada dos parámetros β.

No caso de que queiramos considerar as diferenzas de orde 1, teremos que a matriz de penalización será

$$S = \begin{pmatrix} 1 & -1 & 0 & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 1 & \end{pmatrix}.$$

Polo tanto o termo da penalización na función que queremos minimizar será

$$\beta^T S \beta = \sum (\beta_{j+1} - \beta_j)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + \dots$$

Aumentando a orde das diferenzas. aumentamos a flexibilidade da estimación. Habitualmente considérase orde dúas ou orde tres.

Podemos ver o efecto de considerar algúns destes suavizadores na regresión na Figura 3.5 que mostramos a continuación.

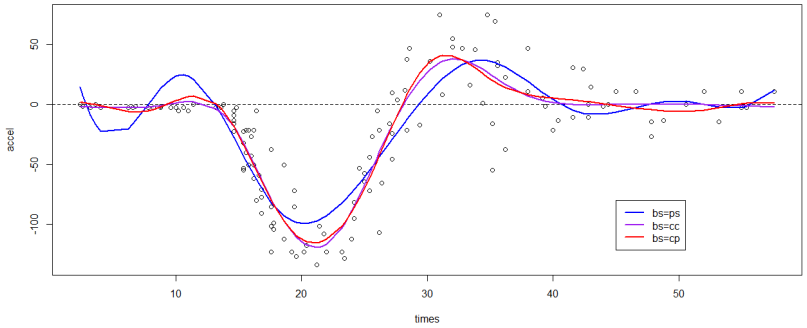


Figura 3.5: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste dunha función spline que trate de explicar a función de regresión de *accel* en función de *times* tomando tres bases de splines distintas: P-Splines (en azul), cubic regression splines (en morado) e unha versión cíclica dos P-spline(en vermello).

Graos de liberdade

Consideremos un modelo aditivo da seguine forma

$$y_i = f(x_i) \tag{3.7}$$

Se representamos a función de suavizado f nunha base podemos expresar (3.7) como un GLM sobreparametrizado como segue:

$$y_i = X_i\,\beta.$$

Para obter o noso vector de parámetros β temos que resolver o seguinte problema de mínimos cadrados penalizados que segue:

$$\hat{\beta} = arg\,min\,\|\sqrt{W}\,(z - X\,\beta)\|^2 + \lambda\,\beta^T\,S\,\beta, \tag{3.8}$$

Se resolvemos (3.8) obtemos a seguinte solución

$$\hat{\beta} = \left(X^T\,W\,X + \lambda\,S\right)^{-1}\,X^T\,W.$$

Daquela a **matriz de proxección** será

$$A = X\,\left(X^T\,W\,X + \lambda\,S\right)^{-1}\,X^T\,W.$$

Deste xeito, en analoxía cos modelos lineais, os **graos de liberdade efectivos** serán

$$edf = tr\left[X\,\left(X^T\,W\,X + \lambda\,S\right)^{-1}\,X^T\,W\right] = tr(A).$$

Algoritmo de aproximación

Para un valor dado λ, a estimación de β pode darse mediante o que se coñece como P-IRLS, que non é máis que unha modificación do algoritmo IRLS que se emprega na estimación dos modelos GLM. Supoñendo λ coñecido, algoritmo *P-IRLS* procede do seguinte xeito

1. Establécense uns valores iniciais $\hat{\mu}_i = y_i + \delta$ e $\hat{\eta}_i = g(\hat{\mu}_i)$, sendo δ un valor próximo ou igual a cero de forma que faga viable o cálculo de $\hat{\eta}_i$.

2. Calculamos os pseudovalores
$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\, g^{'}(\hat{\mu}_i)$$
e os pesos
$$\omega_i = \frac{1}{g^{'}(\hat{\mu}_i)^2\, V(\hat{\mu}_i)}$$
sendo $V(\mu) = Var_{\mu}(Y)$.

3. Obtemos β como o estimador por mínimos cadrados penalizados
$$\hat{\beta} = arg\, min \left[\sum_{i=1}^{n} \omega_i\, (z_i - x_i^{'}\, \beta)^2 + \sum_{j} \lambda_j\, \beta_j^{'}\, S_j\, \beta_j \right].$$
Podemos escribilo de forma matricial como
$$\hat{\beta} = arg\, min\, \|\sqrt{W}\, (z - X\, \beta)\|^2 + \lambda\, \beta^T\, S\, \beta,$$
onde S é o que coñecemos como **matriz de penalización**.

4 Actualizamos $\hat{\eta}_i = x_i^{'}\, \hat{\beta}$ e $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

5 Repetir os pasos 2, 3 e 4 ata converxencia.

Xa vimos como obter β coñecido λ, pero á hora de traballar con datos reais este parámetro non vai ser coñecido, polo que tamén terá que ser aproximado. Obter ambas estimacións ao mesmo tempo dá lugar a dúas formas de proceder:

1. Introducir a estimación de λ dentro do algoritmo *P-IRLS*, obtendo así un algoritmo máis eficiente pero con problemas de converxencia.

2. Obter a estimación de λ fóra do algoritmo *P-IRLS*, que nos proporciona un algoritmo menos eficiente pero asegura converxencia. Deste xeito para cada estimación que fagamos de λ imos ter que aplicar o algoritmo *P-IRLS* para obter a estimación de β. É un procedemento anidado.

Selección automática do parámetro λ

Dado λ (ou equivalentemente para un valor de *edf*):
$$\hat{f} = S_{\lambda}\, z = S_{edf}\, z$$
A selección do parámetro de suavización óptimo, λ_{opt}, basearase na minimización dalgún criterio aproximado de erro:

- Generalized Cross-Validation (GVC). Este valor defínese como segue:
$$GCV(\lambda) = \frac{n \times \text{Deviance}}{[n - \gamma\, tr(S_{\lambda})]^2}$$
sendo γ un parámetro de escala para evitar valores de λ altos. Acostuma a empregarse para modelos con parámetro de escala descoñecido (como por exemplo o modelo gaussiano).

- Unbiased Risk Estimator (UBRE): Criterio AIC reescalado que vén dado por

$$UBRE(\lambda) = \frac{\text{Deviance}}{n} + \frac{2\,\gamma\,\phi\,tr(S_\lambda)}{n} - \phi.$$

  Este criterio acostuma a empregarse para modelos con parámetro de escala coñecido (como poden ser o modelo binomial ou o modelo poisson).

- Restricted Maximum Likelihood (REML). Un suavizador penalizado de forma cuadrática (como é o caso dos P-splines), pode ser representado como un modelo mixto (para máis información consultar (7) e (28)). Deste xeito podemos formular e implementar os P-Splines no ámbito dos GLM mixtos e así empregar o Restrictec Maximum Likelihood (REML) que nos permitirá traballar co problema de selección do parámetro de suavización dende unha visión diferente.

Observación 3.3. Os criterios *GCV* e *UBRE* poden ser computacionalmente costosos (λ_{opt} obténse a través dunha busca refinada nunha reixilla)

Procedemos agora a mostrar gráficamente os axustes realizados ao considerar os métodos de selección automática do parámetro de suavizado vistos.

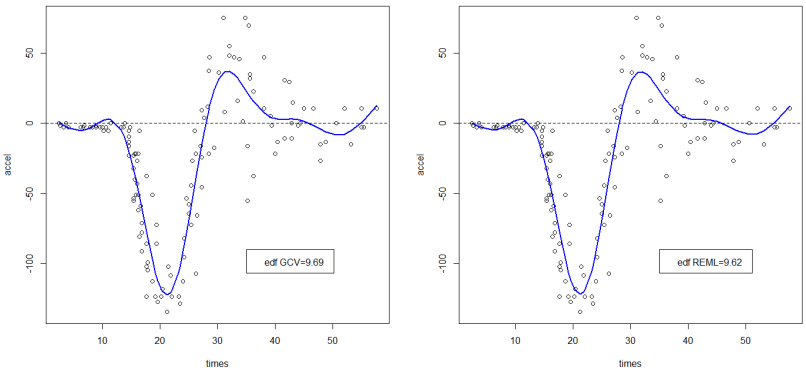


Figura 3.6: Representación nun diagrama de dispersión da variable *accel* frente á variable *times* da base de datos `mcycle` xunto cun axuste dun modelo aditivo que trata de explicar a función de regresión de *accel* en función de *times*. Na parte esquerda considerouse un criterio de validación cruzada (GCV) para seleccionar o parámetro de suavizado λ, mentres que na dereita se empregou máxima verosimilitude restrinxida (REML).

Vemos así como o axuste apenas varía neste caso e gráficamente parece a mesma curva. No referente aos graos de liberdade efectivos axustados por cada método, tamén son moi similares pois só difiren en 0.07.

3.2. Modelos aditivos xeralizados

Consideremos agora de novo un modelo aditivo xeralizado coa seguinte estrutura:

$$g(\mu_i) = A_i\,\theta + f_1\,(x_{1i}) + f_2\,(x_{2i}) + f_3\,(x_{3i}, x_{4i}) \tag{3.9}$$

onde $\mu_i \equiv \mathbb{E}(Y_I)$ e $Y_i \sim EF(\mu_i, \phi)$. Y_i é a variable resposta, $EF(\mu_i, \phi)$ denota a familia exponencial de distribución coa media μ_i e parámetro de dispersión ϕ. A_i é a fila da matriz do modelo paramétrico, θ o correspondente vector de coeficientes e f_j a función de suavizado da covariable x_k.

O problema de identificación

Se supoñemos que temos agora dúas covariables, x e v, para explicar a variable resposta y, entón o modelo aditivo resultante sería:

$$y_i = \alpha + f_1(x_i) + f_2(v_i) + \varepsilon_i \tag{3.10}$$

onde α é o intercepto, f_j son as funcións de suavizado e os ε_i son variables aleatorias independentes $N(0, \sigma^2)$.

Como agora o modelo contén máis dunha función de suavizado aparece o que se coñece como problema de identificación: as funcións f_1 e f_2 son estimables salvo unha constante. Isto quere dicir que si f_1 e f_2 son dúas funcións suavizadoras estimadas para o noso modelo, entón $f_1 - c$ e $f_2 + c$ tamén o serán.

Para solucionar este problema empregamos a seguinte restricción sobre unha das funcións de suavizado

$$\sum_{i=1}^{n} f_1(x_1) = 0, \text{ou equivalentemente } 1^T f_1 = 0, .$$

Graos de liberdade efectivos

Procedendo de forma análoga á vista para o caso univariante, considerando agora un modelo aditivo xeralizado

$$g(\mu_i) = A_i\gamma + \sum_j f_j(x_{ji}), \quad y_i \sim EF(\mu_i, \phi) \tag{3.11}$$

onde A_i é a fila i-ésima da matriz de deseño da parte paramétrica do modelo, cos correspondentes parámetros γ, f_j son as funcións de suavizado do vector x_j e $EF(\mu_i, \phi)$ denota a familia exponencial de distribucións con media μ_i e parámetros de escala ϕ.

Se representamos a función de suavizado f nunha base podemos expresar (3.11) como un GLM sobreparametrizado como segue:

$$g(\mu i) = X_i\,\beta, \quad y_i \sim EF(\mu_i, \phi)$$

Deste xeito, en analoxía cos modelos lineais, os **graos de liberdade efectivos** serán

$$edf = tr\left[X\ \left(X^T\,W\,X + \lambda\,S\right)^{-1}\ X^T\,W\right] = tr(A).$$

Ademais, podemos coñecer os graos de liberdade efectivos de cada termo suavizador sen máis que sumar os valores dos termos A_{ii} correspondentes aos coeficientes β_i do suavizador.

Selección automática do parámetro de suavizado

No caso dos modelos aditivos xeralizados (3.11), no canto de ter un único parámetro de suavizado teremos un vector de parámetros de suavizado (unha compoñente por cada función de suavizado do noso modelo) como mostramos a continuación:

$$\lambda = (\lambda_1, \ldots, \lambda_p)$$

Para obter o valor óptimo do parámetro de suavizado, λ_{opt}, minimizaremos as versións multivariantes dos criterios de erro vistos anteriormente como poden ser o GCV e UBRE (consultar (13) e (33) para máis detalles) ou REML (consultar (28) ou (2) para máis detalle).

Algoritmos de aproximación

Buscar a estimación do parámetro de suavizado é a parte computacionalmente máis costosa da estimación dos GAMS, pois necesitamos manter ao mesmo tempo a eficiencia dos métodos e a estabilidade dos mesmos. Chegados a este punto, e de forma similar ao caso univariante, temos tres formas de proceder.

1. Escoller directamente un dos criterios de selección automática multivariante presentados anteriormente. Esta opción require realizar unha iteracción externa en cada paso do algoritmo para opter empregando o PIRLS os coeficientes estimados do modelo correspondentes ao parámetro de suavizado da iteración actual. Para optimizar a rentabilidade deste procedemento, empregaremos o método de Newton na iteración exterior.

2. Aplicar a versión do modelo aditivo gaussiano do método de selección de parámetros de suavizado elixido o método lineal penalizado ao modelo lineal penalizado axustado en cada iteración do PIRLS. Este enfoque é o que se coñece como método PQL e foi proposto por primeira vez por Gu en (11)

3. Empregar o que se coñece como o método xeralizado de Fellner-Schall (ver con máis detalle en (33)), que actualiza os parámetros de suavizado en cada iteración do PIRLS.

As dúas últimas opcións teñen a vantaxe dunha maior simplicidade e eficiencia pero non existe ningunha garantía de converxencia.

3.3. Aplicación a datos biomédicos

Unha vez formulado o modelo aditivo xeralizado que podemos ver na Ecuación (3.11), procederemos a axustar distintos modelos para os datos presentados no Capítulo 1 sobre a presión arterial en función de diversas variables explicativas. Iremos axustando distintos modelos considerando distintas variables explicativas e tratando de ver cales son signiticativas para as dúas variables resposta, presión arterial sistólica e diastólica. Para este axustes empregarase a función `gam` do paquete `mgcv` de R.

3.3.1. Buscamos a mellor distribución para a variable resposta

Como mostramos na Figura 1.1, na que puidemos ver un histograma das dúas variables resposta que imos considerar nos axustes dos modelos aditivos xeralizados, parecen non seguir unha distribución normal. Por este motivo probaremos a axustar un modelo sen covariables para cada unha das variables resposta considerando primeiro que seguen unha distribución normal e posteriormente que seguen a unha distribución Gamma. Calcularemos o AIC e BIC dos modelos para ver con cal obtemos un menor valor.. Amósanse os resultados na Táboa 3.1.

	PA sistólica		**PA diastólica**	
Distribución considerada	AIC	BIC	AIC	BIC
Normal	21910.48	21922.14	19506.82	19518.48
Gamma	21784.08	21795.74	19442.95	19454.62

Táboa 3.1: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os modelos de distribución normal e Gamma. .

Neste caso tanto o AIC e BIC foron calculados axustando un modelo GAM sen variables explicativas. Comprobamos como para ambas variables resposta o valor máis baixo obtense considerando a distribución Gamma. Ademáis, no caso da presión arterial sistólica a baixa do valor do AIC e do BIC é máis que notable (moito máis que no caso da presión arterial sistólica). Polo tanto, en todos os modelos axustados posteriormente incluirase sempre o argumento `family=Gamma(link="log")` na función `gam`. Isto simplemente o facemos a modo orientativo pois realmente os que teñen que seguir unha distribución específica son os erros e non a variable resposta. Coa validación do modelo final ao remate deste capítulo veremos se realmente os modelos axustados son válidos.

3.3.2. Introducimos as variables explicativas idade e sexo

As primeiras variables que imos incluír sera o sexo e a idade, a segunda delas mediante unha función de suavizado e separando o seu efecto por sexos. Estas dúas variables estarán incluidas en todos os modelos posteriores que axustemos. Polo tanto, procedemos a axustar un modelo para cada variable resposta coas seguintes estruturas:

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS} Z_{sexo} + s_1^{PAS}(X_{Idade}) Z_{sexo}, \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD} Z_{sexo} + s_1^{PAD}(X_{Idade}) Z_{sexo}, \end{aligned} \tag{3.12}$$

onde μ é un parámetro da distribución de Y que reflexa a súa media, g é unha función de enlace que neste caso é a función log, o $s_1(\cdot)$ representa a función de suavizado da variable *Idade* tomando como base os P-Splines e separando o efecto por sexos. A variable categórica esta representada por Z_{sexo} e aparece co termo de suavizado da idade para representar a interacción existene entre ambas variables. Ademais, os superíndices PAS e PAD empréganse para diferenciar os coeficientes e funcións suavizadoras dos dous modelos, pois aínda que involucren as mesmas variables, ao traballar con variables resposta diferentes trátase de funcións de suavizado diferente. Unha vez axustado o modelo podemos ver na Figura 3.7 os efectos da función de suavizado da variable *idade* separada por sexos para explicar a PAS.

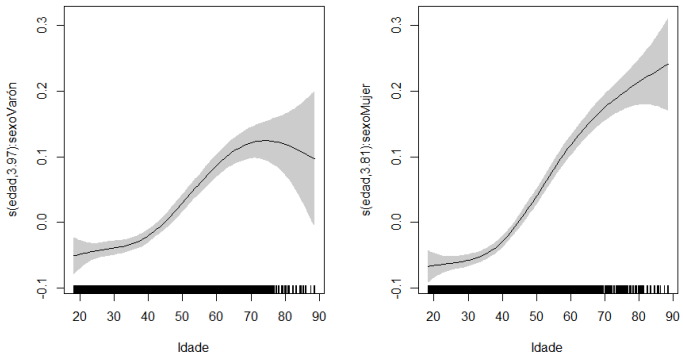


Figura 3.7: Función de suavizado correspondente ao efecto da variable *Idade* separada mediante a variable *sexo* (varón na parte esquerda e muller na parte dereita) tratando de explicar a media condicionada da presión arterial sistólica (a variable *pas*) mediante un modelo aditivo xeralizado considerando como base suavizadora os P-Splines.

Comprobamos efectivamente neste gráfico a necesidade de separar o efecto da idade sobre a variable resposta en función do sexo do individuo considerado, pois o comportamento da función de suavizado é moi diferente nas dúas gráficas. Mentres que no caso dos homes a medida que aumenta a idade aumenta a presión arterial sistólica de forma paulatina pero constante, ata chegar un punto no que

descende pasados os 70 anos; no caso das mulleres a idade non parece ter apenas efecto entre os 18 e os 40 anos (vemos que a curva representada é case recta) pero posteriormente comeza a presentar unha gran subida a partir deste punto de inflexión (que coincide coa premenopausia) e a diferenza dos homes non sufre unha baixada nos últimos anos de vida.

Procedemos agora a realizar o axuste do mesmo modelo pero considerando agora como variable resposta a presión arterial diastólica no canto da sistólica. Temos de novo os efectos da función de suavizado da variable *Idade* separada por sexos para explicar a variable respota *pad* na Figura 3.8.

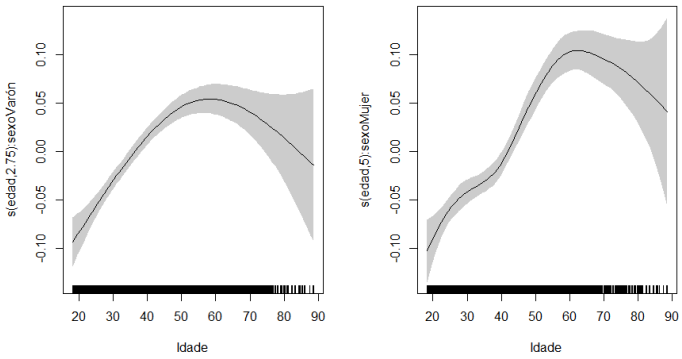
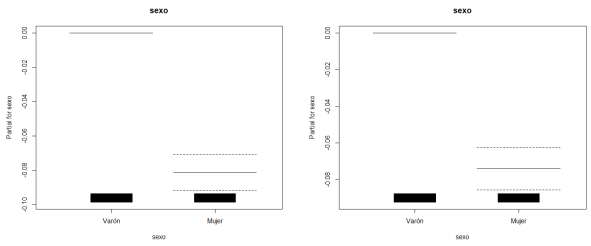


Figura 3.8: Efectos da función de suavizado correspondente a variable *Idade* separada mediante a variable *sexo* (varón na parte esquerda e muller na parte dereita) tratando de explicar a media condicionada da presión arterial diastólica (a variable *pad*) mediante un modelo aditivo xeralizado considerando como base suavizadora os P-Splines.

Neste caso, podemos ver que o comportamento dos homes é moi similar ao visto no caso da presión arterial diastólica, coa diferenza de que non se produce un primeiro tramo onde apenas aumenta coa idade, se non que o aumento é continuo ata chegar ata unha idade aproximada de 60 anos e posteriormente descende paulatinamente. No caso das mulleres nesta gráfica aínda se pode observar de forma máis clara o cambio de comportamento que se produce na función de suavizado unha vez superados os 40 anos, momento a partir do cal a PAD aumenta rapidamente coa idade, chegando a estancarse este efecto chegados os 60 e sufrir un lixeiro descenso nos últimos anos de vida.

Destacar tamén que tanto nas gráficas da Figura 3.7 como da Figura 3.8 temos que ter coidado á hora de extrapolar o comportamento das funcións de suavizado a partir dos 80 anos, pois os datos correspondentes a ese intervalo de idade na nosa mostra son escasos e polo tanto as funcións de suavizado teñen moita variabilidade (como podemos ver na zona sobreada de cor gris nas catro gráficas). Porén, o que si parece claro é que existe unha forte relación entre a idade (separando por sexo) e as dúas presións arteriais medidas no estudo que estamos a considerar.

Mostramos tamén agora na Figura 3.9 unha representación dos coeficientes asociados a cada sexo (realmente unicamente se mostra o coeficiente asociado á muller, pois tomamos como referencia ao home) en cada un dos dous modelos axustados.



(a) PA sistólica (b) PA diastólica

Figura 3.9: Representación dos coeficientes asociados (xunto cos intervalos correspondentes aos constrastes pertinente) á cada un dos grupos nos que se divide a variable *sexo* dentro do modelo aditivo xeralizado axustado tomando como variable resposta a PAS (na esquerda) e a PAD (na dereita) e variables explicativas o sexo e a idade (separada por sexos e mediante unha función de suavizado). O grupo de referencia en ambos casos son os varóns.

Podemos ver como en ambos casos o coeficiente correspondente ás mulleres indica que estas teñen un valor da presión arterial menor que os homes, tanto sistólica como diastólica. Notemos tamén que a diferenza entre sexos é maior no caso da presión arterial sistólica que no da diastólica. Con estes gráficos parece clara a necesidade de incluír esta variable en todos os nosos modelos, tanto de forma independente como para separar o efecto da suavización da variable *idede*.

3.3.3. Incorporamos as variables sociodemográficas

Imos engadir agora ao modelo xa axustado as variables demográficas que nos indican se un individuo vive na costa ou no interior (*int_cost*) e se vive nun entorno urbano ou rural (*urb_rur*). Polo tanto, empregando as mesmas funcións de R que anteriormente, procedemos a axustar os modelos aditivos que presentamos a continuación:

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{urb\_int} + \beta_3^{PAS}\, X_{int\_cost} + s_1^{PAS}(X_{Idade})\, Z_{sexo}, \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{urb\_int} + \beta_3^{PAD}\, X_{int\_cost} + s_1^{PAD}(X_{Idade})\, Z_{sexo}, \end{aligned} \tag{3.13}$$

onde β_0 é o intercepto do modelo, os β_i para $i \in \{1, 2, 3\}$ son os coeficientes das variables categóricas que acompañan e $s_1(\cdot)$ representa as funcións de suavizado (destacar que como este modelo inclúe novas variables, aínda que manteñamos a nomenclatura do modelo (3.12), estas funcións poden ser diferentes).

Unha vez presentado o modelo que se vai axustar, pasaremos a representar algúns dos seus compoñentes e interpretar os valores dos coeficientes obtidos. Posto que os efectos das funcións de suavizado da variable *idade* e os efectos da variable categórica *sexo* apenas varían con respecto ao modelo anterior, non engadiremos eses gráficos e limitarémonos a representar os efectos das novas variables categóricas. Coma sempre, comezaremos polo caso da PAS, vendo os efectos que teñen as novas variables na Figura 3.10.

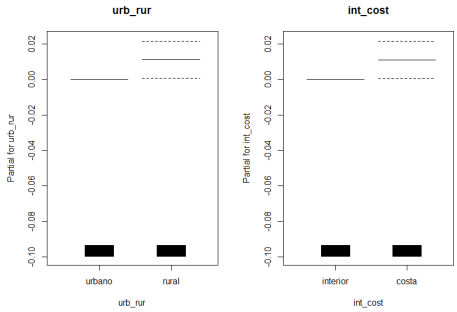


Figura 3.10: Representación dos coeficientes asociados (xunto cos intervalos correspondentes aos constrastes pertinente) á cada un dos grupos nos que se dividen as variable *int_cost* (dereita) e *urb_rur* (esquerda) dentro do modelo aditivo xeralizado axustado tomando como variable resposta a PAS e variables explicativas o sexo e a idade (separada por sexos e mediante unha función de suavizado), ademais das variables xa mencionadas. O grupo de referencia en ambos casos é o que se mostra en primeiro lugar.

Vemos que en ambos casos o efecto das dúas variables parece bastante cuestionable. No caso do coeficiente correspondente á primeira variable representada sae non signficativo mentres que no segundo si é significativo pero cun p-valor bastante alto. Con todo podemos dicir que o feito de vivir no rural ou na costa produce, de xeito discreto, un aumento da PAS.

Mostramos agora na Figura 3.11 as mesmas gráficas sustituíndo agora a PAS por PAD.

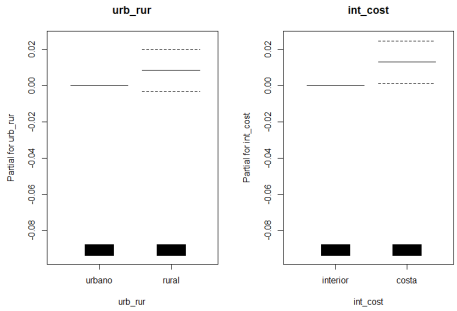


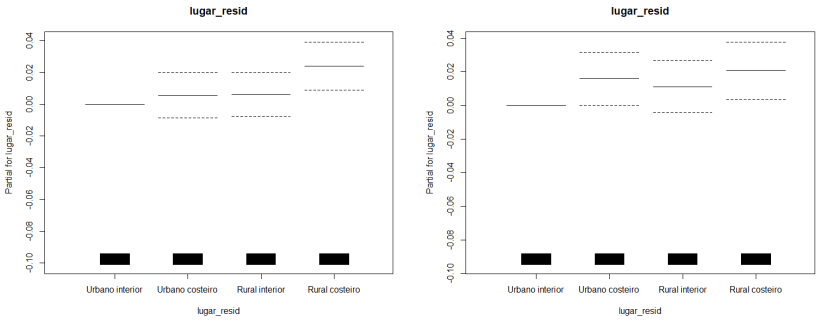
Figura 3.11: Representación dos coeficientes asociados (xunto cos intervalos correspondentes aos constrastes pertinente) á cada un dos grupos nos que se dividen as variable *int_cost* (dereita) e *urb_rur* (esquerda) dentro do modelo aditivo xeralizado axustado tomando como variable resposta a PAD e variables explicativas o sexo e a idade (separada por sexos e mediante unha función de suavizado), ademais das variables xa mencionadas. O grupo de referencia en ambos casos é o que se mostra en primeiro lugar.

Vemos que neste caso o coeficiente correspondente ao grupo rural dentro da variable *urb_rur* é non significativo mentres que o grupo correspondente á costa dentro da variable *int_cost* é lixeiramente significativo. De novo, aqueles individuos que residen na costa teñen unha maior presión arterial diastólica.

Visto que a influencia destas variables que acabamos de introducir nos modelos é bastante baixa, procedemos a sustituílas por unha nova que combine ambas e teña catro categorías. Esta información está recollida na variable *lugar_resid*, polo que modificaremos o modelo (3.13) para axustar agora

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{lugar\_resid} + s_1^{PAS}(X_{Idade})\, Z_{sexo}, \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{lugar\_resi} + s_1^{PAD}(X_{Idade})\, Z_{sexo}. \end{aligned} \tag{3.14}$$

onde β_2 é un vector dos coeficientes correspondentes a tres categorías da variable *lugar_resid* (agás o grupo de referencia, que sería *urbano-interior*) e X_{lugar_resid} representa un vector de tres compoñentes que toma o valor 1 na posición do grupo correspondente ao individuo (e un vector de ceros se o individuo pertence ao grupo de referencia).Unha vez axustado ao modelo procedemos a representar o valor dos coeficientes correspondentes á variable *lugar_resid* obtendo os resultados que se mostran nos gráficos da Figura 3.12.



(a) PA sistólica (b) PA diastólica

Figura 3.12: Representación dos coeficientes asociados (xunto cos intervalos correspondentes aos constrastes pertinente) á cada un dos grupos nos que se divide a variable *lugar_resid* dentro do modelo aditivo xeralizado axustado tomando como variable resposta a PAS (na esquerda) e a PAD (na dereita) e variables explicativas sexo, idade (separada por sexos e mediante unha función de suavizado) e lugar de residencia. O grupo de referencia en ambos casos é o primeiro dos que se mostra.

Vemos en ambos casos que o único coeficiente significativo é o que se corresponde á dupla rural-costa. Isto era esperable pois se ambas características por separado aumentaban de forma lixeira as presións arteriais sistólica e diastólica, as dúas ao mesmo tempo terán un aumento maior. Podemos destacar tamén que o efecto é maior na PAS que na PAD.

Engadimos o nivel de estudos

Inicialmente creamos un modelo considerando unicamente as variable sexo, o efecto suavizado da variable idade separada por sexo e o nivel de estudos, que consta de cinco categorías explicadas no Capítulo 1. Mostramos a continuación a estrutura dos modelos axustados:

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{estud1} + s_1^{PAS}(X_{Idade})\, Z_{sexo}, \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{estud1} + s_1^{PAD}(X_{Idade})\, Z_{sexo}. \end{aligned} \tag{3.15}$$

Aínda que non mostraremos os resultados deste axuste, puidemos ver como no caso do modelo axustado para a PAS existían diferenzas estatísticamente significativas que diferenciasen os grupos III, IV e V do grupo I. Pola contra, no caso do modelo para a PAD, ningún dos coeficientes saíu significativo.

En vista a este resultado, decidimos engadir esta variable ao modelo axustado anteriormente que tamén consideraba o lugar de residencia como variable explicativa. Axustamos un modelo polo tanto como o que segue, anidado co modelo (3.14)

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{lugar\_resid} + \beta_4^{PAS}\, X_{estud1} + s_1^{PAS}(X_{Idade})\, Z_{sexo}, \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{lugar\_resi} + \beta_4^{PAD}\, X_{estud1} + s_1^{PAD}(X_{Idade})\, Z_{sexo}. \end{aligned} \tag{3.16}$$

Os efectos das variables anteriores (sexo, idade e lugar de residencia) apenas se ve modificado por incluír esta nova variable categórica. Mostramos a continuación na Figura 3.13 unha representación gráfica dos coeficientes asociados a cada grupo da variable xunto cos intervalos de confianza para o contraste sobre a diferenza existente contra o grupo de referencia, o grupo I.

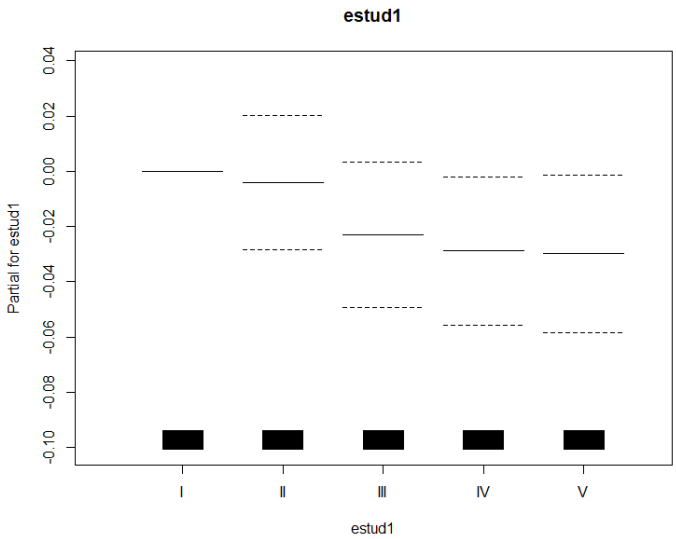


Figura 3.13: Representación dos coeficientes asociados (xunto cos intervalos correspondentes aos constrastes pertinentes) á cada un dos grupos nos que se divide a variable *estud1* dentro do modelo aditivo xeralizado axustado tomando como variable resposta a PAS e variables explicativas sexo, idade (separada por sexos e mediante unha función de suavizado), o lugar de residencia e o nivel de estudos. O grupo de referencia é o primeiro dos que se mostra, o nivel de estudos I.

Vemos claramente como os coeficientes dos grupos III, IV e V diferénzanse considerablemente do do grupo de referencia, mentres que o do grupo II parece ser bastante similar. En vista a este gráfico parece que se pode realizar unha diferenza entre as persoas que teñen un nivel de estudos dos grupos I e II por unha parte e as persoas que teñen un nivel de estudos III, IV e V por outra parte.

Non incluiremos o caso da variable resposta PAD posto que os coeficientes resultaron ser todos non significativos e polo tanto non incluiremos esta variable nos modelos correspondentes á presión arterial diastólica de aquí ata o remate dos modelos GAM.

3.3.4. Sumamos agora o efecto da diabetes

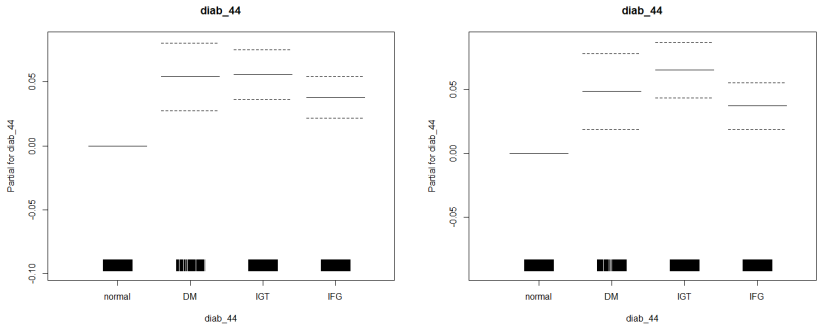
Unha vez consideradas as variables sociodemográficas, podemos incluír tamén unha nova variable categórica referente ao estado da diabetes do paciente. Recordemos que temos dúas posibilidades, unha que estaba dividida en catro grupos e outra en cinco. Neste proceso decidíuse escoller a que esta dividida en catro grupos pois á hora de realizar os axustes (tanto considerando só esa variable como incluíndoa nos modelos anteriores) o AIC e BIC máis baixos se correspondían sempre con *diab_44*.

Por outra banda, aínda que non incluiremos os gráficos correspondentes, cando incluímos a variable referente á diabetes das persoas os efectos das variables categóricas sociodemográficas vólvense non significativos. Por este motivo comprobamos o valor do AIC e o BIC para os modelos que consideraban unicamente as variable sociodemográficas, unicamente a variable *diab_44* e os que consideraban ambas variables. En todos os casos o valor do BIC e AIC máis baixos se correspondían co modelo máis sinxelo, o que só incluía o efecto da nova variable explicativa. Polo tanto decidimos eliminar esas variables do modelo (e non incluílas nos modelos posteriores) e axustar un novo modelo considerando as variables

categóricas de sexo e diabetes xunto co efecto da variable continua idade suavizada separado por sexos, polo que modificamos o modelo (3.16) para axustar agora un modelo aditivo coa seguinte estrutura

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{diab\_44} + s_1^{PAS}(X_{Idade})\, Z_{sexo}, \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{diab\_44} + s_1^{PAD}(X_{Idade})\, Z_{sexo}. \end{aligned} \tag{3.17}$$

Podemos ver unha representación dos coeficientes do vector β_2^{PAS} e β_2^{PAD} correspondentes á variable *diab_44* na Figura 3.14.



(a) PA sistólica (b) PA diastólica

Figura 3.14: Representación dos coeficientes asociados (xunto cos intervalos correspondentes aos constrastes pertinente) á cada un dos grupos nos que se divide a variable *diab_44* dentro do modelo aditivo xeralizado axustado tomando como variable resposta a PAS (na esquerda) e a PAD (na dereita) e variables explicativas sexo e idade (separada por sexos e mediante unha función de suavizado). O grupo de referencia en ambos casos é o primeiro dos que se mostra.

Vemos claramente como en ambos caso pertencer a calquer grupo que non sexa o grupo de referencia (aquel que non padece diabetes nin ningún tipo de prediabetes) provoca un aumento da presión arterial, tanto sistólica como diastólica. Este aumento é maior no caso da sistólica que na diastólica, destacando tamén que para o primeiro caso o grupo con maior presión arterial sería o dos diabéticos (DM) mentres que para o segundo caso o grupo con maior presión arterial sería aqueles que teñen unha prediabetes diagnosticada mediante IGT.

3.3.5. Variables antropomórficas

Temos agora varias variables relacionadas coas medidas antromórficas dos pacientes como son a altura (recollida en *talla*), o peso, a cadeira, a cintura e o IMC. Podemos ver que todas son continuas polo que incluiremos o seu efecto suavizado como xa fixemos coa idade inicialmente. Ademais sabemos que as catro primeiras depende do sexo da persoa, polo que tamén empregaremos esta variable para diferenciar o efecto delas sobre a variable resposta.

O primeiro que se fixo foi introducir cada unha das variables por separado ao modelo axustado en (3.17) e así ver cales producían un efecto sobre as variables respostas. Unha vez comprobado que o efecto da altura non é significativo e en vista a que as outras catro variables están fortemente relacionadas, procedemos a axustar modelos con diferentes combinacións desas catro variables (non incluiremos a estrutura de todos estes modelos posto que simplemente nos quedaremos cun deles, o que teña menor valor do AIC e do BIC) e comprobar cales nos devolven un AIC e BIC máis baixos para continuar engadindo variables. Mostramos os resultados obtidos na Táboa 3.2

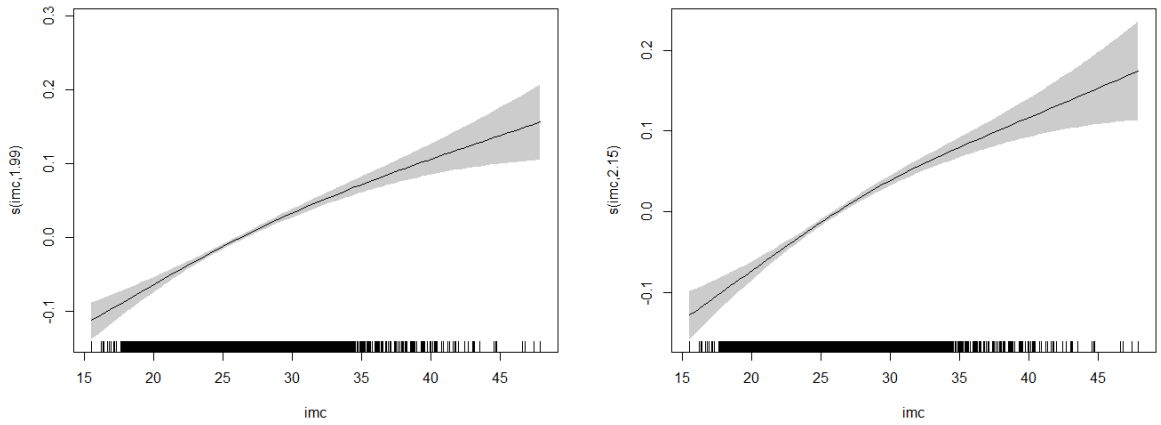
	PA sistólica		**PA diastólica**	
Modelo axustado	AIC	BIC	AIC	BIC
Sexo+s(Idade)+diab_44+s(Peso)	20736.23	20828.62	18716.29	18810.07
Sexo+s(Idade)+diab_44+s(Cadeira)	20850.07	20954.88	18777.59	18884.78
Sexo+s(Idade)+diab_44+s(Cintura)	20805.06	20908.28	18759.02	18877.35
Sexo+s(Idade)+diab_44+s(IMC)	20714.21	20805.13	18712.73	18805.97
Sexo+s(Idade)+diab_44+s(Peso)+s(IMC)	20712.67	20815.53	18701.45	18816.2
Sexo+s(Idade)+diab_44+s(Peso)+s(Cadeira)	20719.28	20829.42	18718.19	18827.81
Sexo+s(Idade)+diab_44+s(Peso)+s(Cintura)	20732.11	20854.29	18708.8	18838.9
Sexo+s(Idade)+diab_44+s(Peso)+s(Cadeira)+s(Cintura)	20714.28	20862.56	18711.63	18857.07

Táboa 3.2: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAM axustados para tratar de explicar a media condicionada das variables resposta PAS e PAD. Na primeira columna indícanse as variables explicativas consideradas en cada un dos modelos, onde $s(\cdot)$ indica unha función de suavizado considerando como base os P-Splines. Todos os efectos suavizadores, agás o do *IMC*, foron separados por sexos.

Vemos así os valores máis baixos do AIC e do BIC se corresponden co modelo que inclúen o efecto suavizado da variable IMC (sin distiguir sexos) e o modelo que inclué ademais o peso mediante unha función suavizadora separando os datos por sexos. Polo tanto imos escoller o primeiro destes modelos por ser o máis simple e ademais tendo en conta que dentro da información do IMC tamén esta recollida en certa medida o peso dunha persoa, sobretodo en relación á súa altura. Por este motivo a partir de agora incluiremos a variable *imc* mediante un efecto de suavizado en todos os modelos que axustemos. Polo tanto os modelos cos que continuaremos a partir de agora serán os que teñen a seguinte estrutura

$$\begin{aligned} g(\mu_{PAS}) &= \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{diab\_44} + s_1^{PAS}(X_{Idade})\, Z_{sexo} + s_2^{PAS}(IMC), \\ g(\mu_{PAD}) &= \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{diab\_44} + s_1^{PAD}(X_{Idade})\, Z_{sexo} + s_2^{PAD}(IMC). \end{aligned} \tag{3.18}$$

onde as funcións s_2^{PAS} e s_2^{PAD} son as funcións de suavizado da variable *IMC* para cada un dos modelos axustados sen separar os datos por sexos, a diferenza do que acontecía co caso das funcións de suavizado da variable *Idade*. Mostramos na Figura 3.15 o efecto suavizado da variable IMC en cada unha das presións arteriais estudadas.



(a) PA sistólica

(b) PA diastólica

Figura 3.15: Efectos da función de suavizado correspondente a variable *IMC* tratando de explicar a media condicionada da presión arterial diastólica (a variable *pad*) mediante un modelo aditivo xeralizado considerando como base suavizadora os P-Splines xunto co efecto das variables categóricas *sexo* e *diab_44* e o efecto suavizado da variable *idade*.

Vemos como as dúas funcións teñen un comportamento similar coa diferenza de que o rango de valores do caso da PAS é maior que o da PAD. Ao aumentar o IMC dun paciente aumentará a presión arterial do mesmo (tanto sistólica como diastólica) pero non o fai de forma lineal. O aumento é moito maior para valores baixos do IMC (entre 15 e 30) mentres que o aumento vai diminuíndo para valores altos do IMC ata case chegar a estabilizarse no caso da PAS.

3.3.6. Incluimos as variables sobre a glucosa e a hemoglobina

Para rematar incluiremos as tres últimas variables da nosa base de datos ao modelo axustado anteriormente. Introduciremos estas variables mediante unha función de suavizado sen facer diferenzas por sexo. Axustaremos distintos modelos con todas as posibles combinacións que se poden facer con estas tres variables, pero non mostraremos o efecto de suavizado de cada un delas, pois posteriormente recolleremos nunha táboa o modelo resultante final para cada variable explicativa.

3.3.7. Resumo dos modelos e discusión final

Unha vez axustados varios modelos considerando distintas variables explicativas dentro da nosa base de datos procedemos a recoller na Táboa 3.3 o valor do AIC e do BIC para cada un deles e así poder tratar de escoller cal é o modelo que mellor se axusta aos nosos datos. Podemos ver en cor verde marcado os valores máis baixos para cada unha das columnas.

	PA sistólica		**PA diastólica**	
Modelo axustado	AIC	BIC	AIC	BIC
Sexo+s(Idade)	20986.59	21049.5	18988.15	19050.85
Sexo+s(Idade)+urb_rur+int_cost	20981.83	21056.41	18985.47	19059.32
Sexo+s(Idade)+lugar_resid	20982.8	21063.33	18987.04	19066.72
Sexo+s(Idade)+estud_1	20983.58	21067.72	18985.2	19070.57
Sexo+s(Idade)+lugar_resid+estud_1	20981.07	21082.84	18985.02	19087.16
Sexo+s(Idade)+diab_44	20939.96	21020.17	18946.16	19026.11
Sexo+s(Idade)+diab_55	20941.17	21027.21	18948.14	19033.91
Sexo+s(Idade)+diab_44++lugar_resid+estud_1	20938.13	21057.42	18946.86	19043.99
Sexo+s(Idade)+diab_44+s(Peso)	20736.23	20828.62	18716.29	18810.07
Sexo+s(Idade)+diab_44+s(Cadeira)	20850.07	20954.88	18777.59	18884.78
Sexo+s(Idade)+diab_44+s(Cintura)	20805.06	20908.28	18759.02	18877.35
Sexo+s(Idade)+diab_44+s(IMC)	20714.21	20805.13	18712.73	18805.97
Sexo+s(Idade)+diab_44+s(Peso)+s(IMC)	20712.67	20815.53	18701.45	18816.2
Sexo+s(Idade)+diab_44+s(Peso)+s(Cadeira)	20719.28	20829.42	18718.19	18827.81
Sexo+s(Idade)+diab_44+s(Peso)+s(Cintura)	20732.11	20854.29	18708.8	18838.9
Sexo+s(Idade)+diab_44+s(Peso)+s(Cadeira)+s(Cintura)	20714.28	20862.56	18711.63	18857.07
Sexo+s(Idade)+diab_44+s(IMC)+s(GLUCOSA)	20704.59	20825.73	18709.24	18825.49
Sexo+s(Idade)+diab_44+s(IMC)+s(G2H)	20697.91	20824.4	18696.38	18795.39
Sexo+s(Idade)+diab_44+s(IMC)+s(GLUCOSA)+s(G2H)	20693.57	20826.88	18700.11	18823.07
Sexo+s(Idade)+diab_44+s(IMC)+s(HBA1C)	20706.88	20818.32	18710.71	18824.07
Sexo+s(Idade)+diab_44+s(IMC)+s(HBA1C)+s(GLUCOSA)	20705.41	20836	18710.55	18829.2
Sexo+s(Idade)+diab_44+s(IMC)+s(HBA1C)+s(G2H)	20694.15	20808.1	18698.15	18804.79
Sexo+s(Idade)+diab_44+s(IMC)+s(HBA1C)+s(GLUCOSA)+s(G2H)	20694.84	20834.75	18699.55	18814

Táboa 3.3: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAM axustados para tratar de explicar a media condicionada das variables resposta PAS e PAD. Na primeira columna indícanse as variables explicativas consideradas en cada un dos modelos, onde $s(\cdot)$ indica unha función de suavizado considerando como base os P-Splines. Todos os efectos suavizadores, agás o do *IMC*, *GLUSCOSA*, *GH2* e *HBA1C*, foron separados por sexos. En cor verde podemos ver marcado o valor máis baixo para cada caso.

Vemos claramente como para o caso da variable resposta *PAD* o modelo que obtén os valores máis

baixos de AIC e BIC é o que presenta a seguinte estrutura

$$g(\mu_{PAD}) = \beta_0^{PAD} + \beta_1^{PAD} Z_{sexo} + \beta_2^{PAD} X_{diab\_44} + s_1^{PAD}(X_{Idade}) Z_{sexo} + s_2^{PAD}(X_{IMC}) + s_3^{PAD}(X_{G2H}), \tag{3.19}$$

onde s_1^{PAD}, s_2^{PAD} e s_3^{PAD} son as función de suavizado do efecto das variables correspondentes sobre a variable resposta PA diastólica. Unha vez escollido o modelo final, representamos as funcións de suavizado para ver o efecto que teñen sobre a variable explicativa como podemos ver nas gráficas da Figura 3.16.

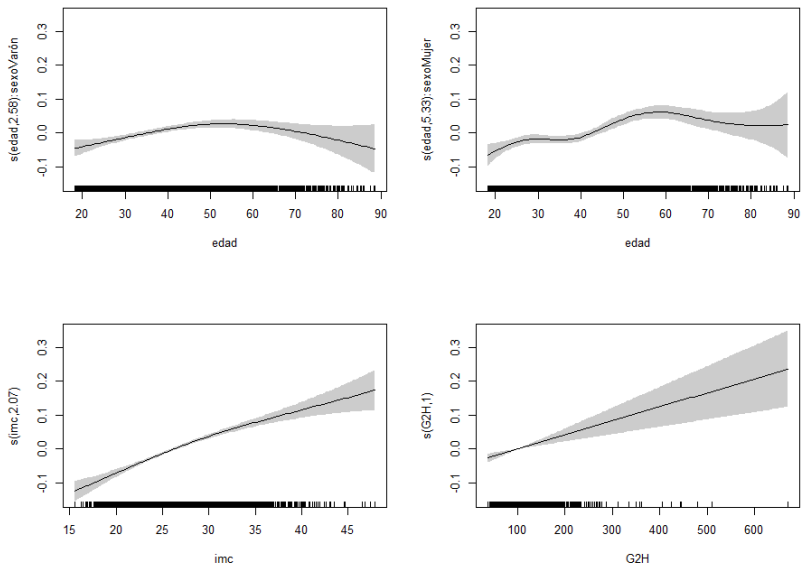


Figura 3.16: Efectos das funcións de suavizado correspondentes as variables *idade*, *imc* e *G2H* tratando de explicar a media condicionada da presión arterial diastólica (a variable *pad*) mediante un modelo aditivo xeralizado considerando como base suavizadora os P-Splines, xunto co efecto das variables categóricas *sexo* e *diab_44*. O efecto da variables *idade* está separado pola variable categórica *sexo*.

Comprobamos así como as dúas primeiras gráficas son moi similares ás vistas para o modelo (3.12) e que podemos ver na Figura 3.8. Para o caso da *Idade* nos homes vemos como a medida que esta aumenta, a PAD vai aumentando tamén de forma paulatina pero que chegada unha idade (aproximadamente os 60 anos), comeza a diminuír. Pola contra, nas mulleres parece manterse máis ou menos estable ata chegar aos 40 anos, onde sofre unha repentina subida en apenas un intervalo de 10 anos para despois volver estabilizarse e incluso descender.

Para o caso do *IMC* vemos que o comportamento deste novo modelo (3.19) é case idéntico ao do modelo (3.16) que podemos ver na Figura 3.15. A media que aumenta o IMC dun individuo tamén aumenta a súa PAD, pero vemos que este aumento é máis acusado para valores máis baixos do IMC que para valores altos.

Por último, vemos o efecto da glucosa en sangue pasadas as dúas horas. Neste caso vemos tamén que a maior cantidade de glucosa en sangue, maior PAD. Notemos que a meirande parte dos datos se atopan antes do valor 300 polo que a estimación do efecto para valores superiores a 300 ten un intervalo de confianza moi grande e pode sufrir moita variabilidade. Pasamos agora a mostrar un resumo do modelo para ver os graos de liberdade correspondentes así como para comprobar se finalmente podemos reducir o efecto de suavizado da variable *G2H* a un efecto lineal.

```
> summary(modelo_pad16)


Family: Gamma
Link function: log

Formula:
pad ~ sexo + s(edad, by = sexo, bs = "ps") + s(imc, bs = "ps") +
diab_44 + s(G2H, bs = "ps")

Parametric coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  4.317849   0.004836 892.886   <2e-16 ***
sexoMujer   -0.059226   0.005701 -10.389   <2e-16 ***
diab_44DM   -0.020337   0.018212  -1.117   0.2642
diab_44IGT   0.010178   0.012241   0.831   0.4058
diab_44IFG   0.016971   0.008708   1.949   0.0514 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                   edf   Ref.df   F    p-value
s(edad):sexoVarón 2.579  3.155  8.971 5.64e-06 ***
s(edad):sexoMujer 5.330  6.121 11.442  < 2e-16 ***
s(imc)            2.065  2.586 88.529  < 2e-16 ***
s(G2H)            1.003  1.006 17.611 2.78e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.255   Deviance explained = 26.4%
GCV = 0.018667  Scale est. = 0.018744  n = 2520
```

Na saída devolta polo `summary` obsérvanse dúas táboas principais. Na primeira, baixo o título *Parametric coefficients* recóllese a estimación do intercepto e os parámetros correspondestes ás variables que teñen un efecto paramétrico, xunto cos seus erros típicos, o estatístico de contraste e o p-valor asociado. Vemos como neste caso o p-valor é menor que $2 \cdot 10^{-16}$ para o caso do intercepto e o coeficiente da variable *sexo*, polo que ambos son significativamente distintos de cero. Para o caso dos coeficientes correspondentes ás categorías, vemos que toman o valor de 0.2642, 0.4058 e 0.0514 respectivamente, polo que ningún resulta significativamente distinto de cero ao 5 % e podemos platexarnos eliminar esta variable do noso modelo.

Por outra banda, a táboa precedida por *Approximate significance of smooth terms* recolle os términos dos efectos suavizados (é dicir, os efectos non paramétricos das variable *idade*, *imc* e *G2H*). Ademais, cada un dos efectos indica os graos de liberdade efectivos da función, así como unha correción dos mesmos (*Red.rf*), o estatístico de contraste así como o p-valor asociado.

Neste sentido, observamos que as curvas asociadas á *idade* separada por sexos, o *imc* e *G2H* teñen 2.579, 5.330, 2.065 e 1.003 graos de liberdade respectivavemente, mentres que os seus p-valores asociados son menores que todos os niveis de significación habituais polo que temos evidencias estatísticamente significativas para rexeitar a hipótese nula de que as covariables non exerzan efecto sobre a variable resposta. Notemos que a curva correspondente a *G2H* parece ter un efecto lineal polo que podemos platexarnos considerar un efecto paramétrico para esta variable.

Finalmente, nas derradeiras liñas do `summary` recollemos o valor do coeficiente de determinación $R^2 = 0.255$, a porcentaxe de deviance explicada, 26.4 %. Vemos que ambos valores son bastante baixos

polo que os modelos axustados non permitirán extraer predicciónss fiables. Finalmente tamén se devolve o valor óptimo (mínimo) do criterio GCV, a varianza do erro e o tamaño mostral.

Procedemos polo tanto a realizar o axuste do modelo (3.19) eliminando a variable *diab_44* e vemos e comparamos o resultado do AIC e BIC de cada modelo que se recolle na Táboa 3.4.

Con	**diab_44**	**Sen**	**diab_44**
AIC	BIC	AIC	BIC
18696.15	18795.08	18694.36	18776.23

Táboa 3.4: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAM axustados para tratar de explicar a media da variable resposta PAD. Nas dúas primeiras columnas esta o modelo correspondente a ecuación (3.19) mentres que nas dúas últimas columnas temos o mesmo modelo quitando a variable *diab_44*.

Vemos que efectivamente en ambos casos obtemos un valor máis baixo para o modelo máis sinxelo, como era de esperar posto que esta variable non tiña coeficientes significativos. Podemos realizar tamén esta comparación mediante un test ANOVA na que consideremos un modelo máis sinxelo (sen a variable *diab_44*) que denominaremos *modelo_pad16B* e o modelo máis complexo (coa variable *diab_44*) que denominaremos *modelo_pad16*. Para iso empregamos a función `anova` de R coa saída que mostramos a continuación.

```
> anova(modelo_pad16B,modelo_pad16,test = "Chisq")

Analysis of Deviance Table

Model 1: pad ~ sexo + s(edad, by = sexo, bs = "ps") + s(imc, bs = "ps") +
s(log(G2H), bs = "ps")
Model 2: pad ~ sexo + s(edad, by = sexo, bs = "ps") + s(imc, bs = "ps") +
diab_44 + s(log(G2H), bs = "ps")
       Resid.Df   Resid. Dev   Df   Deviance Pr(>Chi)
1    2505.1     46.518
2    2502.1     46.442    2.9117 0.075997   0.2435
```

Como vemos, o p-valor devolto é de 0.2435, moi superior a todos os niveis de significación habituais polo que é preferible optar polo modelo máis sinxelo.

Do mesmo xeito que coa variable *diab_44* imos considerar agora un modelo que supoña o efecto paramétrico da variable *G2H*. Recollemos o AIC e BIC dos dous modelos na Táboa 3.5.

Efecto	**paramétrico**	**Efecto non**	**paramétrio**
AIC	BIC	AIC	BIC
18694.36	18776.22	18694.36	18776.23

Táboa 3.5: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAM axustados para tratar de explicar a media da variable resposta PAD. Nas dúas primeiras columnas esta o modelo correspondente a ecuación na que se considera un efecto paramétrico da variable *G2H* e as dúas últimas considerando un efecto non paramétrico.

Vemos que os resultados obtidos son case idénticos polo que parece recomendable obtar por un efecto paramétrico para esta variable. Quedamos así finalmente co modelo seguinte

$$g(\mu_{PAD}) = \beta_0^{PAD} + \beta_1^{PAD}\, Z_{sexo} + \beta_2^{PAD}\, X_{G2H} + s_1^{PAD}(X_{Idade})\, Z_{sexo} + s_2^{PAD}(X_{IMC}). \tag{3.20}$$

Procedemos a validar agora o modelo empregando a función `gam.check` da libraría `gam`. A saída obtida recollémola na Figura 3.17.

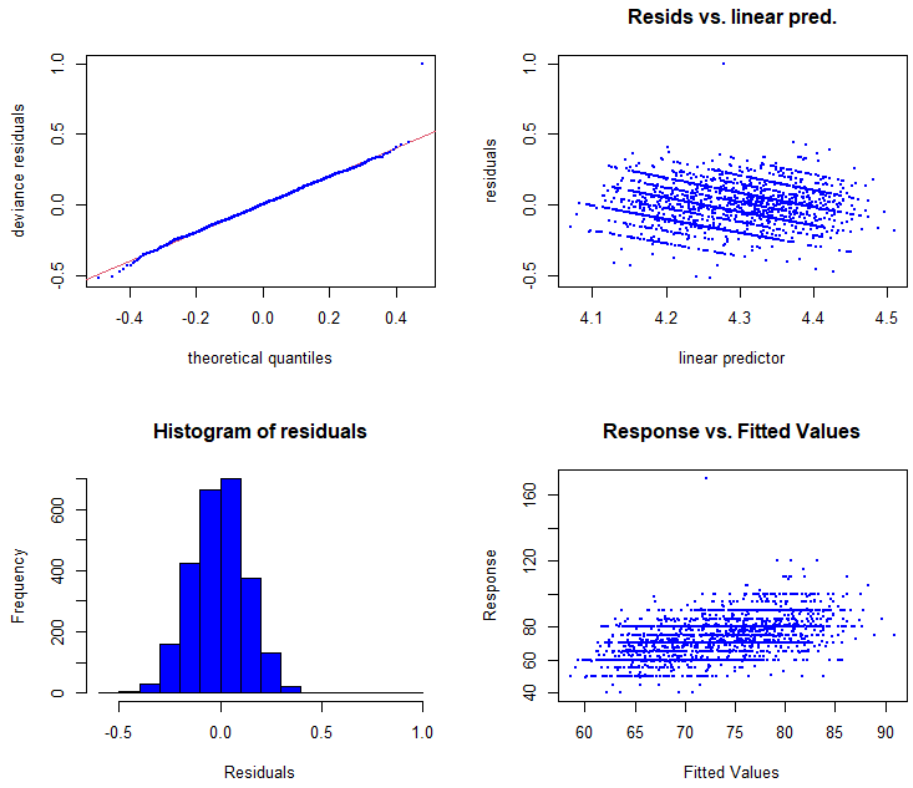


Figura 3.17: Gráficos devoltos pola función `gam.check` da libraría `gam` de R que permiten a validación do modelo axustado en (3.20).

Na primeira gráfica (cadro superior esquerdo) constitúe un QQ-plot dos residuos da deviance. Se os datos se axustan correctamente ao modelo, estes residuos deben situarse próximos a liña de referencia, coincidindo cos seus cuantís esperados. Como vemos, este axuste parece bastante satisfactorio en toda a recta, polo que neste sentido non encontramos ningún motivo para rexeitar o modelo. A misma conclusión pódese sacar do gráfico inmediantamente inferior, que recolle un histograma dos residuos de dito modelo.

Por último, o gráfico da esquina inferior dereita constitúe unha representación gráfica da resposta frente aos valores axustados. Neste caso, debería de observarse unha tendencia lineal nos puntos do gráfico. Se nos fixamos, é certo que se observa unha certa tendencia crecente, aínda que os residuos do modelo son tan grandes que dificultan vela con claridade (recordemos que tanto o coeficiente de determinación como a porcentaxe de deviance explicadas eran moi baixas).

Finalmente, comentar que a función `gam.check`, además dos gráficos xa presentados, devolve a seguinte saída por consola.

```
> gam.check(modelo_pad16C,col="blue",pch = 20,cex = 0.5)
```

```
Method: GCV   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [8.953843e-13,4.719122e-11]
(score 0.0186519 & scale 0.01873501).
Hessian positive definite, eigenvalue range [6.743763e-06,9.997805e-06].
Model rank =  30 / 30

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                    k'  edf    k-index p-value
s(edad):sexoVarón 9.00 2.62    0.99    0.25
s(edad):sexoMujer 9.00 5.32    0.99    0.26
s(imc)            9.00 2.09    1.03    0.94
```

Na primeira columna da táboa anterior (titulada `k'`) recóllese o número de funcións da base que se tomaron para a estimación da curva (e polo tanto, o número máximo de graos de liberdade que se poden empregar na súa construcción), mentres que na segunda columna (`edf`) mostranse os grados de liberdade cos que finalmente se estimaron os efectos non paramétricos. A última columna recolle os p-valores asociados ao contraste de si foron ou non suficientes ditos graos de liberdade para estimar eses efectos non paramétricos. Como nos tres casos se obteñen p-valores elevados. non é necesario ampliar o número de funcións básicas empregadas.

Unha vez visto o caso da variable resposta *PAD*, vemos que para o caso da variable resposta *PAS* o modelo que ten un menor AIC e o que ten un menor BIC non coinciden. Decantarémonos polo modelo do BIC máis baixo, posto que o valor do AIC deste modelo é casi idéntico ao valor máis baixo do AIC.Este modelo é o que presentamos a continuación:

$$g(\mu_{PAS}) = \beta_0^{PAS} + \beta_1^{PAS} Z_{sexo} + \beta_2^{PAS} X_{diab\_44} + s_1^{PAS}(X_{Idade}) Z_{sexo} + s_2^{PAS}(X_{IMC}) + s_4^{PAS}(X_{HBA1C}) + s_4^{PAS}(X_{G2H}), \tag{3.21}$$

onde s_1^{PAS},s_2^{PAS},s_3^{PAD} e s_4^{PAD} son as función de suavizado do efecto das variables correspondentes sobre a variable resposta PA sistólica. Representamos inicialmente as funcións de suavizado das dúas primeiras variables continuas do noso modelo, como podemos ver na Figura 3.18.

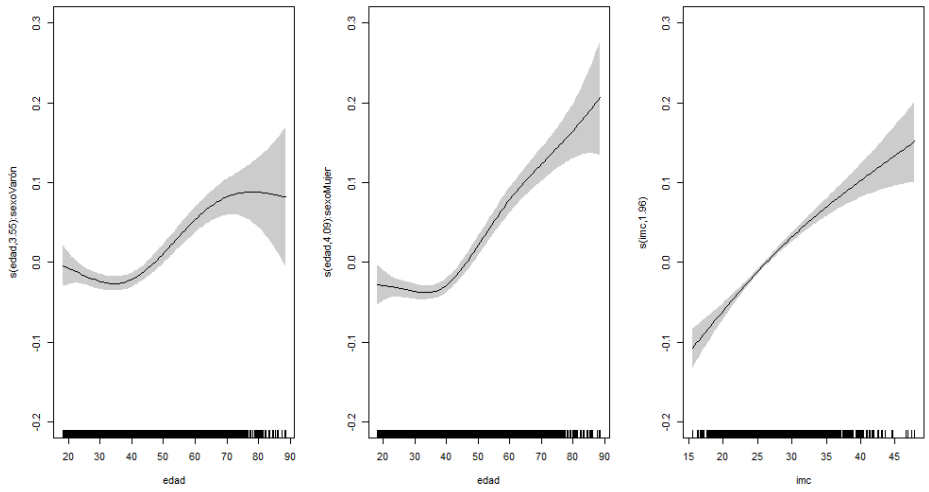


Figura 3.18: Efectos das funcións de suavizado correspondentes as variables *idade* e *imc* tratando de explicar a media condicionada da presión arterial istólica (a variable *pas*) mediante un modelo aditivo xeralizado considerando como base suavizadora os P-Splines xunto co efecto das variables categóricas *sexo* e *diab_44* e o efecto suavizado tamén das variable *G2H* e *HBA1C*. O efecto da variables *idade* está separado pola variable categórica *sexo*.

Vemos que este comportamento é similar ao visto ao longo dos modelos que se foron axustando (os dados por (3.12) e (3.18)) e que podemos ver nas Figura 3.7 e 3.15. Polo tanto non comentaremos os resultados. Procedemos agora a ver as funcións de suavizado das dúas últimas variables continuas da Ecuación (3.21) cuxa representación gráfica podemos ver na Figura 3.19.

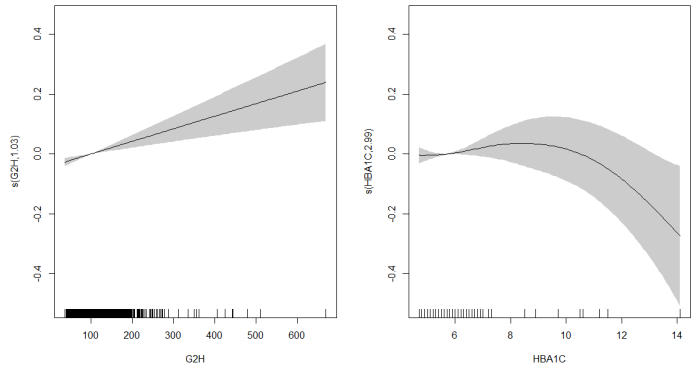


Figura 3.19: Efectos das funcións de suavizado correspondentes as variables *G2H* e *HBA1C* tratando de explicar a media condicionada da presión arterial sistólica (a variable *pas*) mediante un modelo aditivo xeralizado considerando como base suavizadora os P-Splines xunto co efecto das variables categóricas *sexo* e *diab_44* e os efectos suavizados das variable *idade* (separada por sexos) e *imc*.

Pasamos agora a mostrar un resumo do modelo para ver os graos de liberdade correspondentes así como para comprobar se finalmente podemos reducir o efecto de suavizado da variable *G2H* a un efecto lineal.

```
> summary(modelo_pas16)
```

```
Family: Gamma
Link function: log

Formula:
pas ~ sexo + s(edad, by = sexo, bs = "ps") + s(imc, bs = "ps") +
diab_44 + s(G2H, bs = "ps") + s(HBA1C, bs = "ps")

Parametric coefficients:
             Estimate  Std. Error t value   Pr(>|t|)
(Intercept)  4.824981   0.004363 1105.774   <2e-16 ***
sexoMujer   -0.067877   0.005107  -13.290   <2e-16 ***
diab_44DM   -0.016549   0.016636   -0.995    0.320
diab_44IGT   0.003258   0.011514    0.283    0.777
diab_44IFG   0.019694   0.007835    2.513    0.012 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                   edf   Ref.df   F    p-value
s(edad):sexoVarón 3.554  4.259 16.986 < 2e-16 ***
s(edad):sexoMujer 4.092  4.870 45.796 < 2e-16 ***
s(imc)            1.957  2.454 85.640 < 2e-16 ***
s(G2H)            1.006  1.011 15.716 7.6e-05 ***
s(HBA1C)          2.930  3.519  2.077  0.0817 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.346   Deviance explained = 35.9%
GCV = 0.015048  Scale est. = 0.015084  n = 2520
```

Na saída devolta polo `summary` obsérvanse dúas táboas principais. Na primeira, baixo o título *Parametric coefficients* recóllese a estimación do intercepto e os parámetros correspondentes ás variables que teñen un efecto paramétrico, xunto cos seus erros típicos, o estatístico de contraste e o p-valor asociado. Vemos como neste caso o p-valor é menor que $2 \cdot 10^{-16}$ para o caso do intercepto e o coeficiente da variable *sexo*, polo que ambos son significativamente distintos de cero. Para o caso dos coeficientes correspondentes ás categorías, vemos que toman o valor de 0.320, 0.777 e 0.012 respectivamente, polo que só o último resulta significativamente distinto de cero ao 5 % (pero non ao 1 %) polo que podemos plantexarnos considerar un modelo sen esta variable

Por outra banda, a táboa precedida por *Approximate significance of smooth terms* recolle os térmos dos efectos suavizados (é dicir, os efectos non paramétricos das variable *idade*, *imc*, *G2H*) e *HBA1C*. Ademais, cada un dos efectos indica os graos de liberdade efectivos da función, así como unha correción dos mesmos (*Red.rf*), o estatítico de contraste así como o p-valor asociado.

Neste sentido, observamos que as curvas asociadas á *idade* separada por sexos, o *imc*, *G2H* e *HBA1C* teñen 3.554 , 4.092 , 1.957, 1.006 e 2.930 graos de liberdade respectivavemente, mentres que os seus p-valores asociados son todos (ágas o correspondente a *HBA1C*) menores que os niveis de significación habituais polo que temos evidencias estatísticamente significativas para rexeitar a hipótese nula de que as covariables non exercan efecto sobre a variable resposta. Notemos que a curva correspondente a *G2H* parece ter un efecto lineal polo que podemos platexarnos considerar un efecto paramétrico para esta variable.

Finalmente, nas derradeiras liñas do `summary` recollemos o valor do coeficiente de determinación $R^2 = 0.346$, a porcentaxe de deviance explicada, 35.9 %. Vemos que ambos valores son bastante baixos polo que os modelos axustados non permitirán extraer predicciónss fiables. Tamén se devolve o valor

óptimo (mínimo) do criterio GCV, a varianza do erro e o tamaño mostral.

Procedemos polo tanto a realizar o axuste do modelo (3.21) eliminando a variable *diab_44* e vemos e comparamos o resultado do AIC e BIC de cada modelo que se recolle na Táboa 3.6.

Con	**diab_44**	**Sen**	**diab_44**
AIC	BIC	AIC	BIC
20694.15	20808.1	20694.95	20803.16

Táboa 3.6: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAM axustados para tratar de explicar a media da variable resposta PAS. Nas dúas primeiras columnas está o modelo correspondente á ecuación (3.21) mentres que nas dúas últimas columnas temos o mesmo modelo quitando a variable *diab_44*.

Vemos que no caso do AIC obtemos un valor similar para ambos modelos mentres que para o caso do BIC o modelo sen a variable *diab_44* ten un valor máis baixo. Podemos realizar tamén esta comparación mediante un test ANOVA na que consideremos un modelo máis sinxelo (sen a variable *diab_44*) que denominaremos *modelo_pas16B* e o modelo máis complexo (coa variable *diab_44*) que denominaremos *modelo_pas16*. Para iso empregamos a función `anova` de R coa saída que mostramos a continuación.

```
> anova(modelo_pas16B,modelo_pas16,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: pas ~ sexo + s(edad, by = sexo, bs = "ps") + s(imc, bs = "ps") +
s(G2H, bs = "ps") + s(HBA1C, bs = "ps")
Model 2: pas ~ sexo + s(edad, by = sexo, bs = "ps") + s(imc, bs = "ps") +
diab_44 + s(G2H, bs = "ps") + s(HBA1C, bs = "ps")
    Resid. Df Resid. Dev  Df   Deviance Pr(>Chi)
1    2499.3     37.407
2    2498.9     37.366 0.36598 0.041108  0.02768 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Como vemos, o p-valor devolto é de 0.02768, vemos que este test é significativo ao 5 % pero non ao 1 % polo que podemos quedarnos co modelo máis sinxelo que non inclúe á variable diabetes.

Do mesmo xeito que se fixo para o caso do modelo considerando a variable resposta PAD, comprobamos se efectivamente podíamos considerar un efecto paramétrico para a variable explicativa *G2H* cos resultados que podemos ver na Táboa 3.7

Efecto paramétrico		**Efecto non paramétrio**	
AIC	BIC	AIC	BIC
20693.22	20792.35	20694.95	20803.16

Táboa 3.7: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAM axustados para tratar de explicar a media da variable resposta PAD. Nas dúas primeiras columnas está o modelo correspondente a ecuación na que se considera un efecto paramétrico da variable *G2H* e as dúas últimas considerando un efecto non paramétrico.

Neste caso vemos de novo que tanto o valor do AIC como do BIC diminúe ao considerar un efecto paramétrico polo que de novo nos quedamos co modelo máis sinxelo.Ademais, eliminaremos o efecto suavizado da variable *HBA1C* por dous motivos, por unha parte vemos que non parece significativo e por outra banda, se nos fixamos na Figura 3.19, na parte que a curva parece non ser constante, temos moita variabilidade e moi poucos datos, polo que non podemos sacar conclusión moi precisas. Con estes comentarios, chegamos ao modelo que presenta a seguinte estrutura:

$$g(\mu_{PAS}) = \beta_0^{PAS} + \beta_1^{PAS}\, Z_{sexo} + \beta_2^{PAS}\, X_{G2H} + s_1^{PAS}(X_{Idade})\, Z_{sexo} + s_2^{PAS}(X_{IMC}). \tag{3.22}$$

Procedemos a validar agora o modelo empregando a función `gam.check` da libraría `gam`. A saída obtida recollémola na Figura 3.20.

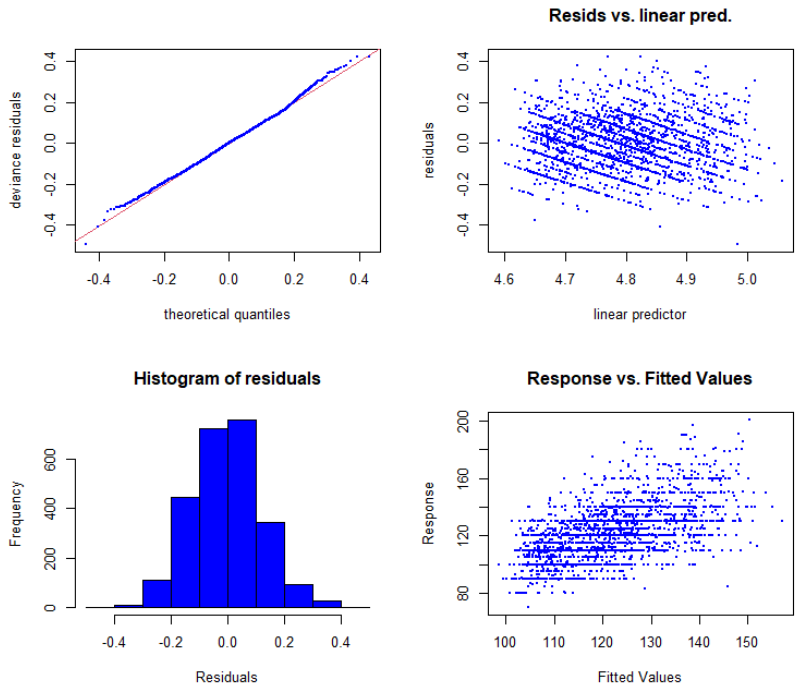


Figura 3.20: Gráficos devoltos pola función `gam.check` da libraría `gam` de R que permiten a validación do modelo axustado en (3.22).

Na primeira gráfica (cadro superior esquerdo) constitúe un QQ-plot dos residuos da deviance. Se os datos se axustan correctamente ao modelo, estes residuos deben situarse próximos a liña de referencia,

coincidindo cos seus cuantís esperados. Como vemos, este axuste parece bastante satisfactorio en toda a recta (agás nos extremos, onde parece que os puntos se separan bastante da recta de referencia), aínda así non encontramos ningún motivo para rexeitar o modelo. A misma conclusión pódese sacar do gráfico inmediantamente inferior, que recolle un histograma dos residuos de dito modelo.

Por último, o gráfico da esquina inferior dereita constitúe unha representación gráfica da resposta frente aos valores axustados. Neste caso, debería de observarse unha tendencia lineal nos puntos do gráfico. Se nos fixamos, é certo que se observa unha certa tendencia crecente, aínda que os residuos do modelo son tan grandes que dificultan vela con claridade (recordemos que tanto o coeficiente de determinación como a porcentaxe de deviance explicadas eran moi baixas).

Finalmente, comentar que a función `gam.check`, además dos gráficos xa presentados, devolve a seguinte saída por consola.

```
> gam.check(modelo_pas16C,col="blue",pch = 20,cex = 0.5)


Method: GCV   Optimizer: outer newton
full convergence after 5 iterations.
Gradient range [6.118757e-12,2.303923e-09]
(score 0.01504276 & scale 0.01509356).
Hessian positive definite, eigenvalue range [3.198997e-06,6.43826e-06].
Model rank =  39 / 39

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                    k'   edf   k-index p-value
s(edad):sexoVarón 9.00 3.61    1.02    0.88
s(edad):sexoMujer 9.00 4.18    1.02    0.85
s(imc)            9.00 1.94    0.98    0.16
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Na primeira columna da táboa anterior (titulada `k'`) recóllese o número de funcións da base que se tomaron para a estimación da curva (e polo tanto, o número máximo de graos de liberdade que se poden empregar na súa construcción), mentres que na segunda columna (`edf`) móstranse os graos de liberdade cos que finalmente se estimaron os efectos non paramétricos. A última columna recolle os p-valores asociados ao contraste de si foron ou non suficientes ditos graos de liberdade para estimar eses efectos non paramétricos. Nos tres casos obteñense p-valores elevados, polo que non é necesario ampliar o número de funcións básicas empregadas.

Capítulo 4

Modelos aditivos xeralizados de localización, escala e forma

Unha vez vistos os modelos aditivos xeralizados no Capítulo 3 que nos permiten modelar o parámetro de localización (μ), xeralizaremos estes modelos para poder tamén modelar un parámetro de escala (ϕ) e para dous parámetros de forma (ν e τ). Damos así lugar aos coñecidos como modelos aditivos xeralizados de localización, escala e forma (GAMLSS). Comezaremos presentando os inicios destes modelos, nos que unicamente se trataba de modelar o parámetro de escala posto que había datos para os cales non era lóxico asumir que este era constante. Posteriormente pasaremos a extender esto para o caso tamén dos parámetros de forma e finalmente aplicaremos os modelos vistos a datos biomédicos empregando o paquete `gamlss` de R.

4.1. GAMLSS

4.1.1. Os seus inicios: modelar o parámetro de escala

A distribución gamma ten dous parámetros: a media, que denotaremos por μ, e o parámetro de escala, que denotaremos por ϕ e que está relacionado coa varianza da variable resposta como indicamos $Var(Y) = \sigma^2\,\phi^2$. Ata agora, unicamente tiñamos formulado un modelo para a explicar μ en función das covariables, pero existen ocasións nas que non é apropiado asumir que o parámetro de escala é constante (como pode ser o caso dos datos da renta de Munich que podemos ver en (30)).

Consideraremos polo tanto o seguinte submodelo dentro dos modelos GAMLLS

$$\begin{aligned} Y &\sim EF(\mu,\phi) \\ \eta_1 &= g_1(\mu) = X_1\,\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \\ \eta_2 &= g_2(\phi) = X_2\,\beta_{2k} + \sum_{j=1}^{J_2} h_{j2}(x_{j2}) \end{aligned} \tag{4.1}$$

onde $EF(\mu,\phi)$ denota calquer distribución de dous parámetros e ambos parámetros son función das variables explicativas (funcións lineais, funcións suavizadas ou ambas).

4.1.2. Modelo aditivos xeralizados de localización, escala e forma (GAMLSS).

Un dos principais problemas das distribucións de dous parámetros é o feito de que a asimetría e a kurtosis da distribución están fixadas para μ e ϕ fixos. Por este motivo, será tamén de utilidade

axustar modelos que nos permitan estruturas máis flexibles para a asimetría e a kurtosis. Chegamos así aos Modelo Aditivos Xeralizados de Localización, Escala e Forma (GAMLSS), que non será máis que unha extensión do modelo (4.1) como segue:

$$\begin{aligned} Y &\sim EF(\mu, \phi, \nu, \tau) \\ \eta_1 &= g_1(\mu) = X_1\,\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \\ \eta_2 &= g_2(\phi) = X_2\,\beta_{2k} + \sum_{j=1}^{J_2} h_{j2}(x_{j2}) \\ \eta_3 &= g_3(\nu) = X_3\,\beta_{3k} + \sum_{j=1}^{J_3} h_{j3}(x_{j3}) \\ \eta_4 &= g_4(\tau) = X_4\,\beta_{4k} + \sum_{j=1}^{J_4} h_{j4}(x_{j4}) \end{aligned} \tag{4.2}$$

onde $EF(\mu, \phi, \nu, \tau)$ é unha distribución de catro parámetros onde ν e τ son parámetros de forma que habitualmente están relacionados coa asimetría e a kurtosis da distribución. Desta forma o modelo (4.2) define un modelo aditivo xeralizado de localización, escala e forma introducido por primeira vez por Rigby e Stasinopoulos en (27).

Axuste do modelo

Un modelo GAMLSS paramétrico (é dicir, un modelo (4.2) sen funcións de suavizado) axústase mediante estimacións por máxima verosimilitude. Tendo en conta que $Y \sim EF(\mu, \phi, \nu, \tau)$, a función de máxima verosimilitude vén dada por

$$L(\mu, \phi, \nu, \tau) = \prod_{i=1}^{n} f(y_i|\mu_i, \phi_i, \nu_i, \tau_i)$$

Polo tanto, a función de log-verosimilitude será a seguinte

$$\ell = \sum_{i=1}^{n} \log\, f(y_i|\mu_i, \phi_i, \nu_i, \tau_i)$$

No caso de que o modelo (4.2) si que teña funcións suavizadoras, o axuste do modelo realizarase considerando a estimación de parámetros mediante máxima verosimilitude penalizada. No capítulo 9 de (30) podemos ver como a maioría das funcións de suavizado empregadas nos modelos GAMLSS se poden expresar como $h(x) = Z\,\gamma$, onde Z é unha matriz de bases evaluada en x e γ é un conxunto de coeficientes da penalización cuadrática do modelo $\gamma^T\,G(\lambda)\,\gamma$ onde λ é un vector ou escalar de hiperparámetros. Rigby an Stasinopoulos en (27) mostraron que o algoritmo empregado para axustar o modelo GAMLSS para valores fixos dos hiperparámetros λ maximiza a función de verosimilitude penalizada dada por

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \gamma_{kj}^T\,G_{kj}(\lambda_{kj})\,\gamma_{kj} \tag{4.3}$$

Chegados ata aquí propuxeron dous algoritmos para axustar o modelo GAMLSS (podemos ver estes algoritmos explicados de forma máis detallada no capítulo 3 de (30)):

- O primeiro deles, o algoritmo CG, é unha xeralización do algoritmo de Cole and Green (que podemos ver en [3]) que emprega as primeiras derivadas e os valores esperados das segundas (exactos ou paroximados) e derivadas cruzadas da función de verosimilitude respecto de $\theta = (\mu, \phi, \nu, \tau)$.

- O segundo algoritmo é o algoritmo RS, que non é máis que unha xeralización do empregado por Rigby e Stasinopoulos en [25] e [26] para axustar modelos MADAM (Modelos Aditivos de Dispersión e Media). Este modelo é máis adecuado que o anterior pois non emprega o valor esperado das derivadas cruzadas, que en determinadas situacións pode ser nulo e provocar problemas de converxencia no algoritmo anterior.

Deste xeito vemos que estes novos modelos proporcionan unha gran flexibilidade para modelos de regresión univariantes, permite calquer distribución para a variable resposta e todos os parámetros desta distribución poden ser modelados como función das variables explicativas. Ademais, extende os modelos estatísticos básicos permitindo unha gran flexibilidade ao poder modelar a sobredispersión, os ceros inflados e a asimetría e kurtosis dos datos.

Os Modelos Aditivos Xeralizados de Localización, Escala e Forma (GAMLSS, *Generalized Additive Model for Location, Scale and Shape*) é unha clase xeral de modelos para unha resposta Y univariante.

Neste tipo de modelos considérase unha resposta Y cunha distribución de parámetros (μ, σ, ν, τ) onde μ se corresponde co parámetro de localización, σ co parámetro de escala e ν e τ cos parámetros de forma.

Se denotamos os parámetros mencionados anteriormente como θ_k con $k = \{1, 2, 3, 4\}$. Deste xeito un GAMLSS é un modelo de regresión semiparamétrico de resposta transformada, θ_k, que se compón de 4 submodelos de regresión GAM da forma

$$g_k(\theta_k) = \eta_k = X_k\,\beta_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk})$$

Deste xeito, con este tipo de modelos, ademais de poder modelar a media da variable resposta, tamén podemos modelar outros parámetros da distribución en función de efectos paramétricos ou non paramétricos das variables explicativas e de posibles efectos aleatorios. Para implementar estes modelos pordemos empregar o paquete `gamlss` xunto con todas as súas extensións.

4.2. Aplicación a datos biomédicos

Unha vez visto a estrutura dos modelos GAMLSS e os algoritmos que se empregan para aproximar os seus parámetros procedemos a axustar dous modelos GAMLSS para tratar de explicar as nosas variables resposta *pas* e *pad* que recollemos no 1 como xa fixemos para o caso dos modelos GAM.

Podemos ver nas ecuacións (4.2) que a nosa variable resposta ten que pertencer a unha distribución que se encontre dentro da familia exponecial. En vista aos resultados obtidos nos axustes dos modelos GAM podemos pensar que o noso parámetro de escala ϕ non é constante e polo tanto tamén depende das covariables, polo que imos axustar novos modelos.

Procedemos agora a tratar de buscar o mellor modelo para explicar as variables resposta consideradas en función das variables explicativas da base de datos do Capítulo 1. Operaremos dunha forma similar ao feito no Capítulo 3. Iremos engandindo variable pouco a pouco e veremos que modelos teñen un menor valor do AIC e do BIC, neste caso teremos que traballar cos parámetros de localización, μ, e escala, σ.

Mostraremos a forma de proceder considerando como variable resposta a variable *pad* e unicamente coas primeiras variables que engadiremos, despois mostraremos só os modelos finais para non estar repetindo sempre o mesmo procedemento. Iniciamos engadindo o efecto da variable categórica *sexo* e

o efecto non paramétrico da *idade*[1] . Consideramos así axustar os seguintes tres modelos:

$$\begin{aligned} Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\ \eta_{1\,PAD} &= g_1(\mu_{PAD}) = \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade}) \\ \eta_{2\,PAD} &= g_2(\phi_{PAD}) = \beta_{0\,\phi}^{PAD} \end{aligned} \tag{4.4}$$

$$\begin{aligned} Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\ \eta_{1\,PAD} &= g_1(\mu_{PAD}) = \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade}) \\ \eta_{2\,PAD} &= g_2(\phi_{PAD}) = \beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD} X_{sexo} \end{aligned} \tag{4.5}$$

$$\begin{aligned} Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\ \eta_{1\,PAD} &= g_1(\mu_{PAD}) = \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade}) \\ \eta_{2\,PAD} &= g_2(\phi_{PAD}) = \beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD} X_{sexo} + s_{1\,\phi}^{PAD}(X_{Idade}) \end{aligned} \tag{4.6}$$

Unha vez axustados estes tres modelos calculamos o AIC e BIC de cada un para comprobar cal ten un valor máis baixo. Mostramos os resultados na Táboa 4.1.

Expresión μ	**Expresión ϕ**	**AIC**	**BIC**
$\beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade})$	$\beta_{0\,\phi}^{PAD}$	19003.56	19049.19
$\beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade})$	$\beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD} X_{sexo}$	19004.96	19056.27
$\beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade})$	$\beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD} X_{sexo} + s_{1\,\phi}^{PAD}(X_{Idade})$	19005.59	19062.96

Táboa 4.1: Valor do AIC (*Criterio de Información de Akaike*) e do BIC (*Criterio de Información Bayesiano*) para os distintos modelos GAMLSS axustados para tratar de explicar a media (μ) e a escala (ϕ) da distribución da variable resposta PAD. Na primeira columna está a estrutura que empregamos para explicar a media e na segunda columna a estrutura que empregamos para explicar a escala.

Vemos así como neste caso o mellor modelo é aquel que considera que o parámetro de escala é constante. A partir de aquí, procedemos da mesma forma co resto de variables explicativas que ten a nosa base de datos e dunha forma análoga ao caso dos modelos GAM. Iremos introducindo sucesivamente variables (tanto na expresión da media como na expresión da escala) e iremos comparando cal é o mellor modelo. En cada etapa quedámonos co mellor modelo e despois será a este modelo ao que lle engadamos as seguintes variables. Chegamos deste xeito ao seguinte modelo final:

$$\begin{aligned} Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\ \eta_{1\,PAD} = g_1(\mu_{PAD}) &= \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + \beta_{2\,\mu}^{PAD} X_{lugar\_resid} + \beta_{3\,\mu}^{PAD} X_{estud1} + \beta_{4\,\mu}^{PAD} X_{diab\_44} \\ &\quad + s_{1\,\mu}^{PAD}(X_{Idade}) + s_{2\,\mu}^{PAD}(X_{imc}) + s_{3\,\mu}^{PAD}(X_{peso}) + s_{4\,\mu}^{PAD}(X_{G2H}) \\ \eta_{2\,PAD} = g_2(\phi_{PAD}) &= \beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD} X_{lugar\_resid} + s_{1\,\phi}^{PAD}(X_{imc}) \end{aligned} \tag{4.7}$$

[1]Neste caso non se separa este efecto por sexos pois a función `gamlss` do paquete `gamlss` de R non permite esta opción.

Unha vez escollido o modelo final fíxose un resumo do mesmo no que se observaba como os coeficientes das variables categóricas *lugar_resid*, *estud1* e *diab_44* non eran significativos polo que se axustaron modelos sen elas para ver se íamos obtendo valores do AIC e BIC máis baixos e finalmente obtouse polo modelo que se mostra a continuación.

$$\begin{aligned} Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\ \eta_{1PAD} = g_1(\mu_{PAD}) &= \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + s_{1\,\mu}^{PAD}(X_{Idade}) + s_{2\,\mu}^{PAD}(X_{imc}) + s_{3\,\mu}^{PAD}(X_{peso}) + s_{4\,\mu}^{PAD}(X_{G2H}) \\ \eta_{2PAD} = g_2(\phi_{PAD}) &= \beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD} X_{lugar\_resid} + s_{1\,\phi}^{PAD}(X_{imc}) \end{aligned} \tag{4.8}$$

Mostramos agora un resumo deste modelo empregando a función `summary` de R.

```
> summary(modeloD46)


Family:  c("GA", "Gamma")

Call:  gamlss(formula = pad ~ sexo + pb(edad, by = sexo) +pb(imc) + pb(peso, by = sexo)
+ pb(G2H), sigma.formula = ~lugar_resid + pb(imc), family = GA, data = GAL2)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)          3.914e+00  1.773e-02 220.696  < 2e-16 ***
sexoMujer           -4.081e-02  7.600e-03  -5.370 8.60e-08 ***
pb(edad, by = sexo)  1.925e-03  2.192e-04   8.780  < 2e-16 ***
pb(imc)              5.912e-03  1.361e-03   4.344 1.46e-05 ***
pb(peso, by = sexo)  1.754e-03  4.633e-04   3.785 0.000157 ***
pb(G2H)              3.270e-04  6.756e-05   4.840 1.38e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
                           Estimate  Std. Error t value Pr(>|t|)
(Intercept)               -1.794937   0.083356 -21.533  < 2e-16 ***
lugar_residUrbano costeiro -0.080277   0.038309  -2.096 0.036226 *
lugar_residRural interior  -0.044445   0.037319  -1.191 0.233797
lugar_residRural costeiro  -0.140462   0.041099  -3.418 0.000642 ***
pb(imc)                    -0.005596   0.003019  -1.854 0.063872 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

------------------------------------------------------------------
NOTE: Additive smoothing terms exist in the formulas:
i) Std. Error for smoothers are for the linear effect only.
ii) Std. Error for the linear terms maybe are not accurate.
------------------------------------------------------------------
No. of observations in the fit:  2520
```

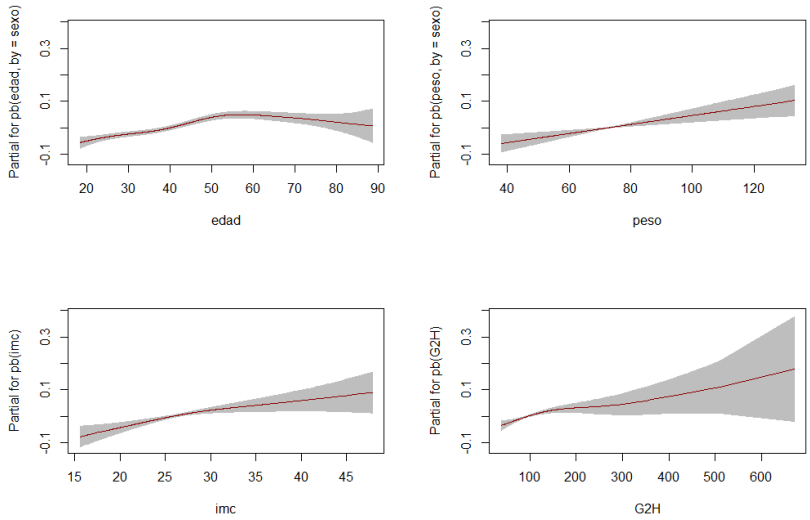
```
Degrees of Freedom for the fit:  17.81656
Residual Deg. of Freedom:  2502.183
at cycle:  4

Global Deviance:     18650.19
AIC:     18685.82
SBC:     18789.73
*****************************************************************
```

Na saída devolta polo `summary` obsérvanse dúas táboas principais. Na primeira, recóllense os coeficientes que afectan ao parámetro de localización (μ). Podemos ver catro columnas, na primeira delas vemos unha estimación do parámetro (no caso dos efectos suavizados únicamene fai referencia ao efecto lineal), xunto cos seus erros típicos, o estatístico de contraste e o p-valor asociado. Vemos como todos os p-valores son máis baixos que os niveis de significación habituais polo que temos evidencias estatísticamente significativas para dicir que son distintos de cero.

Na segunda táboa recóllense os coeficientes que afectan ao parámetro de escala (ϕ) xunto coas mesmas 4 columnas que para o parámetro anterior. Neste caso vemos que existen p-valores máis altos que algúns dos niveis de significación habituais. Aínda así, axustáronse modelos sen esas variables e os AIC e BIC devoltos eran máis altos polo que continuamos considerando este modelo.

Na parte final da saída podemos ver o número de observacións do modelo, os graos de liberdade estimados (17.81656), os graos de liberdade dos residuos, así como a deviance global (18650.19) ou o AIC (18685.82).

Unha vez visto o resumo do modelo podemos representar gráficamente o efecto das funcións suavizadoras das covariables para o parámetro de localización sen máis que empregar a función `term.plot` do paquete `gamlss` de R. Vemos os gráficos devoltos na Figura 4.1.



(a) Idade e IMC (b) Peso e G2H

Figura 4.1: Efectos da función de suavizado correspondentes ás variables *idade*, *peso*, *imc* e *G2H* tratando de explicar o parámetro de localización (μ) da distribución da familia exponencial que segue á variable resposta *pad* mediante un modelo aditivo xeralizado de localización, escala e forma (GAMLSS) considerando como base suavizadora os P-Splines xunto co efecto da variable categórica *sexo*.

Podemos ver na primeira gráfica (arriba á esquerda) como o comportamento para a variable *idade* é moi similar ao visto nos modelos GAM do Capítulo 3. Inicialmente aumenta de forma paulatina ata chegar a estancarse antes dos 40 e despois subir de forma rápida e rematar baixando ao final. Notemos que neste caso non podemos ver tan claramente o efecto da menopausia, posto que a variable non está separada por sexos (probablemente se o separásemos a subida entorno aos 40 para as mulleres sería moito maior).

Na segunda gráfica vemos que o efecto do *peso* parece case lineal. Por este motivo decidíuse axustar un modelo no que considerásemos o efecto paramétrico do *peso* e efectivamente obtivemos mellores resultados en termos do AIC e do BIC polo que se optou por este modelo. Despois na terceira gráfica (abaixo á esquerda) vemos o efecto do *imc* que aumenta rapidamente para valores moi baixos e despois ralentiza este aumento.

Por último temos a gráfica correspondente a *G2H* na que vemos un forte aumento para valores baixos (dende case 0 ata 200) e despois continua un ascenso pero cunha gran variabilidade debido á falta de datos nese rango de valores. Isto probablemente sexa debido a que eses valores de glucosa en sangue tan altos non son habituais e só se poden dar en persoas diabéticas non diagnosticadas (como efectivamente se comproba se miramos na base de datos).

Unha vez vistas as gráficas correspondentes aos efectos suavizadores para o parámetro de localización (μ) podemos facer o mesmo para o caso do parámetro de escala (ϕ) sen máis que empregar o argumento `what="sigma"` da función `term.plot` empregada antes. Podemos ver os resultados obtidos na Figura 4.2.

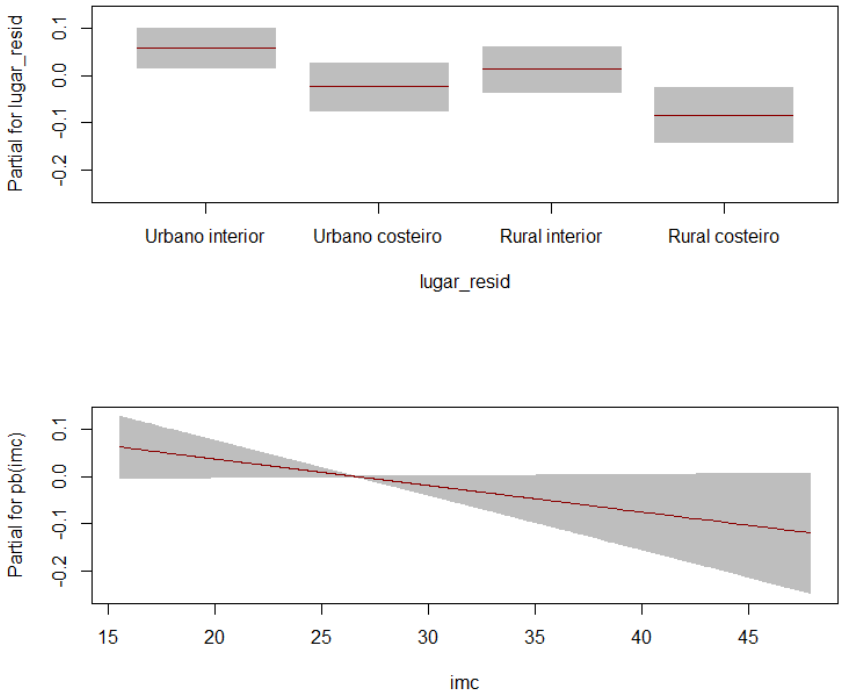


Figura 4.2: Efectos da variable *lugar_resid* e da función de suavizado correspondente á variable *imc* tratando de explicar o parámetro de escala (ϕ) da distribución da familia exponencial que segue a variable resposta *pad* mediante un modelo aditivo xeralizado de localización, escala e forma (GAMLSS) considerando como base suavizadora os P-Splines.

Vemos claramente na gráfica superior como parece que pertencer ao grupo urbano interior aumenta o valor do parámetro de escala mentres que pertencer ao grupo rural costeiro provoca unha diminución deste parámetro. Vemos tamén como o *imc* parece ter un efecto lineal (a maior imc, menor valor do parámetro de escala) polo que axustamos tamén un modelo no que supoñamos o efecto paramétrico desta variable. Destaquemos tamén que para valores extremadamente altos do *imc* a variabilidade é moi grande polo que non podemos sacar conclusións de forma clara.

Deste xeito, o modelo co que finalmente nos quedamos é o que se mostra:

$$
\begin{aligned}
Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\
\eta_1^{PAD} = g_1(\mu_{PAD}) &= \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD}\, X_{sexo} + \beta_{3\,\mu}^{PAD}\, X_{peso} + s_{1\,\mu}^{PAD}(X_{Idade}) + s_{2\,\mu}^{PAD}(X_{imc}) + s_{3\,\mu}^{PAD}(X_{G2H}), \\
\eta_2^{PAD} = g_2(\phi_{PAD}) &= \beta_{0\,\phi}^{PAD} + \beta_{1\,\phi}^{PAD}\, X_{lugar\_resid} + \beta_{2\,\phi}\, X_{imc}.
\end{aligned}
\tag{4.9}
$$

Co modelo (4.6) xa axustado podemos facer unha diagnose dos residuos como se fixo no caso dos modelos GAM. Para isto empregamos o comando *plot* de R sobre o obxecto no que temos gardado o modelo e obtemos os gráficos que podemos ver na Figura 4.3.

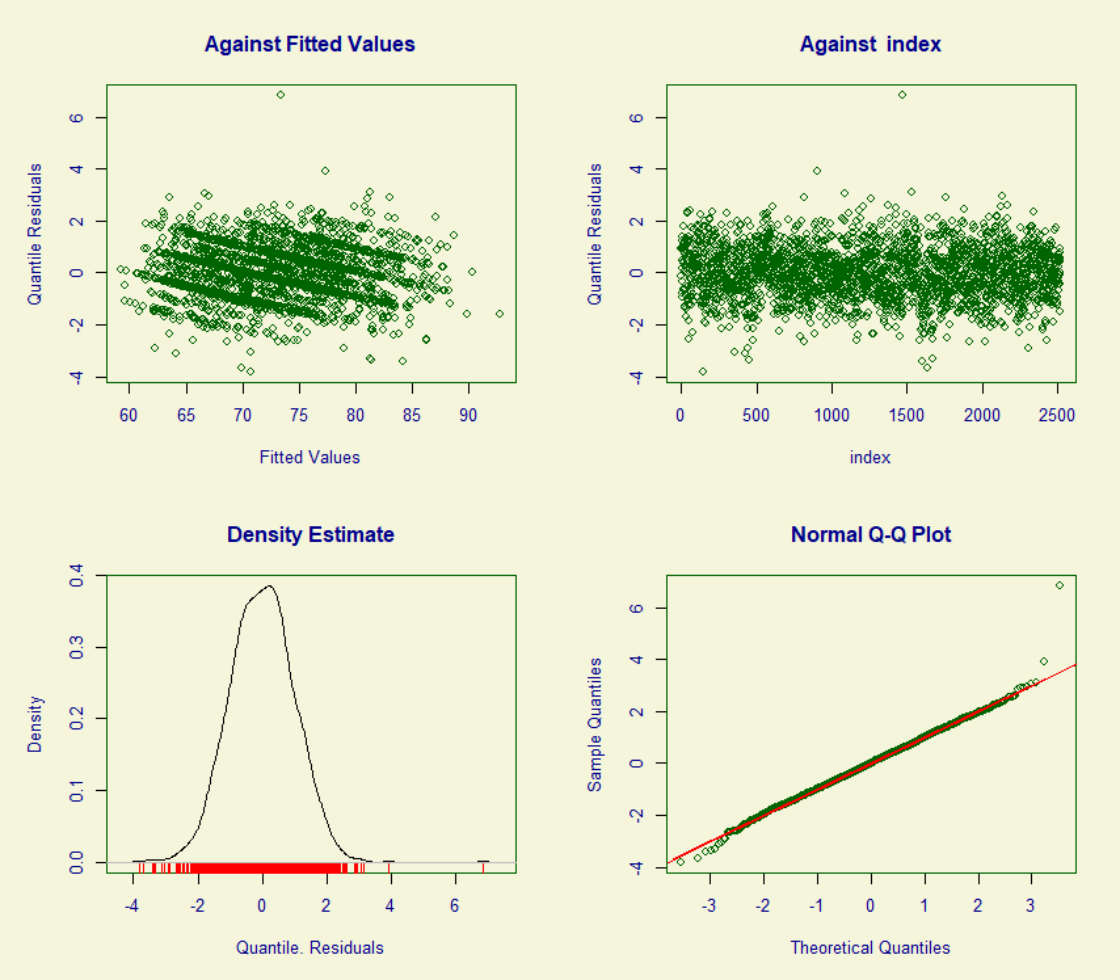


Figura 4.3: Gráficos devoltos pola función `plot` de R executada sobre o modelo axustado coa función `gamlss` do paquete `gamlss` que permiten a validación do modelo axustado en (4.9).

Nestes gráficos podemos ver que os residuos se comportan adecuadamente (aínda que poida apreciarse algo de tendencia no primeiro dos gráficos). Nos dous gráficos superiores móstranse os residuos fronte aos valores axustados de μ e fronte aos índices dos datos. En ambos casos parece verse unha dispersión aleatoria dos residuos entorno ao valor 0. Notemos que tamén parece haber un residuo cun valor excesivamente alto. Nos dous gráficos inferiores podemos ver unha estimación da densidade tipo núcleo na parte esquerda e un QQ-plot na parte dereita. Ambos gráficos parecen mostrar un comportamento adecuado, destacando que no segundo os puntos dos cuantís se sitúan sobre a recta vermella de forma case perfecta agás nos extremos.

Ademais destes gráficos, a función `plot` tamén imprime a seguinte información por consola.

```
plot(modeloD48)
Summary of the Quantile Residuals
mean   =  -0.001709973
variance   =  1.000092
coef. of skewness  =  0.0831356
coef. of kurtosis  =  3.847989
Filliben correlation coefficient  =  0.9973563
```

Aquí vemos un resumo dos cuantís dos residuos. Vemos que a súa media está próxima a cero (-0.001709973) e a súa varianza próxima a un (1.000092). Por outra parte o coeficiente de asimetría sepárase lixeiramente de cero (0.0831356) e o seu coeficiente e kurtosis tamén se separa considerablemente de tres (3.847989). Polo que debemos de dubidar de se a distribución dos nosos residuos será unha normal estándar.

Unha vez rematado o estudo para a variable *pad* procedemos a mostrar o modelo final seleccionado para a presión arterial sistólica. Procedendo de forma análoga chegamos a un modelo similar ao visto para o caso da variable resposta *pad* e novamente tiñamos coeficientes que non eran significativos polo que se foron eliminando ata chegar ao modelo que presentamos a continuación:

$$\begin{aligned} Y_{PAS} &\sim EF(\mu_{PAS}, \phi_{PAS}) \\ \eta_{1\,PAS} = g_1(\mu_{PAS}) &= \beta^{PAS}_{0\,\mu} + \beta^{PAS}_{1\,\mu} X_{sexo} + s^{PAS}_{1\,\mu}(X_{Idade}) + s^{PAS}_{2\,\mu}(X_{imc}) + s^{PAS}_{3\,\mu}(X_{peso}) \\ &\quad + s^{PAS}_{4\,\mu}(X_{cadeira}) + s^{PAS}_{5\,\mu}(X_{G2H}) \\ \eta_{2\,PAS} = g_2(\phi_{PAS}) &= \beta^{PAS}_{0\,\phi} + \beta^{PAS}_{1\,\phi} X_{diab\_44} + s^{PAS}_{1\,\phi}(X_{idade}) \end{aligned} \tag{4.10}$$

Mostramos agora un resumo deste modelo empregandoa función `summary` de R.

```
> summary(modeloS49)

Family:  c("GA", "Gamma")

Call:  gamlss(formula = pas ~ sexo + pb(edad) + pb(imc) +
pb(cadeira, by = sexo) + pb(peso, by = sexo) +
pb(G2H), sigma.formula = ~pb(edad) + diab_44, family = GA,      data = GAL2)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.534e+00  3.179e-02 142.618  < 2e-16 ***
sexoMujer             -4.875e-02  6.984e-03  -6.980 3.76e-12 ***
```

```
pb(edad)                3.121e-03  2.086e-04  14.967  < 2e-16 ***
pb(imc)                 6.875e-03  1.282e-03   5.364 8.89e-08 ***
pb(cadeira)            -2.117e-03  5.008e-04  -4.228 2.45e-05 ***
pb(peso)                1.980e-03  4.469e-04   4.431 9.80e-06 ***
pb(G2H)                 4.168e-04  6.469e-05   6.443 1.40e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) -2.287680   0.043012 -53.187  < 2e-16 ***
pb(edad)     0.003914   0.001042   3.757 0.000176 ***
diab_44DM    0.052537   0.072876   0.721 0.471032
diab_44IGT   0.114257   0.052543   2.175 0.029758 *
diab_44IFG  -0.018106   0.043745  -0.414 0.678974
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

------------------------------------------------------------------
NOTE: Additive smoothing terms exist in the formulas:
i) Std. Error for smoothers are for the linear effect only.
ii) Std. Error for the linear terms maybe are not accurate.
------------------------------------------------------------------
No. of observations in the fit:  2520
Degrees of Freedom for the fit:  22.06687
Residual Deg. of Freedom:  2497.933
at cycle:  5

Global Deviance:     20596.19
AIC:     20640.33
SBC:     20769.02
```

Na saída devolta polo `summary` obsérvanse dúas táboas principais. Na primeira, recóllense os coeficientes que afectan ao parámetro de localización (μ). Podemos ver catro columnas, na primeira delas vemos unha estimación do parámetro (no caso dos efectos suavizados unicamente fai referencia ao efecto lineal), xunto cos seus erros típicos, o estatístico de contraste e o p-valor asociado. Vemos como todos os p-valores son máis baixos que os niveis de significación habituais polo que temos evidencias estatísticamente significativas para dicir que son distintos de cero.

Na segunda táboa recóllense os coeficientes que afectan ao parámetro de escala (ϕ) xunto coas mesmas 4 columnas que para o parámetro anterior. Neste caso vemos que existen p-valores máis altos que algúns dos niveis de significación habituais. Aínda así, axustáronse modelos sen esas variables e os AIC e BIC devoltos eran máis altos polo que continuamos considerando este modelo.

Na parte final da saída podemos ver o número de observacións do modelo, os graos de liberdade estimados (22.06687), os graos de liberdade dos residuos, así como a deviance global (20640.33) ou o AIC (20769.02).

Unha vez visto o resumo do modelo podemos representar gráficamente o efecto das funcións suavizadoras das covariables para o parámetro de localización sen máis que empregar a función `term.plot` do paquete `gamlss` de R. Vemos os gráficos devoltos na Figura 4.4.

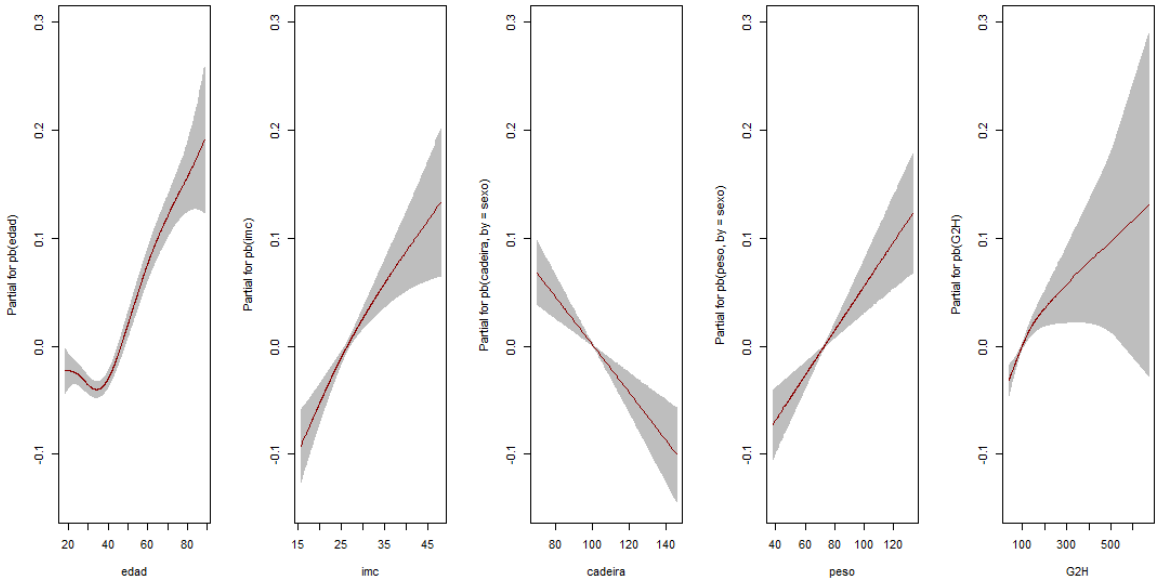


Figura 4.4: Efectos da función de suavizado correspondentes ás variables *idade*, *peso*, *imc*, *cadeira* e *G2H* tratando de explicar o parámetro de localización (μ) da distribuciónda familia exponencial que segue a variable resposta *pas* mediante un modelo aditivo xeralizado de localización, escala e forma (GAMLSS) considerando como base suavizadora os P-Splines xunto co efecto da variable categórica *sexo*.

Na primeira gráfica comezando pola esquerda vemos o efecto da *idade*, que inicialmente produce unha lixeira baixada (entre os 20 e os 40 anos) para despois subir rápidamente a medida que aumenta a idade. Na segunda gráfica vemos o efecto do *imc* sobre a presión arterial sistólica que se corresponde co xa visto no caso dos modelos GAM, inicialmente aumenta rápidamente ata chegar máis ou menos ao seu valor medio onde comeza a aumentar de forma máis lenta (podemos ver que para valores altos temos moita variabilidade polo que non podemos extraer moitas conclusións). Tras isto podemos ver os gráficos correspondentes ao efecto da variable *cadeira* e da variable *peso*, que ambos parecen ser lineais. No primeiro caso descendente (a medida que aumenta o tamaño da cadeira diminúe a PAS) e o segundo ascendente (a medida que aumenta o peso aumenta a PAS). Por último, vemos o efecto da glucosa en sangue pasadas dúas horas, na que claramente se produce un aumento rápido a medida que aumenta este valor .Temos que diferenciar dúas zonas: nunha primeira zona con valores por debaixo de 300 onde se atopan a meirande parte dos datos e temos pouca variabilidade e unha seginda zona por encima de 300 onde temos presentes poucos datos (e todos se corresponden con persoas con diabetes non diagnosticada) e moitisima variabilidade debido á pouca densidade de datos nese rango de valores, polo que tampouco sería correcto extraer moitas conclusión neste caso.

Unha vez vistas as gráficas correspondentes aos efectos suavizadores para o parámetro de localización (μ) podemos facer o mesmo para o caso do parámetro de escala (ϕ) sen máis empregar o argumento `what="sigma"` da función `term.plot` empregada antes. Podemos ver os resultados obtidos na Figura 4.5.

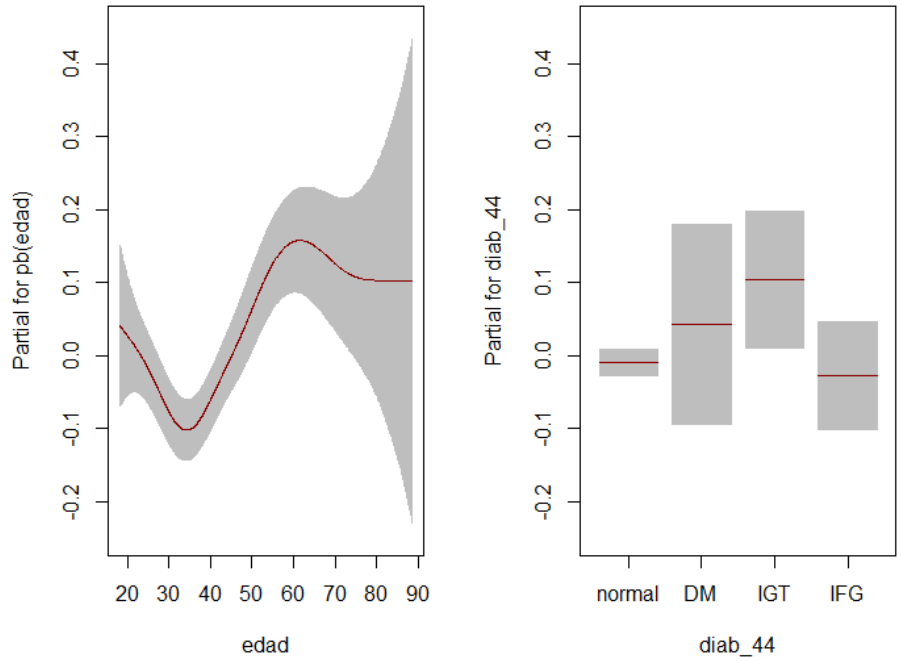


Figura 4.5: Efectos da variable *diab_44* e da función de suavizado correspondente á variable *idade* tratando de explicar o parámetro de escala (ϕ) da distribución da familia exponencial que segue a variable resposta *pas* mediante un modelo aditivo xeralizado de localización, escala e forma (GAMLSS) considerando como base suavizadora os P-Splines.

Vemos como na primeira gráfica se nos mostra o efecto suavizado da variable *Idade* sobre o parámetro de escala (ϕ) da distribución que segue a variable resposta *pas*. Este parámetro diminúe claramente e inicialemente a medida que aumenta a idade mentres que a partir dos 40 aproximadamente aumenta ata chegar aos 60 que se estabiliza e comeza a mostrar gran variabilidade. Este aumento do parámetro de escala entre os 40 e os 60 pode ser debido a que non se diferenciou por sexos (o paquete `gamlss` presenta problemas á hora de considerar interaccións entre variable) e polo tanto temos o efecto da menopausia nas mulleres e non nos homes. No segundo gráfico vemos o efecto da variable categórica *diab_44* no cal vemos claramente como o grupo pertencente a *IGT* presenta un maior valor do parámetro ϕ que o resto.

Como na Figura 4.4 vimos que o efecto das variable *cadeira* e *peso* parecía ser lineal, axustamos un novo modelo considerando estas variables sen efecto de suavizado e comprobamos que efectivamente este modelo nos devolvía un valor do AIC máis baixo polo que simplificamos o modelo para obter o seguinte:

$$\begin{aligned} Y_{PAS} &\sim EF(\mu_{PAS}, \phi_{PAS}) \\ \eta_1^{PAS} = g_1(\mu_{PAS}) &= \beta_{0\,\mu}^{PAS} + \beta_{1\,\mu}^{PAS} X_{sexo} + \beta_{2\,\mu}^{PAS} X_{peso} + \beta_{2\,\mu}^{PAS} X_{cadeira} \\ &\quad + s_{1\,\mu}^{PAS}(X_{Idade}) + s_{2\,\mu}^{PAS}(X_{imc}) + s_{3\,\mu}^{PAS}(X_{G2H}) \\ \eta_2^{PAS} = g_2(\phi_{PAS}) &= \beta_{0\,\phi}^{PAS} + \beta_{1\,\phi}^{PAS} X_{diab\_44} + s_{1\,\phi}^{PAS}(X_{idade}) \end{aligned} \tag{4.11}$$

Co modelo (4.11) xa axustado podemos facer unha diagnose dos residuos como se fixo no caso dos modelos GAM. Para isto empregamos o comando *plot* de R sobre o obxecto no que temos gardado o

modelo e obtemos os gráficos que podemos ver na Figura 4.6.

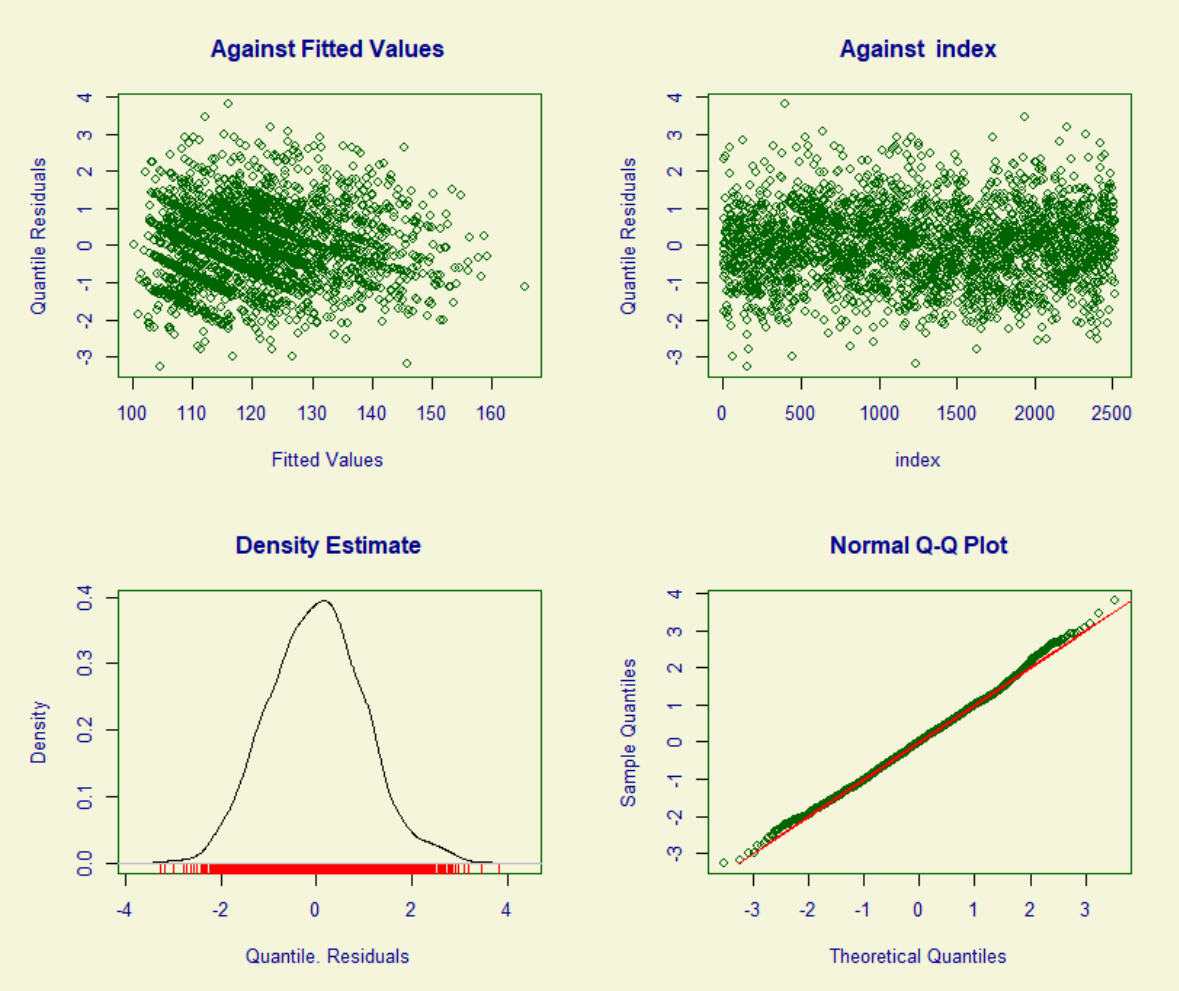


Figura 4.6: Gráficos devoltos pola función `plot` de R executada sobre o modelo axustado coa función `gamlss` do paquete `gamlss` que permiten a validación do modelo axustado en (4.11).

Nestes gráficos podemos ver que os residuos se comportan adecuadamente (aínda que poida apreciarse algo de tendencia no primeiro dos gráficos). Nos dous gráficos superiores móstranse os residuos fronte aos valores axustados de μ e fronte aos índices dos datos. En ambos casos parece verse unha dispersión aleatoria dos residuos entorno ao valor 0. Nos dous gráficos inferiores podemos ver unha estimación da densidade tipo núcleo na parte esquerda e un QQ-plot na parte dereita. Ambos gráficos parecen mostrar un comportamento adecuado, destacando que no segundo os puntos dos cuantís se sitúan sobre a recta vermella de forma case perfecta agás nos extremos.

Ademais destes gráficos, a función `plot` tamén imprime a seguinte información por consola.

```
plot(modeloS51)
```

```
      Summary of the Quantile Residuals
mean   =  3.880802e-05
variance   =  1.000393
coef. of skewness  =  0.1227169
```

```
coef. of kurtosis  =  3.069454
Filliben correlation coefficient  =  0.9991409
```

Aquí vemos un resumo dos cuantís dos residuos. Vemos que a súa media está próxima a cero ($3.880802 \cdot 10^{-5}$) e a súa varianza próxima a un (1.000393). Por outra parte o coeficiente de asimetría sepárase lixeiramente de cero (0.1227169) e o seu coeficiente e kurtosis tamén se separa un pouco de tres (3.069454). Polo que tampouco parece que debamos dubidar se a distribución dos nosos residuos será unha normal estándar. Rematamos deste xeito os axustes dos modelos GAMLSS e avanzaremos a uns novos modelos nos cales trataremos de relacionar o valor das dúas presións arteriais estudadas mediante un tipo de modelos multivariantes moi novedosos que se coñecen como CGAMLSS e que se engloban dentro dos coñecidos como Joint Modelling.

Capítulo 5

Modelos Joint Modelling

Neste capítulo introduciremos os Joint Modelling (seguindo (9) e (35)) a través do que coñecemos como CGAMLSS (*bivariate copula generalised additive models for location, scale and shape*), que non é máis que unha extensión ás respostas múltiples do caso dos GAMSLSS. De forma máis específica, comezaremos presentando o que se coñece como CGAMLSS e introduciremos o concepto de cópula, de gran importancia neste capítulo, mostrando algúns dos tipos de cópula máis importantes (para máis información ver (18)) que poderemos empregar posteriormente na parte práctica do traballo. Posteriormente formularemos os modelos CGAMLSS a partir de modelos xa coñecidos. Para rematar, trataremos a inferencia dos parámetros do modelo dende un punto de vista frecuentista ((19)) e dende un punto de vista bayesiano ((17)). Este tipo de modelos non serán máis que unha extensión ao caso multivariante dos modelos GAMLSS ((30)) vistos no capítulo anterior.

5.1. Modelos de regresión de cópulas bivariadas

Un modelo CGAMLSS ((19) e (17)) modela a distribución conxunta dun par de variables resposta (y_1, y_2) dadas un conxunto de covariables e unha cópula que especifique a estrutura de dependencia entre estas dúas variables resposta. Comezaremos considerando que as variables resposta y_1 e y_2 son continuas. Nestes modelos CGAMLSS, a función de distribución acumalada conxunta (cdf) de y_1 e y_2, dada a información das covariables (recollida en ν), expresámola en termos das funcións de distribucións marxinais e unha función cópula C que as une.

5.1.1. Cópulas

Chegados a este punto, podemos preguntarnos, que é unha cópula? Seguindo (22), podemos interpretar unha cópula dende dous puntos de vista: "Dende un punto de vista, as copulas son funcións que unen ou copulan funcións de distribución multivariadas cara as funcións de distribucións marxinais univariadas. Tamén se pode consultar (14) para máis información. Alternativamente, as cópulas son funcións de distribución multivariadas cuxas funcións de distribución marxinais univariadas son uniformes no intervalo $(0, 1)$".

Definamos agora cando unha función é unha cópula de forma máis específica:

Definición 5.1. Unha **cópula bivariada** $C : [0,1]^2 \longrightarrow [0,1]$ é unha función de distribución dun vector aleatorio (Y_1, Y_2) con marxes uniformes en $(0, 1)$:

$$C(\mathbf{y}) = \mathbb{P}[Y_1 \leq y_1.Y_2 \leq y_2]$$

onde

$$\mathbb{P}[Y_i \leq y_i] = y_i \quad \text{para } i = 1, 2 \text{ e } 0 \leq y_i \leq 1.$$

Outra posibilidade é definir as funcións cópulas en termos dunhas determinadas propiedaes que cumpren, como se indica en (8).

Definición 5.2. Dada unha función $C : [0,1]^2 \longrightarrow [0,1]$, diremos que C é unha **función cópula** se verifica as seguintes propiedades (que podemos ver en (23)):

1. $C(y_1, y_2) = 0$, $\forall\,(y_1, y_2) \in [0,1]^2 \Leftrightarrow y_1 = 0$ e/ou $y_2 = 0$.

2. $C(y_1, 1) = y_1$ e $C(1, y_2) = y_2$, $\forall\,(y_1, y_2) \in [0,1]^2$.

3. $\forall\,(a_1, a_2)\,(b_1, b_2) \in [0,1]^2$ verifícase que

$$C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) + C(a_1, b_1) \geq 0.$$

Pódemos ver máis propiedades das funcións cópula en (1)

Observación 5.3. Pódense entender como unha forma alternativa que busca correxir os erros ou problemas que presenta o coeficiente de correlación de dúas variables. Con estas funcións conseguimos modelar a dependencia entre variables fóra do contexto lineal e gaussiano.

Como consecuencia das propiedades vistas nesta definición, podemos concluír que un función de distribución acumulativa bivariada definida co cadrado unidade e con distribución marxinal uniforme é unha función cópula. Podemos asi definir unha función cópula de forma máis exacta, en virtude do Teorema de Sklar, como indicamos a continuación

$$F(y_1, y_2|\nu) = C(F_1(y_1|\nu), F_2(y_2|\nu), \rho) \tag{5.1}$$

onde $F_1(y_1|\nu)$ e $F_2(y_2|\nu)$ son as funcións de distribución marxinais de $y_1|\nu$ e $y_2|\nu$ que toman valores en $(0,1)$, $C(\cdot,\cdot|\nu)$ é a función de cópula que contén a información sobre a asociación entre as dúas variables respota e ρ é o parámetro de asociación da cópula que mide a dependencia entre as variables resposta.

Existen numerosas funcións cópula que permiten modelar diferentes tipos de estruturas de dependencia entre as variables respostas (podemos ver a Táboa 5.1 e a Figura 5.1 para ver algunhas delas). Por exemplo, a cópula Clayton permítenos considerar unha estrutura de dependencia asimétrica cando dúas variables aleatorias mostran unha maior asociación positiva para valores pequenos que para valores grandes. Pola contra, a cópula de Gumbel ou a cópula de Joe representan a situación contraria, dúas variables resposta que mostran unha maior dependencia para valores grandes. Tamén existen modificacións destas cópulas que nos permiten modelas estruturas de dependencia negativas.

A continuación recollemos as cópulas clásicas dun só parámetro xunto coa súa definición, o rango de valores do parámetro correspondente á cópula e a función enlace.

Cópula	$C(u, v; \rho)$	**Rango de ρ**	**Función Link**
Clayton	$(u^{-\rho} + v^{-\rho} - 1)^{\frac{-1}{\rho}}$	$\rho \in (0, \infty)$	$\log(\rho - \epsilon)$
FGM	$u\, v\{1 + \rho\,(1-u)\,(1-v)\}$	$\rho \in [-1, 1]$	$tanh^{-1}(\rho)$
Frank	$-\rho^{-1} \log\{1 + (e^{-\rho\, u} - 1)\,(e^{-\rho\, v} - 1)/(e^{-\rho} - 1)\}$	$\rho \in \mathbb{R} - \{0\}$	-
Gaussian	$\Phi_2\,(\Phi^{-1}(u), \Phi^{-1}(v); \rho)$	$\rho \in [-1, 1]$	$tanh^{-1}(\rho)$
Gumbel	$exp\left[-\{(-\log(u))^\rho + (-\log(v))^\rho\}^{1/\rho}\right]$	$\rho \in (1, \infty)$	$\log(\rho - 1)$
Joe	$1 - \{(1-u)^\rho + (1-v)^\rho - (1-u)^\rho\,(1-v)^\rho\}^{1/\rho}$	$\rho \in (1, \infty)$	$\log(\rho - 1 - \epsilon)$

Táboa 5.1: Cópulas máis habituais xunto co correspondene rango do parámetro de asociación ρ e a función link. $\Phi_2(\cdot, \cdot; \rho)$ denota a función de distribución bivariada dunha normal estándar co coeficiente de correlación ρ, mentres que Φ denota a función de distribución univariada dunha normal estándar. Por último, ϵ toma un valor de 10^{-7} e é usado para asegurarse de que as restricións do espazo de ρ se manteñen.

Vexamos unha representación gráfica dunha simulación de datos para cada unha das cópulas para ver así a relación de dependencia existente entre as dúas variables da función que comentabamos anteriormente.

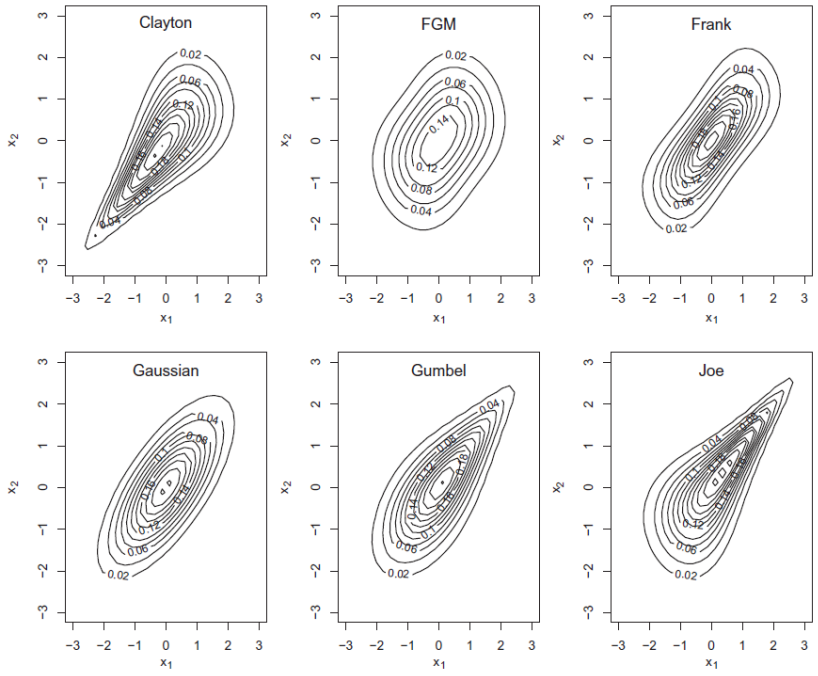


Figura 5.1: Gráficas do contorno de algunhas das funcións cópulas máis usuais para datos simulados considerando un valor de 0.5 para a τ de Kendall.

Podemos simular tamén datos considerando distintos valores para o parámetro dunha mesma función cópula e así ver o efecto que ten o parámetro na distribución dos mesmos. Comecemos considerando

os casos das cópulas gaussianas e FGM que podemos ver na Figura 5.2.

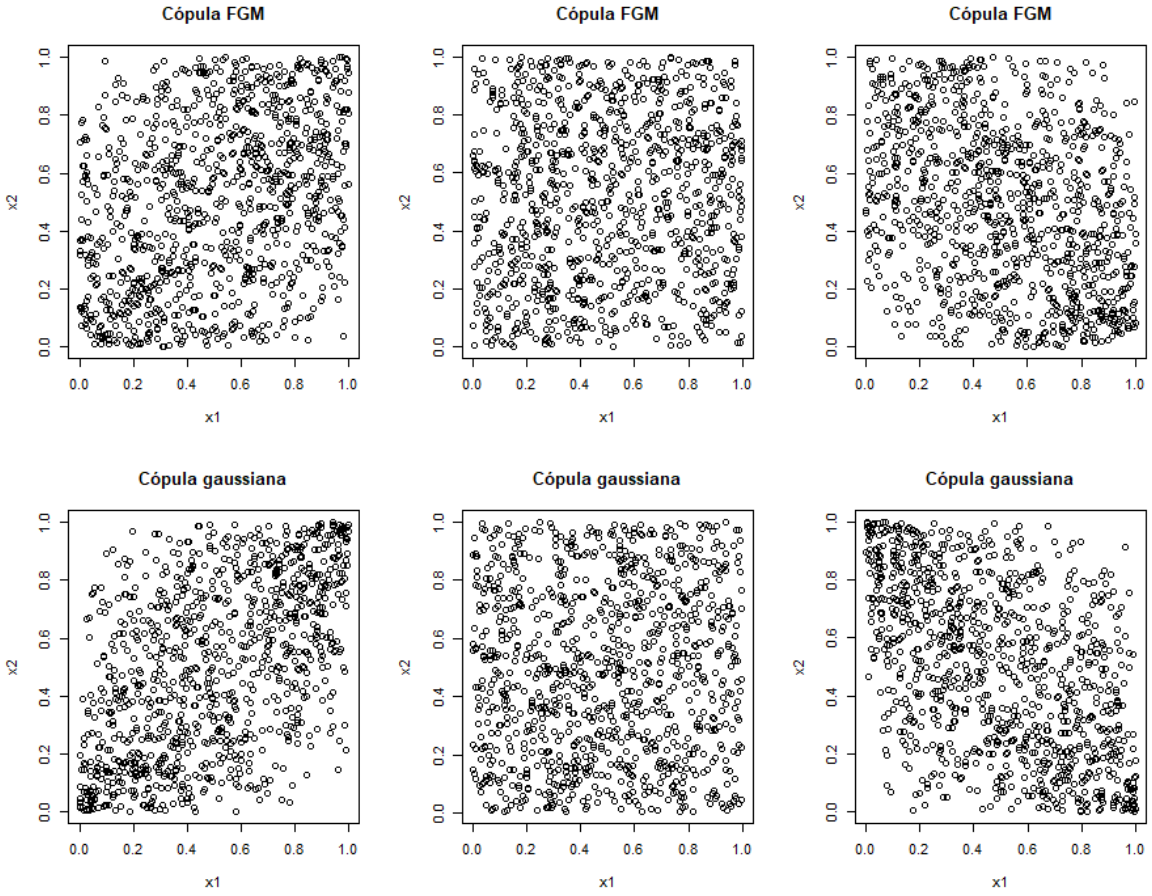


Figura 5.2: Diagramas de dispersión para datos simulados correspondentes ás funcións copulas FGM (na parte superior), con parámetro $\rho = 1$,$\rho = 0$ e $\rho = -1$ de esquerda a dereita, e Gaussiana (na parte inferior), con parámetro $\rho = 0.5$,$\rho = 0$ e $\rho = -0.5$.

Vemos así como para un ρ positivo, valores baixos dunha variable se corresponden con valores baixos da outra variable mentres que para un ρ negativo, valores baixos dunha variable se corresponden con valores altos da outra variable. Notemos tamén que esta relación de dependencia é moito maior no caso da cópula gaussiana que no caso da cópula FGM, como xa podíamos intuír na Figura 5.1. Isto mesmo acontece se consideramos a cópula Frank, aínda que non se incluirá de forma gráfica.

Vexamos tamén como afecta o valor de ρ ás cópulas que teñen un rango de valores para este parámetro máis grande e sempre positivo, como poden ser Clayton, Gumbel e Joe (como podemos ver na Táboa 5.1). Para isto mostramos a Figura 5.3.

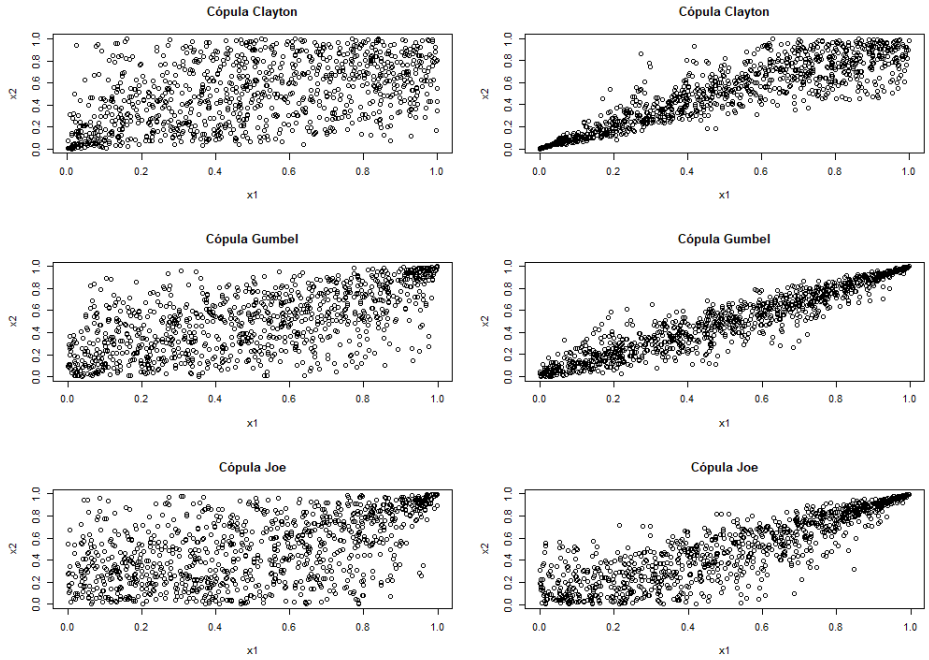


Figura 5.3: Diagramas de dispersión para datos simulados correspondentes ás funcións copulas Clayton (na parte superior), con parámetro $\rho = 1$ (na esquerda) e $\rho = 5$ (na dereita); Gumbel (na parte central), con parámetro $\rho = 2$ (na esquerda) e $\rho = 5$ (na dereita) e Joe (na parte inferior), con parámetro $\rho = 2$ (na esquerda) e $\rho = 5$ (na dereita).

Vemos así que nas tres funcións cópula consideradas aumenta a relación de dependencia entre as variables ao aumentar o valor do parámetro.

Ata agora todas as cópulas que vimos tiñan unicamente un parámetro e pódese clasificar en dous tipos:

- **Cópulas arquimedianas**: formadas por todas as cópulas bivariadas paramétricas que poden ser expresadas como suma de dúas funcións marxinais dependentes de cada unha das variables como segue (consultar (22)):
$$\phi(C(u,v)) = \phi(u) + \phi(v). \tag{5.2}$$
Pero como estamos interesados en coñecer a expresión de $C(u,v)$ debemos definila a través da función inversa de ϕ que, despexando de (5.2), temos que:
$$C(u,v) = \phi^{-1}(\phi(u) + \phi(v)). \tag{5.3}$$
A función ϕ é coñecida como función xeradora da cópula. Dentro desta familia atópanse as cópulas Clayton, Joe, Gumbel e Frank.

- **Cópulas elípticas:** as cópulas elípticas obtéñense de distribucións multivariantes elípticas aplicando a transformación inversa que se recolle no Teorema de Sklars (que se pode consultar en (5)). Sabemos entón que un vector aleatorio ten unha distribución elíptica multivariada se pode ser expresado do seguinte xeito:
$$X \overset{d}{=} \mu + R\,A\,U, \tag{5.4}$$
onde $\mu \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times k}$ con $\Sigma = A\,A^T$ e $rank(\Sigma) = k \leq d$, U é un vector aleatorio d-dimensional uniformemente distribuido na esfera $S^{d-1} = \{u \in \mathbb{R}^d : u_1^2 + \cdots + u_d^2 = 1\}$ e R é un unha variable aleatoria positiva indepedente de U. Dentro deste tipo de cópulas temos a cópula gaussiana vista anteriormente.

Cópula t-Student

Imos presentar agora unha cópula elíptica que ten dous parámetros pois será a que mellor se axuste aos nosos datos posteriormente. Deste xeito definimos a cópula t-Student como indicamos:

$$X \overset{d}{=} \mu + \Sigma^{1/2}\sqrt{W}\,Z, \tag{5.5}$$

onde $Z \in N_d(0, I_d)$ é unha distribución normal, $\Sigma = \Sigma^{1/2}\,\Sigma^{1/2}$ é definida positiva. Ademais, W e Z son independentes e W segue unha distribución Gamma inversa con parámetros $(\nu/2, \nu/2)$. Deste xeito a cópula bivariada t-Student vén dada por:

$$C(u, v) = t_{2,\zeta}\left(t_{\zeta}^{-1}(u), t_{\zeta(v)}^{-1}; \zeta, \rho\right)$$

onde $\rho \in (-1, 1)$, t_2 deota unha distribución t-Student bivariada e t_{ζ}^{-1} denota a inversa de distribución t-Student estándar con ζ graos de liberdade. Esta cópula convertese nunha gaussiana se $\zeta \to \infty$.

5.2. Formulación do modelo

Unha vez vistas as funcións cópulas, que xogarán un gran papel nestes novos modelos que imos presentar, procedemos a formulalos dunha forma máis teórica. Traballando con modelos de regresión CGAMLLS aumentamos a flexibilidade con respecto aos modelos considerados anteriormente pois non temos só flexibilidade á hora de escoller as distribucións marxinais da resposta bivariable se non que tamén cando escollemos a estrutura de dependencia existente entre elas cando seleccionamos a cópula.

Consideremos que temos un conxunto de observacións $\{y_i = (y_{i1}, y_{i2}), i \in \{1, \ldots, n\}\}$, onde y_{i1},y_{i2} son as variables resposta do noso modelo, e o vector de covariable $\nu_i, i \in \{1, \ldots, n\}$. Sexa p_1 e p_2 a densidade marxinal de y_1 e y_2 respectivamente,

$$p_{1,i} \equiv p_1(Y_{1i}|\mu_{1i}, \sigma_{1i}),$$
$$p_{2,i} \equiv p_2(Y_{2i}|\mu_{2i}, \sigma_{2i}).$$

Observación 5.4. Neste caso estamos supoñendo que $p_{1,i}$ e $p_{2,i}$ dependen de dous parámetros cada un, pero en xeral poderiamos considerar distribucións máis complexas.

No modelos CGAMLSS, todos os parámetros se poden expresar como preditores aditivos do mesmo xeito que no caso dos modelos lineais xeralizados, empregando unha función link bixectiva adecuada, g, que asegura que as restricións nos espazos de parámetros se manteñen. Daquela podemos escribir o seguinte:

$$\eta_i^{\mu_1} = g_{\mu_1}(\mu_{1i});\ \eta_i^{\sigma_1} = g_{\sigma_1}(\sigma_{1i}),$$
$$\eta_i^{\mu_2} = g_{\mu_2}(\mu_{2i});\ \eta_i^{\sigma_2} = g_{\sigma_2}(\sigma_{2i}).$$

A elección da función link ben determinada polas restricións que se apliquen ao espazo de parámetros do correspondente parámetro en función da distribución asignada á variable correspondente (consultar (19) e (16) para ver de forma máis detallada as funcións enlace).

Para simplificar a notación, imos considerar unicamente cópulas cun só parámetro (como podemos ver na Táboa 5.1). Ademais, o parámetro da cópula, que denotaremos por ρ_i, tamén estará relacionado cun preditor aditivo, η_i^{ρ}, como mostramos

$$\eta_i^{\rho} = g_{\rho}(\rho_i),$$

onde a eleción da función link da cópula, $g_{\rho}(\cdot)$, depende do tipo de cópula e pode ser consultada na Táboa 5.1.

Deste xeito vemos que o número total de parámetros da función de distribución conxunta que desexamos modelar no CGAMLSS é a suma do número de parámetros de cada distribución marxinal e o

número de parámetros da función cópula. No caso particular que estamos considerando, temos un total de cinco parámetros (consideraremos unicamente dous parámetros para as distribucións marxinais, localización e escala, e un único parámetro para a función cópula). Sexa $\vartheta = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\}$, a estimación dos predictores lineais deste espazo de parámetros vén dada por:

$$\begin{aligned} \eta_i^{\mu_k} &= \beta_0^{\mu_k} + \sum_{j=1}^{J^{\mu_k}} f_j^{\mu_k}(\nu_i), \quad k \in \{1,2\}, \\ \eta_i^{\sigma_k} &= \beta_0^{\sigma_k} + \sum_{j=1}^{J^{\sigma_k}} f_j^{\sigma_k}(\nu_i), \quad k \in \{1,2\}, \\ \eta_i^{\rho} &= \beta_0^{\rho} + \sum_{j=1}^{J^{\rho}} f_j^{\rho}(\nu_i), \end{aligned} \tag{5.6}$$

onde $\beta_0^{(\cdot)}$ é o intercepto xeral de cada preditor, as funcións $f_j^{(\cdot)}$ representan os diferentes efectos das covariables e $J^{(\cdot)}$ denota o número total de efectos das covariables consideradas para cada preditor. Notemos que cada un dos preditores lineais definidos en (5.6), tanto dos parámetros das distribucións marxinais como da cópula, pode depender de diferentes covariables e diferentes efectos de cada unha delas.

De forma xeral e para simplificar a notación, deixando de lado a dependencia dos parámetros, podemos usar a seguinte forma xenérica para referírnos a ecuación (5.6):

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(v_i). \tag{5.7}$$

Ademais, do mesmo xeito que acontecía nos modelos aditivos xeralizados, cada función f_j da ecuación (5.7) pode ser expresada mediante unha combinación lineal de D_j funcións bases escollidas de forma axeitada como indicamos a continuación:

$$f_j(\nu_i) = \sum_{d_j}^{D_j} \beta_{j,d_j} B_{j,d_j}(\nu_i). \tag{5.8}$$

Da ecuación (5.8) deducimos que o vector de avaliacións $(f_j(\nu_1), \ldots, f_j(\nu_n))^T$ pódese escribir como $Z_j\, b_j$ con $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,D_j})^T$, onde β_j recolle todos os coeficientes das base e as entradas $Z_j[i, d_j] = B_{j,d_j}(\nu_i)$ da matriz de deseño Z son as funcións base evaluadas nos valores observados das covariables. Deste xeito, podemos concluír escribindo a ecuación (5.7) como segue:

$$\eta = \beta_0\, 1_n + Z_1\, \beta_1 + \cdots + Z_J\, \beta_J, \tag{5.9}$$

onde $\eta = (\eta_1, \ldots, \eta_n)^T$ representa o vector de preditores para todas as observacións e 1_n é un vector $n-$dimensional de uns.

Igual que acontecía nos modelos GAM, para regular á hora da estimación a posible sobreparametrización do modelo que axustamos teremos que facer determinadas consideracións.

Por unha parte, dentro do contexto frecuentista, cada vector β_j ten asociado unha penalización cuadrática (dentro do contexto da verosimilitude penalizada que levamos considerando ao longo do traballo), De forma máis exacta, consideraremos penalizacións cuadráticas que tomen a forma $\lambda\, \beta^T\, K\, \beta$ (non estamos considerando o índice do parámetro e o índice da función por simplicidade), onde se escolle a matriz de penalización semidefinica positiva K de forma que cumpra as propiedades desexables do efecto da función correspondente (por exemplo, suavidade). O parámetro de suavizado, $\lambda \in [0, \infty)$ mantén o equilibrio entre un bo axuste e a suavidade do mesmo. Xogará un papel moi importante á hora de conseguir efectos adecuados das variables explicativas consideradas para manter a suavidade neste efectos se caer na sobreestimación.

Considerando agora o contexto bayesiano, o termo de penalización é substituido por

$$p(\beta|\tau^2) \propto \exp\left(-\frac{1}{2\,\tau^2}\,\beta^T\,K\,\beta\right),$$

onde a matriz K é agora a matriz de precisión previa e τ^2 reemplaza ao parámetro de suavizado da verosimilitude penalizada. Podemos relacionar o modelo bayesiano a posteriori coa verosimilitude penalizada considerando $\lambda = \frac{1}{2\,\tau^2}$.

Podemos modelar moitos tipos de efectos asumindo distintas suposicións sobre as bases de funcións e a matriz de precisión da penalización ou a matriz de precisión a priori (para máis detalles, consultar (33) e (10))

5.3. Inferencia

Pasaremos agora a discutir sobre a inferencia, dende un punto de vista frecuentista e dende un punto de vista bayesiano, para o caso particular de modelos de regresión con cópulas entre dúas variables continuas presentadas na Sección 5.2. Neste caso, a función da log-verosimilitude dun modelo CGAMLSS con dúas variables marxinais continuas pódese escribir como

$$\ell(\theta) = \sum_{i=1}^{n} \log\{c(F_1(y_{1i}|\mu_{1i},\sigma_{1i}), F_2(y_{2i}|\mu_{2i},\sigma_{2i}),\rho)\} + \sum_{i=1}^{n}\sum_{d=1}^{2} \log\{p_d(y_{di}|\mu_{di},\sigma_{di})\}, \tag{5.10}$$

para $d = 1, 2$ onde $c(\cdot,\cdot,\rho)$ é a función de densidade da función cópula (que vén dada por $\frac{\partial^2 C(F_1(y_{1i}),F_2(y_{2i}))}{\partial F_1(y_{1i})\partial F_2(y_{2i})}$) e $p_d(y_d|\mu_d,\sigma_d)$ a densidade marxinal da $d-$ésima variable resposta. Neste caso o vector de parámetros θ vén definido como $(\beta_{\mu_1}^T, \beta_{\mu_2}^T, \beta_{\sigma_1}^T, \beta_{\sigma_2}^T, \beta_{\rho}^T)$, que é o vector de coeficientes asociado con $\eta_i^{\mu_1}$,$\eta_i^{\mu_2}$,$\eta_i^{\sigma_1}$,$\eta_i^{\sigma_2}$ e η_i^{ρ} respectivamente.

5.3.1. Inferencia bayesiana

Traballando con CGAMLSS, a inferencia bayesiana realízase empregando un algoritmo modular basado en simulacións de cadeas de Markov por Monte Carlo, a través da sucesiva actualización de todos os parámetros do modelo durante as iteracións realizadas. Segundo (17) a distribución log-posterior de $p(\theta|y)$ vén dada por:

$$\log(p(\theta|y)) \propto \ell(\theta) + \sum_{k=1}^{K}\sum_{j=1}^{J_k} \log(p(\beta_{j,k}|\tau_{j,k}^2)\,p(\tau_{j,k}^2)) \tag{5.11}$$

onde θ é o vector de parámetros do modelo e y denota a matriz resposta. Porén, a expresión (5.11) non será manexable para os modelos cos que imos traballar. Por este motivo, Klein e Kneib en (17) empregan un algoritmo de Metropolis-Hastings no cal as densidades propostas para os vectores de coeficientes $\beta_{j,k}$ se obteñen aproximando o logaritmo completo $\log(p(\beta_{j,k}|\cdot))$ mediante unha extensión de Taylor de segunda orde. Para o caso dos parámetros da varianza $\tau_{j,k}^2$, asumiremos que $\tau_j^2 \sim IG(a_j, b_j)$ con $a_j = b_j = 0.001$.

5.3.2. Inferencia por máxima verosimilitude penalizada

Se consideramos agora un punto de vista frecuentista, a estimación dos parámetros de regresión basease na optimización directa da función de verosimilitude penalizada coa selección automática do parámetro de suavizado. Neste tipo de modelos, o uso de algoritmos de optimización sen penalizar produce estimacións con demasiada curvatura. Marra e Radice (19) propuxeron maximizar a seguinte expresión:

$$\ell_p(\theta) = \ell(\theta) - \frac{1}{2}\theta^T\,S\,\theta \tag{5.12}$$

onde $\ell(\theta)$ é a log-verosimilitude do modelo de regresión CGAMLSS con distribucións marxinais continuas (definida na ecuación (5.10)), $\ell_p(\theta)$ denota a log-verosimilitude penalizada do modelo e $S = \text{diag}(K_{\mu_1}, K_{\mu_2}, K_{\sigma_1}, K_{\sigma_2}, K_\rho)$. Os parámetros de suavizado contidos nas compoñentes K forman un vector de parámetros de suavizado que denotaremos por $\lambda = (\lambda_{\mu_1}, \lambda_{\mu_2}, \lambda_{\sigma_1}, \lambda_{\sigma_2}, \lambda_\rho)^T$.

Para estimar θ e λ, Marra e Radice ((19)) propuxeron usar un algoritmo estable e eficiente que integra de forma automática a selección múltiple de parámetros de suavizado. (Consultar (19) para máis detalles sobre o funcionamento do algoritmo).

5.4. Aplicación a datos biomédicos

Para finalizar o traballo procedemos agora a axustar un CGAMLSS considerando como variables resposta a presión arterial sistólica e diastólica xunto coas variables explicativas recollidas no Capítulo 1. Do mesmo xeito que fixemos nos modelos anteriores, faremos o axuste e o estudo do modelo dende un punto de vista frecuentista para o cal usaremos de novo o Criterio de Información de Akaike (AIC) e o Criterio de Información Bayesiano (BIC) que podemos ver en (19) como Marra e Radice mostraron a través de estudos de simulación que, dende o punto de vista frecuentista, o AIC e o BIC eran criterios axeitados para identificar de forma correcta a función cópula do modelo CGAMLSS (e polo tanto ver tamén cal é o tipo de dependencia dos datos considerados).

5.4.1. Distribucións marxinais

Escoller unha distribución marxinal errada para algunha das dúas variables resposta pode afectar na selección da función cópula axeitada para o modelo final. Polo tanto, para tratar de evitar isto, iniciaremos o proceso de axuste deste modelo seleccionando as dúas distribucións marxinais axeitadas para cada unha das variables resposta mediante dous modelos GAMLSS independentes.

A diferenza do que fixemos nos Capítulos 3 e 4, nos cales íamos introducindo as variables de forma manual e comparando os modelos, agora imos facer a selección do modelo a través da función `stepGAICAll.A` do paquete `gamlss` de R. Recordemos que neste capítulo consideraremos máis variables que nos capítulos anteriores. Os modelos devoltos para cada unha das variables resposta son os que mostraremos a continuación. Para o caso da variable *pas*:

$$\begin{aligned} Y_{PAS} &\sim EF(\mu_{PAS}, \phi_{PAS}) \\ \eta_1^{PAS} = g_1(\mu_{PAS}) &= \beta_{0\,\mu}^{PAS} + \beta_{1\,\mu}^{PAS} X_{sexo} + \beta_{2\,\mu}^{PAS} X_{fumador} + \beta_{3\,\mu}^{PAS} X_{hipertiroide} + \beta_{3\,\mu}^{PAS} X_{diab\_55} \\ &\quad + s_{1\,\mu}^{PAS}(X_{Idade}) + s_{2\,\mu}^{PAS}(X_{imc}) + s_{3\,\mu}^{PAS}(X_{trig}) + s_{4\,\mu}^{PAS}(X_{colesterol}), \\ \eta_2^{PAS} = g_2(\phi_{PAS}) &= \beta_{0\,\phi}^{PAS} + s_{1\,\phi}^{PAS}(X_{idade}). \end{aligned} \tag{5.13}$$

Mentres que para o caso da variable *pad* obtemos o seguinte modelo:

$$\begin{aligned} Y_{PAD} &\sim EF(\mu_{PAD}, \phi_{PAD}) \\ \eta_1^{PAD} = g_1(\mu_{PAD}) &= \beta_{0\,\mu}^{PAD} + \beta_{1\,\mu}^{PAD} X_{sexo} + \beta_{2\,\mu}^{PAD} X_{fumador} + s_{1\,\mu}^{PAD}(X_{Idade}) + s_{2\,\mu}^{PAD}(X_{cintura}) + s_{3\,\mu}^{PAD}(X_{imc}) \\ &\quad + s_{4\,\mu}^{PAD}(X_{colesterol}) + s_{5\,\mu}^{PAD}(X_{g2h}) + s_{6\,\mu}^{PAD}(X_{insulina}) + s_{7\,\mu}^{PAD}(X_{alcohol}) + s_{8\,\mu}^{PAD}(X_{trig}), \\ \eta_2^{PAD} = g_2(\phi_{PAD}) &= \beta_{0\,\phi}^{PAD} + \beta_{1\,\mu}^{PAD} X_{lugar\_resid} + s_{1\,\phi}^{PAD}(X_{colesterol}). \end{aligned} \tag{5.14}$$

Unha vez obtidos ambos modelos imos ver que distribución se axusta mellor a eles dentro da familia de distribucións que temos dispoñibles nos modelos GAMLSS. Para iso compararemos o valor do AIC de cada unha delas (axustando os modelos que podemos ver en (5.13) e (5.14) especificando o argumento `family`) mediante a función `chooseDist`. Para o caso da variable *pas* mostramos os resultados na Táboa 5.2.

Distribución	BCCG	BCT	BCPE	LOGNO	LNO	GG	IG	GIG	BCTo	GA
AIC	21982.58	21984.44	21984.57	21991.92	21991.92	21993.84	21993.94	21994.31	21995.66	21999.95

Táboa 5.2: Valores do AIC devoltos pola función `chooseDist` de R para algunhas das distintas distribucións para axustar un modelo GAMLSS para a variable presión arterial sistólica (PAS) tomando como variables explicativas para os seus parámetros as que podemos ver na Ecuación (5.13)

Mentres que para o caso da variable *pad* obtemos os resultados que se mostran na Táboa 5.3

Distribución	BCCG	BCT	BCPE	LOGNO	LNO	IG	GIG	BCTo	GA
AIC	19840.42	19830.47	19840.24	19850.07	19850.07	19861.32	-	19834.11	19850.07

Táboa 5.3: Valores do AIC devoltos pola función `chooseDist` de R para algunhas das distintas distribucións para axustar un modelo GAMLSS para a variable presión arterial diastólica (PAD) tomando como variables explicativas para os seus parámetros as que podemos ver na Ecuación (5.14)

Posto que no paquete `GJRM` de R que posteriormente usaremos para axustar o modelo CGAMLSS non están dispoñibles todas as distribucións marxinais que nos ofrece o paquete `gamlss` imos escoller as distribucións Gamma e Log-Normal que se atopan entre as 10 distribucións con AIC e BIC e polo tanto proporcionan un bo axuste do modelo. Vemos tamén que as diferenzas existentes entre as distintas distribucións que mostramos tampouco son moi considerables polo que calquera destas distribucións sería unha boa escolla para as distribucións marxinais do modelo CGAMLSS que axustaremos finalmente.

5.4.2. Selección da función cópula

Para a selección da función cópula imos proceder dunha forma moi similar á vista para a escolla dos modelos GAMLSS e das distribucións marxinais das nosas variables. Estimaremos o modelo considerando diferentes cópulas a través do paquete `GJRM` e recollemos o valor do AIC e BIC na Táboa 5.4.

Cópula	Clayton	Joe	Gumbel	Frank	Gaussiana	T-Student
AIC	-1619.897	-1908.553	-1564.44	-1831.918	-1967.865	-1991.661
BIC	-1614	-1902.655	-1558.542	-1826.021	-1961.968	-1979.866

Táboa 5.4: Valores do Criterio de Información de Akaike (AIC) e Criterio de Información Bayesiana (BIC) para tratar de explicar a estrutura de dependencia existente entre a presión arterial sistólica (PAS) e a presión arterial diastólica (PAD).

Imos comprobar se esta cópula recolle de forma axeitada a correlación existente entre as dúas presións que queremos considerar. Para iso comezamos facendo unha representación da cópula seleccionada que podemos ver na Figura 5.4 para entender que tipo de correlación entre as variables resposta nos modela este tipo de cópula.

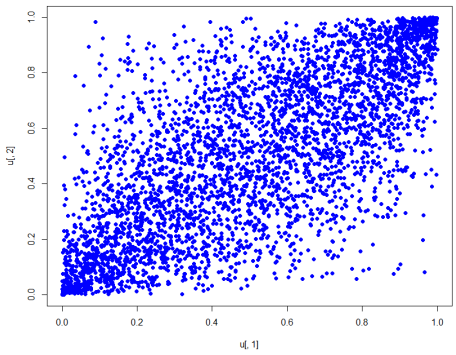


Figura 5.4: Diagrama de dispersión dos datos xerados aleatoriamente mediante a distribución que segue a función cópula T-Student con parámetro $\rho = 0.7299015$ e 12.38472 graos de liberdade.

Vemos así como parece que a función t-cópula ten un especial interés nos valores extremos, é dicir, é unha función adecuada para modelar o fenómeno de correlación directa nos valores extremos (os que se corresponden coas colas das distribucións). Notemos tamén que se trata dunha función simétrica. Representemos agora os valores dunha mostra aleatoria desta distribución xunto cos valores das nosas variables resposta mediante un gráfico de dispersión e vexamos se teñen un comportamento similar na Figura 5.5.

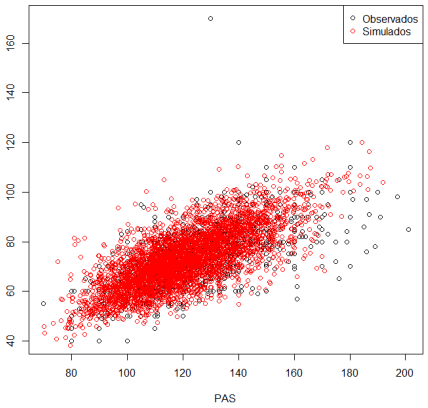


Figura 5.5: Representación do diagrama de dispersión da presión arterial diastólica (*pad*) frente a presión arterial sistólica (*pas*) en cor negra xunto co diagrama de dispersión dos datos simulados pola distribución da función t-cópula con parámetro $\rho = 0.7299015$ e 12.38472 graos de liberdade en cor vermella.

Vemos así como efectivamente a información da estrutura de relación entre a variable *pas* e a variable *pad* esta correctamente capturada pola cópula que seleccionamos.

5.4.3. Axuste do modelo CGAMLSS

Unha vez visto que a función cópula axustada é a adecuada e coas distribucións marxinais xa especificadas e modeladas, procedemos a realizar o axuste do modelo CGAMLSS. Para iso temos que ter en conta que neste tipo de modelos teremos que realizar, por así decilo, tres pasos:

1. Primero será necesario que definamos as covariables que teñen efecto sobre cada un dos parámetros do noso modelo. Tanto das dúas variables resposta (que non teñen que ter as mesmas covariables) como dos parámetros da función cópula.

2. Posteriormente temos que seleccionar a distribución marxinal que seguen as nosas variables resposta como xa fixemos para o caso dos modelos GAMLSS, pois será un argumento que nos pida a función `gjrm` de R.

3. Seleccionar a cópula que mellor axuste a relación de dependencia existente entre as dúas variables resposta consideradas.

Deste xeito, variando cada un destes puntos podemos axustar distintos modelos e comparalos en términos de AIC e BIC. Para comezar imos mostrar o axuste consierado para os parámetros, que non variaremos en ningún dos nosos axustes.

$$\begin{aligned}
\eta^{\mu_1} &= \beta_0^{\mu_1} + \beta_1^{\mu_1} X_{sexo} + \beta_2^{\mu_1} X_{fumador} + \beta_3^{\mu_1} X_{hipertiroide} + s_1^{\mu_1}(X_{idade}) + s_2^{\mu_1}(X_{imc}) + s_3^{\mu_1}(X_{g2h}) \\
&\quad + s_4^{\mu_1}(X_{trig}) + s_5^{\mu_1}(X_{diab_55}) + s_6^{\mu_1}(X_{colesterol}), \\
\eta^{\phi_1} &= \beta_0^{\phi_1} + s_1^{\phi_1}(X_{idade}), \\
\eta^{\mu_2} &= \beta_0^{\mu_2} + \beta_1^{\mu_2} X_{sexo} + \beta_2^{\mu_2} X_{fumador} + s_1^{\mu_2}(X_{idade}) + s_2^{\mu_2}(X_{imc}) + s_3^{\mu_2}(X_{g2h}) + s_4^{\mu_2}(X_{trig}) \\
&\quad + s_5^{\mu_2}(X_{cintura}) + s_6^{\mu_2}(X_{colesterol}) + s_7^{\mu_2}(X_{insulina}) + s_8^{\mu_2}(X_{alcohol}), \\
\eta^{\phi_2} &= \beta_0^{\phi_2} + \beta_1^{\mu_1} X_{lugar_resid} + s_1^{\phi_2}(X_{colesterol}) \\
\eta^{\rho} &= \beta_0^{\rho}.
\end{aligned} \tag{5.15}$$

Unha vez selecionadas as covariables para cada un dos parámetros do noso modelo (neste caso un total de cinco parámetros) imos axustar modelos variando as distribucións marxinais consideradas (log-normal ou gamma como dixéramos anteriormente) xunto coa función cópula (que aínda que escolleramos unha t-cópula tamén imos realizar axustes considerando unha gaussiana, pois recollen estruturas similares). Mostramos os valores do AIC e BIC obtidos para cada un dos modelos na Táboa 5.5.

Cópula	**Distribucións marxinais**	**AIC**	**BIC**
T-Student	Gamma	41018.93	41378.12
Gaussiana	Gamma	94511.85	94767.73
T-Student	Log-Normal	78871.72	79815.68
Gaussiana	Log-Normal	41183.87	41602.86

Táboa 5.5: Valores do Criterio de Información de Akaike (AIC) e Criterio de Información Bayesiana (BIC) para os CGAMLSS axustados coa función cópula que se indica na primeira columna, as distribucións marxinais da segunda e as expresións de cada un dos parámetros que vemos en (5.15).

Vemos así como o mellor modelo é aquel que considera unha T-Student como función cópula e unha Gamma como distribución marxinal para cada unha das variables resposta.

5.4.4. Resultados

Realizando un `summary` do modelo axustado podemos ver os coeficientes obtidos e analizalos como xa fixemos no caso dos modelos GAM e GAMLSS. Posto que o valor dos coeficientes non é moi diferente aos casos xa vistos, non introduciremos a saída de R. Mostramos tamén un scatterplot da presión arterial sistólica e diastólica na Figura 5.6.

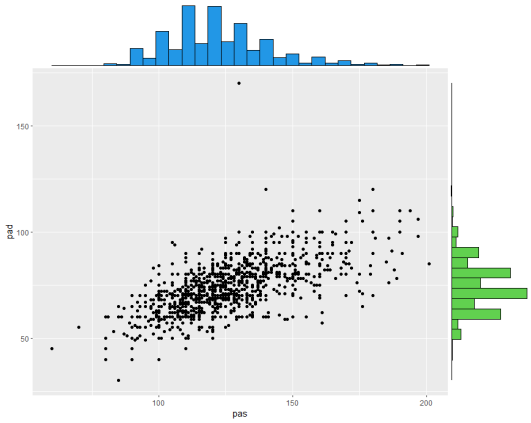


Figura 5.6: Scatterplot da presión arterial sistólica e diastólica no que se representa simultáneamente un diagrama de dispersión da presión diástolica frente a sistólica xunto cos histogramas de cada variable.

Para comprobar o bo axuste realizado das distribucións marxinais podemos empregar o comando `post.check` que nos devolve os gráficos da Figura 5.7.

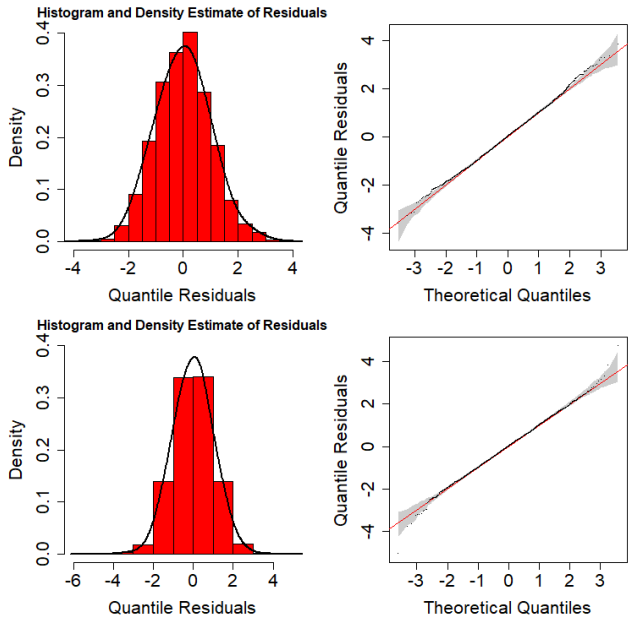


Figura 5.7: Histograma e QQ-plot dos cuantís residuais normalizados para a presión arterial sistólica (parte superior) e a presión arterial diastólica (parte inferior) para o modelo axustado.

Vemos como en ambos casos o histograma xunto coa función de densidade estimada parece corresponderse adecuadamente co corportamento esperado de normalidade. No caso dos QQ-plot ambos

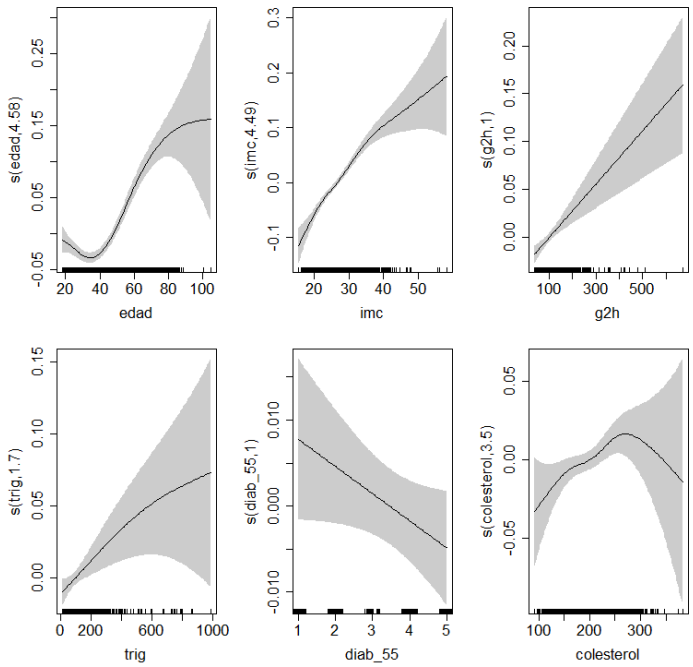


Figura 5.8: Efectos de suavizado sobre a media da presión arterial sistólica das variables *idade*, *imc*, *g2h*, *trigliceridos*, *diab_55* e *colesterol* dentro do modelo GJRM indicado en (5.15) tomando como función cópula unha T-Student e con distribucións marxinais gammas.

teñen un bo comportamento, se ben é certo que para o caso da PAS os cuantís se escapan algo máis nos extremos mentres que no caso da PAD se manteñen dentro dos intervalos de confianza e preto da liña de referencia en todo o rango de valores.

Aínda que non se van comentar, pois terán unha interpretación análoga á xa vista en modelos anteriores, imos representar os efectos das funcións de suavizado de cada unha das covariables sobre cada un dos parámetros. Podemos ver o caso do parámetro de localización (μ) para a distribución da variable *pas* na Figura 5.8 e o mesmo para o caso da distribución da variable *pad* na Figura 5.9. O mesmo acontece para o parámetro de escala (ϕ) para ambas distribucións que podemos ver na Figura 5.10.

Aínda que non o fagamos de forma detallada, podemos ver na Figura 5.8 como os triglicéridos teñen un efecto ascendente na media a medida que aumentamos o valor desta variable (a pesar de que para valores altos teñamos moita variabilidade e non poidamos facer ningún comentario preciso) e tamén como o colesterol parace ter un efecto ascendente para valores baixos (entre 100 e 220), mentres que despois se estabiliza arredor de 200 e sube de novo (da baixada final non extraemos ningunha conclusión pola gran variabilidade que presenta e o escaso número de observacións nese intervalo que temos).

No caso da presión arterial diastólica (Figura 5.9) podemos ver como o colesterol non parece ter efecto para valores baixos e altos (mantese a función de suavizado case estable), mentres que en valores intermedios si sufre unha forte subida ao pasar de niveis baixos a niveis altos de colesterol. Por outra parte a insulina parece non ter efecto mentres que os triglicéridos e o alcohol teñen un efecto de aumento na media da distribución da nosa variable. Vemos na Figura 5.10 para o caso do parámetro de escala como no caso da distribución da presión arterial sistólica esta diminúe inicialmente coa idade (ata os corenta) mentres que despois comeza a aumentar lentamente. Para o caso da presión arterial diastólica aumenta inicialmente para valores baixos do colesterol ata chegar a un pico entre 150 e 200 para despois comezar a descender e estabilizarse finalmente.

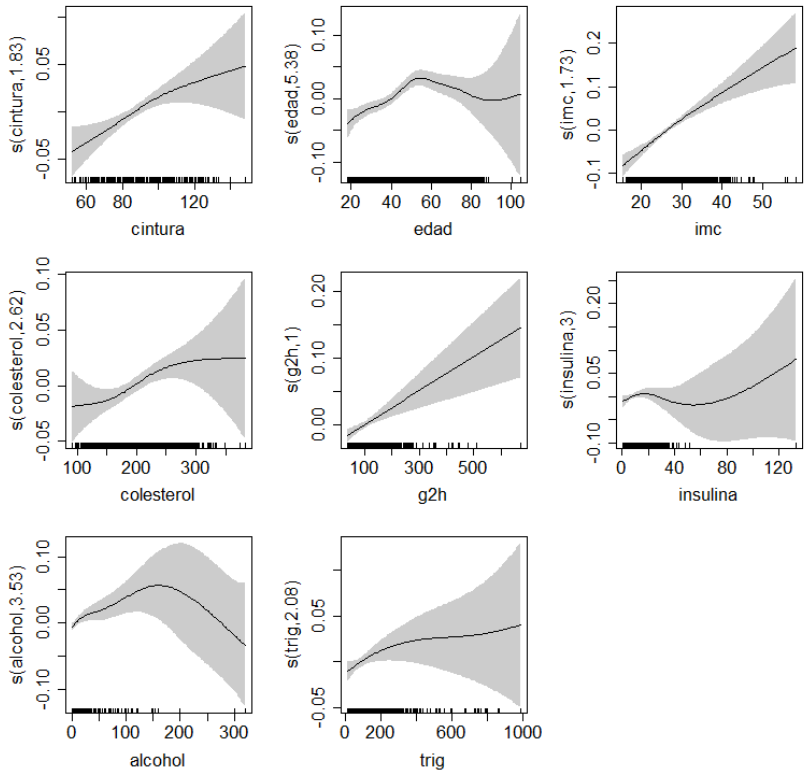
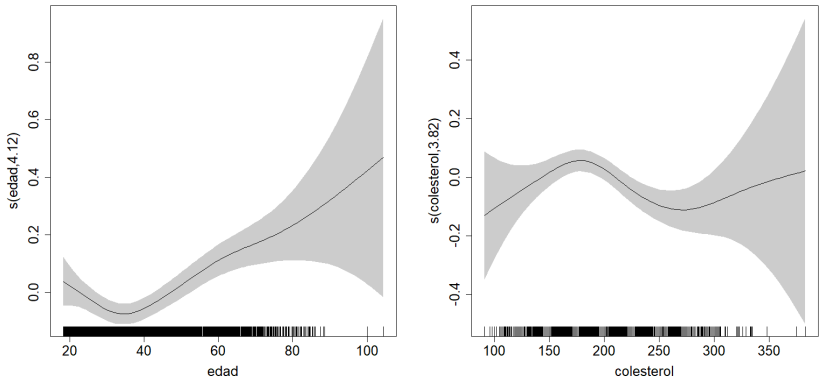


Figura 5.9: Efectos de suavizado sobre a media da presión arterial diastólica das variables *cintura*, *idade*, *imc*, *colesterol*, *g2h*, *insulina*, *alcohol* e *triglicéridos* dentro do modelo GJRM indicado en (5.15) tomando como función cópula unha T-Student e con distribucións marxinais gammas.



(a) PAS (b) PAD

Figura 5.10: Efectos de suavizado sobre o parámetro de escala da presión arterial sistólica da variable *idade* na esquerda e efecto sobre a presión arterial diastólica da variable *colesterol* na dereita dentro do modelo GJRM indicado en (5.15) tomando como función cópula unha T-Student e con distribucións marxinais gammas.

Por último veremos as funcións de suavizado do efecto das covariable para o parámetro ρ da función cópula, que mostramos na Figura 5.11.

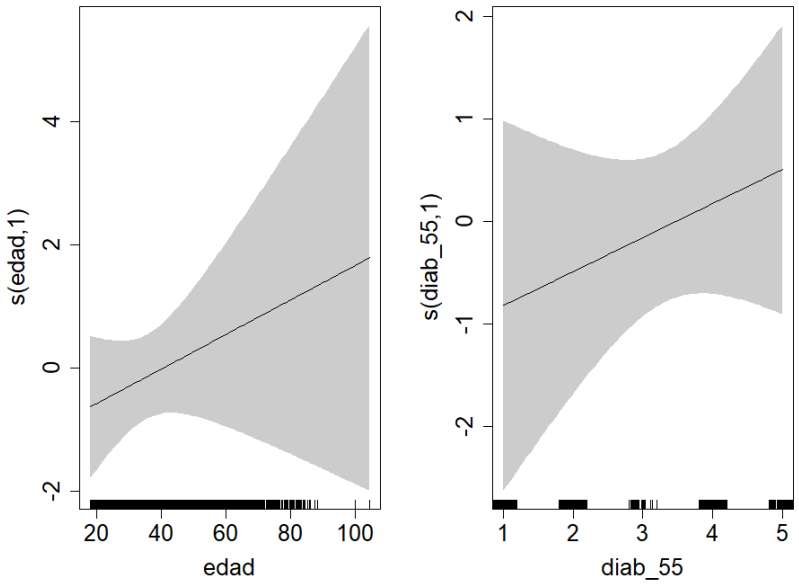


Figura 5.11: Efectos de suavizado sobre o parámetro ρ da función cópula da variable *idade* na esquerda e da variable *diab_55* na dereita dentro do modelo GJRM indicado en (5.15) tomando como función cópula unha T-Student e con distribucións marxinais gammas.

Vemos así como aínda que presentan unha gran variabilidade, polo que debemos ter coidado coas interpretacións que fagamos, parece que o valor do parámetro aumenta a medida que aumenta a idade. Isto quere dicir que para canto maior sexa a idade maior será a correlación existente entre a presión arterial sistólica e a presión arterial diastólica. O mesmo parece que acontece cos grupo da variable *diab_55*, a medida que aumentamos de grupo, a correlación entre as dúas presión é maior.

Bibliografía

[1] Balakrishnan, N., & Lai, C. D. (2009). Continuous bivariate distributions. Springer Science & Business Media.

[2] Brezger, A., & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. Computational Statistics & Data Analysis, 50(4), 967-991.

[3] Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. Statistics in medicine, 11(10), 1305-1319.

[4] De Boor, C., & De Boor, C. (1978). A practical guide to splines (Vol. 27, p. 325). New York: springer-verlag.

[5] Durante F. & Sempi C. (2015) Principles of copula theory. New York: Chapman and Hall/CRC.

[6] Durbán, M. (2009). An introduction to smoothing with penalties: P-splines. Boletín de Estadística e Investigación Operativa, 25(3), 195-205.

[7] Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89-121.

[8] Espasandín-Domínguez, J., Cadarso Suárez, C., Kneib, T., Marra, G., Klein, N., Radice, R., Lado-Baleato, O., González-Quintela, A. & Gude, F. (2019). Assessing the relationship between markers of glycemic control through flexible copula regression models. Statistics in Medicine, 38(27), 5161-5181.

[9] Espasandín Domínguez, J. (2019). Contributions to distributional regression models. Applications in biomedicine. Tese. Universidade de Santiago de Compostela

[10] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression. Models, methods and applications*. Heidelberg, Germany: Springer.

[11] Gu, C. (1992). Cross-validating non-Gaussian data. Journal of Computational and Graphical Statistics, 1(2), 169-179.

[12] Hastie, T. and Tibshirani, R. (1990). Generalized additive models. Boca Raton, USA: Chapman and Hall/CRC Monographs on Statistics and Applied Probability.

[13] Hastie, T. J., & Tibshirani, R. J. (2017). Generalized additive models. Routledge.

[14] Hofert, M., & Mächler, M. (2011). Nested Archimedean copulas meet R: The nacopula package. Journal of Statistical Software, 39(9), 1-20.

[15] Joe, H. (2014). Dependence modeling with copulas. CRC press.

[16] Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. The Annals of Applied Statistics, 9:1024-1052.

[17] Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach. Statistics and Computing, 26:841-860.

[18] Lado Baleato, O. (2017). Bivariate copula regression models in diabetes research. Traballo Final de Máster. Universidade de Santiago de Compostela.

[19] Marra, G. & Radice, R. (2017). Bivariate copula additive models for location, scale and shape. Computational Statistics & Data Analysis, 112, 99-113.

[20] McCullagh, P. and Nelder, J. A. (1989). Generalized linear models. London, New York: Chapman and Hall/CRC Monographs on Statistics and Applied Probability.

[21] Nelder, J. A., and Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.

[22] Nelsen, R. (2006). An introduction to copulas. Heidelberg, Germany: Springer- Verlag.

[23] Palaro, H. P. and Hotta, L. K. (2006). Using conditional copula to estimate value at risk. Journal of Data Science, 4:93-115.

[24] Perez-Fernandez, R., Marino, A. F., Cadarso-Suarez, C., Botana, M. A., Tome, M. A., Solache, I., Rego-Iraeta, A & Mato, A. J. (2007). Prevalence, awareness, treatment and control of hypertension in Galicia (Spain) and association with related diseases. Journal of human hypertension, 21(5), 366-373.

[25] Rigby, R. A., & Stasinopoulos, D. M. (1996a). A semi-parametric additive model for variance heterogeneity. Statistics and Computing, 6(1), 57-65.

[26] Rigby, R. A., & Stasinopoulos, M. D. (1996b). Mean and dispersion additive models. In Statistical theory and computational aspects of smoothing (pp. 215-230). Physica-Verlag HD.

[27] Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), 507-554.

[28] Ruppert , D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric regression. Cambridge University Press.

[29] Rigby, RA y Stasinopoulos, DM (2005). Modelos aditivos generalizados para ubicación, escala y forma. Revista de la Royal Statistical Society: Serie C (Estadística aplicada) , 54 (3), 507-554.

[30] Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). Flexible regression and smoothing: using GAMLSS in R. CRC Press.

[31] Wahba, G. (1990). Spline models for observational data. Society for industrial and applied mathematics.

[32] Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 95-114.

[33] Wood, S. N. (2006). Generalized additive models: an introduction with R. chapman and hall/CRC.

[34] Wood, S. N. (2017). Generalized additive models: An introduction with R. Boca Raton, USA: Chapman and Hall/CRC Texts in Statistical Science.

[35] Yee, T. W. (2015). Vector generalized linear and additive models: with an implementation in R (Vol. 10, pp. 978-1). New York: springer.