Trabajo Fin de Máster

Modelos Aditivos Xeneralizados para Localización, Escala e Forma (GAMLSS)

María Jesús Pérez Pena

Máster en Técnicas Estadísticas

Curso 2015-2016

Proposta de Traballo Fin de Máster

Título en galego: Modelos Aditivos Xeneralizados para Localización, Escala e Forma (GAMLSS)

Título en castellano: Modelos Aditivos Generalizados para Localización, Escala y Forma (GAMLSS)

English title: Generalized Additive Model for Location, Scale and Shape (GAMLSS)

Modalidade: Modalidade B

Autora: María Jesús Pérez Pena, Universidade de Vigo.

Directora: Carmen Cadarso Suárez, Universidade de Santiago de Compostela.

Tutor: Francisco Gude Sampedro, Hospital Clínico Universitario de Santiago.

Breve resumo do traballo:

Determinar a distribución de marcadores clínicos en poboación xeral é importante para poder interpretar valores de referencia. As guías para definir e determinar valores de referencia fan unha serie de recomendacións dirixidas a considerar subgrupos definidos tales como xénero, distribución por grupos de idade e exposicións habituais como consumo de tabaco e alcohol. Os análises denominados "Generalized Additive Model for Location, Scale and Shape" (GAMLSS) refírense a un grupo de modelos estatísticos desenvolvidos por Rigby e Stasinopoulos, permiten modelar a media ou mediana da variable dependente e, ademais, a variabilidade e a asimetría, en relación coas variables independentes. A tarefa será a de realizar aproximacións ás distribucións de biomarcadores para o establecemento de valores de referencia utilizando GAMLSS.

Recomendacións:

Outras observacións:

Dona Carmen Cadarso Suárez, Catedrática de Bioestatística da Universidade de Santiago de Compostela, e don Francisco Gude Sampedro, Adxunto da Unidade de Epidemioloxía do Hospital Clínico Universitario de Santiago, informan que o Traballo Fin de Máster titulado

Modelos Aditivos Xeneralizados para Localización, Escala e Forma (GAMLSS)

foi realizado baixo a súa dirección por dona María Jesús Pérez Pena para o Máster en Técnicas Estadísticas. Estimando que o traballo está acabado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 1 de xullo de 2016.

A directora: O tutor:

Dona Carmen Cadarso Suárez Don Francisco Gude Sampedro

A autora:

Dona María Jesús Pérez Pena

Agradecementos

Ós meus titores Francisco Gude e Carmen Cadarso por axudarme, guiarme e aconsellarme durante todos estes meses.

A todos aqueles que pasaron pola Unidade de Epidemioloxía Clínica do Complexo Hospitalario Universitario de Santiago para traernos os seus casos e permitirme así aprender e realizar este traballo.

Ós meus amigos, ós de sempre, por apoiarme aínda que non entendesen o meu amor polas matemáticas e ós que comparten ese amor comigo porque sempre estaremos unidos por iso, sobre todo a Daniel Mato por estar aí dende o primeiro ata o último día.

Á miña familia por apoiarme en todo momento, sobre todo á miña irmá por axudarme en todo aquilo que precisei.

Índice xeral

Resumo

Resumo en galego

Os modelos aditivos xeneralizados para a localización, escala e forma (GAMLSS) son uns modelos de regresión univariante que se definen co obxectivo de superar varias das limitacións dos modelos xa existentes neste campo como poden ser os modelos lineais xeneralizados (GLM) ou os modelos aditivos xeneralizados (GAM). Os GAMLSS permiten levar a cabo diversas melloras, entre as máis importantes están poder abandonar a familia exponencial para escoller a distribución dos datos de entre unha ampla gama de opcións, incluídas algunhas que presentan unha forte asimetría e/ou curtose. Outra das vantaxes destes modelos é a posibilidade de modelar todos os parámetros da distribución directamente a partir das variables explicativas.

Neste traballo intentarase dar unha idea do que son os GAMLSS. Comezarase expoñendo cal é a forma dun GAMLSS e cales son as súas partes para, posteriormente, explicar a súa implementación en R. Remataremos o traballo motivando a necesidade dos GAMLSS coa utilización destes modelos para tratar varios exemplos reais.

English abstract

The Generalised Additive Models for Location, Scale and Shape (GAMLSS) are a kind of univariate regression models designed to overcome several of the limitations of the preexisting models in this framework, such as the Generalised Linear Models (GLM) and the Generalised Additive Models (GAM). The GAMLSS feature several improvements over the aforementioned models, the most remarkable one being the possibility of choosing the data distribution between a broad selection, instead of the usual exponential family, including distributions that show high asymmetry and/or kurtosis. Another remarkable advantage of the GAMLSS is that it allows us to model all the parameters of the distribution directly from the explicative variables.

The aim of this work is to provide an insight of what the GAMLSS are. We start by introducing what a GAMLSS is and in which parts it consists of, to then explain its implementation in R. We wrap things up by motivating the necessity of the GAMLSS by using these models to study several real examples.

Prefacio

No campo das técnicas estatísticas relacionadas co modelado de regresións univariantes os modelos máis utilizados foron o modelo lineal xeneralizado (GLM) e o modelo aditivo xeneralizado (GAM), véxase Nelder e Wedderburn (1972) e Hastie e Tibshirani (1990) respectivamente. Ambos modelos asumen unha distribución pertencente á familia exponencial para a variable resposta $\boldsymbol{y}$ no que a media μ se modela en función das variables explicativas e a varianza de $\boldsymbol{y}$ vén dada por unha función do seguinte tipo $Var(\boldsymbol{y}) = \phi v(\mu)$, que depende dun parámetro de dispersión constante ϕ e da media. Ademais, tanto a asimetría como a curtose, son modeladas a través da media e do parámetro ϕ. Polo tanto, nos modelos GLM e GAM a varianza, a asimetría e a curtose non son modeladas explicitamente a partir das variables explicativas senón implicitamente a partir da súa relación coa media.

Outra clase importante de modelos son os modelos mixtos lineais (efectos aleatorios) que proporcionan un marco moi amplo para o modelado de datos dependentes do espazo ou do tempo. Estes asumen normalidade na distribución condicionada de $\boldsymbol{y}$ dados os efectos aleatorios, polo tanto non se poden modelar nin a asimetría nin a curtose de forma explícita.

O modelo mixto lineal xeneralizado (GLMM) combina un modelo lineal xeneralizado (GLM) cun modelo mixto lineal mediante a introdución dun termo, polo xeral normal, de efectos aleatorios no preditor lineal para a media do GLM. De igual xeito, pódese combinar un modelo aditivo xeneralizado (GAM) cun modelo mixto dando lugar a un modelo mixto aditivo xeneralizado (GAMM).

Aínda que os modelos GLMM e GAMM son máis flexibes que o GLM e o GAM, estes tamén asumen unha familia exponencial para a distribución condicionada da variable reposta $\boldsymbol{y}$ e tampouco permiten o modelado dos parámetros distintos da media (ou parámetro de localización) a partir das variables explicativas.

Dadas estas limitacións viuse a necesidade de definir uns novos modelos onde ambos feitos quedasen flexibilizados, é dicir, necesítanse uns modelos que permitan utilizar distribucións distintas ás pertencentes á familia exponencial e, tamén, que permitan axustar os parámetros distintos da media directamente a partir da relación entre as variables explicativas e os distintos parámetros en vez de utilizar a relación das variables explicativas coa media e desta, á súa vez, cos restantes parámetros.

Coa finalidade de superar estas limitacións defínense os modelos aditivos xeneralizados de localización, escala e forma (GAMLSS). Estes modelos conseguen flexibilizar non só estas limitacións senón que engaden outras melloras.

A día de hoxe, os modelos GAMLSS dispoñen dun amplo abanico de distribucións (sobre 80) ademais da posibilidade de construír outras novas. Tamén permiten modelar todos os parámetros da distribución directamente a partir das covariables. Ademais, entre outras aportacións, amplían o abanico de suavizadores utilizados nos GAM.

Neste Traballo Fin de Máster comezarase expoñendo a necesidade de utilizar os modelos aditivos xeneralizados para a localización, escala e forma (GAMLSS) a través dun exemplo, vendo neste a necesidade de evolucionar dende os modelos LM aos modelos GAMLSS debido ás limitacións de cada un dos modelos intermedios. Máis adiante explicarase detalladamente que son os GAMLSS, como se pode traballar con eles en R e, finalmente, exemplificarase todo o exposto a partir de varios exemplos reais ós que se lle deu solución durante o período de prácticas na Unidade de Epidemioloxía do CHUS.

Nese último capítulo práctico expoñeranse 3 exemplos que son proba da necesidade dos GAMLSS. Os 2 primeiros exemplos poden englobarse dentro do proxecto AEGIS. Este proxecto é un estudo

transversal que consta de 2 partes, unha de inflamación e outra de glicación. Para a primeira parte disponse de 1516 participantes mentres que na segunda se dispón de 581. Os 2 exemplos expostos neste traballo están relacionados cada un con cada unha destas partes. Na parte de inflamación traballarase coa velocidade de sedimentación globular (VSG), variable relacionada coa inflamación. Para a parte de glicación traballarase con algúns índices de variabilidade da glicosa.

Por último, introducirase un exemplo que non está enmarcado no proxecto AEGIS pero que se considerou de interese introducir para mostrar a utilización de modelos mixtos xunto cos GAMLSS.

Capítulo 1

Que había antes do GAMLSS?

1.1. Introdución.

Neste primeiro capítulo trataremos de motivar a aparición dos GAMLSS. Para isto comezaremos cunha primeira sección, Sección 1.2, onde expoñeremos os modelos máis importantes existentes antes do GAMLSS. Nunha segunda sección, Sección 1.3, traballaremos cun exemplo que nos permita ver a necesidade de cada un dos modelos expostos na Sección 1.2 ata chegar, por último, aos GAMLSS.

O recorrido que se fará na seguinte sección pasará polos modelos lineal (LM), lineal xeneralizado (GLM) e aditivo xeneralizado (GAM). O modelo de regresión univariante máis sinxelo co que nos podemos atopar é o LM. Este modelo, aínda que sinxelo, logra cubrir un amplo espectro de casos a pesar das súas estritas restricións como poden ser o feito de que a variable resposta debe adaptarse a unha distribución normal ou ben que a relación entre a variable, ou as variables explicativas, e a variable resposta debe asumirse lineal.

Para flexibilizar lixeiramente o modelo LM aparece o modelo GLM. Este modelo segue a considerar lineais as relacións entre a variable resposta e as covariables do modelo. Non obstante, permite que a distribución asumida pola variable resposta non sexa necesariamente a distribución normal, podendo ser esta unha Poisson ou unha binomial, entre outras.

O seguinte paso, antes de chegar aos modelos GAM, foi intentar construír relacións non lineais entre a variable resposta e as covariables do modelo utilizando, por exemplo, os polinomios. Aínda así, as flexibilizacións feitas puntualmente sobre os GLM non semellaban suficientes e deron lugar a aparición dos modelos GAM. Estes últimos permiten a utilización de funcións suavizadoras para describir a relación entre a variable resposta e as variables explicativas dun xeito non paramétrico. Non obstante, e a pesar das modificacións feitas sobre os GAM, todos estes modelos quedan obsoletos para aquelas distribucións que teñen máis dun parámetro (que son a maioría delas) xa que todos os modelos mencionados ata agora só son capaces de modelar o parámetro de localización directamente a partir das covariables. Nalgunhas ocasións tamén se logra modelar o parámetro de escala, pero sempre a través da súa relación co parámetro de localización; por esta razón apareceron os GAMLSS. Ditos modelos permiten a modelización do parámetro de escala, e incluso dos parámetros de forma (asimetría e curtose) da distribución utilizada directamente a partir das variables explicativas.

Nótese que ao longo desta introdución só se mencionaron modelos de regresión univariante, a pesar de que existen modelos de regresión multivariante previos ós modelos GAMLSS. Pero dado que os modelos deste traballo nos que máis fincapé faremos son os modelos GAMLSS, e estes son modelos de regresión univariante, deixaremos de lado os modelos de regresión multivariante.

Vexamos na seguinte sección a evolución dos modelos mencionados anteriormente.

1.2. Evolución, de LM a GAMLSS.

Nesta sección, como xa se mencionou na introdución, faremos un pequeno repaso da evolución que se levou a cabo dende os modelos LM aos modelos GAMLSS, pasando polos modelos GLM e GAM.

- Modelos lineais (LM)

  Os modelos lineais son un tipo de modelos de regresión univariante nos que, entre outras, se fan dúas asuncións. A variable resposta considérase que segue unha distribución normal e, ademais, que a relación existente entre as variables explicativas e a variable resposta ten unha forma lineal. Podemos expresar o modelo LM do seguinte xeito:

  $$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ri} + \epsilon_i$$

  onde $\epsilon_i \sim N(0, \sigma^2)$ para $i = 1, 2, \ldots, n$. Para estes modelos asúmese que os erros, ϵ_i, son independentes e seguen unha distribución normal de media cero e desviación típica constante. Podemos reescribir o modelo do seguinte xeito:

  $$Y_i \sim N(\mu_i, \sigma^2)$$
  $$\mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{pi}$$

  para $i = 1, 2, \ldots, n$. Utilizando unha notación matricial podemos de novo reescribir o modelo da seguinte maneira:

  $$\boldsymbol{y} \sim N(\mu, \sigma^2)$$
  $$\mu = \boldsymbol{X}\boldsymbol{\beta}$$

  onde $\boldsymbol{X}$ é unha matriz $n \times p$ ($p = r + 1$) que contén todas as variables explicativas (ademais dunha columna de 1 se se necesita a constante) e $\boldsymbol{\beta}$ é un vector descoñecido de lonxitude p que será estimado a partir das variables explicativas. Esta última notación será a utilizada para relacionar os modelos LM cos modelos GAMLSS.

  A pesar de que a distribución normal é adecuada para numerosos e importantes exemplos estudados dende que se empezou a traballar co concepto de regresión, tamén se viu que a distribución normal non pode abarcar todos os casos existentes. Atendendo á necesidade de que a variable resposta dun modelo de regresión non siga unha distribución normal aparecen os GLM, que son introducidos no seguinte punto.

- Modelos lineais xeneralizados (GLM)

  A principios da década dos 70 Nelder e Wedderburn (1972) propuxeron os modelos lineais xeneralizados (GLM), que englobaban varios modelos de regresión xa existentes, proporcionando un marco unificador para aqueles modelos nos que a distribución da variable resposta pertence á familia exponencial como poden ser, por exemplo, os modelos de regresión lineal ou de regresión loxística.

  Pódese considerar que as melloras máis importantes introducidas polos modelos GLM foron as seguintes:

  - A suposición de que a variable resposta segue unha distribución normal foi substituída pola suposición de que a variable resposta segue unha distribución da familia exponencial, podendo incluír así distribucións como a gamma ou a Poisson.
  - Comezouse a utilizar unha función link monótona, denotada por g(.), para modelar a relación existente entre as variables explicativas e a media da distribución escollida para a variable resposta.

Atendendo a estas consideracións podemos escribir, utilizando unha notación matricial, o modelo GLM do seguinte xeito:

$$\boldsymbol{y} \sim ExpF(\mu, \phi)$$
$$g(\mu) = \boldsymbol{X}\boldsymbol{\beta}$$

onde a denominación $ExpF$ se refire á familia exponencial.

Para cada observación, a función de densidade de probabilidade da familia exponencial pode escribirse como segue:

$$f_Y(y; \mu, \sigma) = \exp^{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)}$$

onde $E(y) = \mu = b'(\theta)$ e $Var(y) = \phi v(\mu)$ sendo $v(\mu) = b''[\theta(\mu)]$.

Coa introdución dos modelos GLM conséguese flexibilizar considerablemente a restrición de que a distribución da variable resposta tivese que ser necesariamente normal. Non obstante, o feito de que a relación entre a variable resposta e as variables explicativas fose non lineal seguía pendente de modelización a pesar de que, como xa se mencionou anteriormente, antes da aparición dos GAM tratouse de traballar con modelos GLM aos que se lle introduciron, por exemplo, polinomios. Polo tanto, a relaxación da suposición de linealidade foi a principal motivación para a aparición dos modelos aditivos xeneralizados (GAM) que se expoñen no seguinte apartado.

- Modelos aditivos xeneralizados (GAM)

  Os modelos GAM fixéronse populares na década dos 80. Os primeiros en introducir estes modelos foron Hastie e Tibshirani (1990) e, posteriormente, foron estendidos por Wood (2006). Ditos modelos pódense considerar unha extensión non paramétrica dos GLM; a idea dos modelos GAM é permitir que sexan os propios datos os que determinen a relación entre o preditor lineal η e as variables explicativas.

  O modelo GAM pódese escribir do seguinte xeito:

$$\boldsymbol{y} \sim ExpF(\mu, \phi)$$
$$g(\mu) = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{J} h_j(x_j)$$

  onde h son funcións de suavizado non paramétricas que se aplican sobre aquelas variables explicativas continuas que non presentan unha relación lineal con respecto á variable resposta.

  Coa introdución dos modelos GAM quedan flexibilizadas as dúas condicións máis restritivas dos modelos LM que eran a imposibilidade de considerar unha familia distinta á normal e de considerar unha relación non lineal entre a variable resposta e as distintas variables explicativas. Aínda así, todo é mellorable e coa idea de introducir novas distribucións fóra da familia exponencial e de poder modelar o parámetro de escala e os parámetros de forma (se a distribución dispuxese deles) directamente a partir das variables explicativas, sen necesidade de utilizar a relación existente entre o parámetro de localización e os restantes parámetros, aparecen os modelos GAMLSS dos que falaremos ao longo do traballo.

Ademais dos modelos expostos nesta sección, cabe mencionar os modelos lineais xeneralizados mixtos (GLMM) e os modelos aditivos xeneralizados mixtos (GAMM) que permiten introducir efectos aleatorios nos modelos GLM e GAM respectivamente. Veremos como ditos efectos aleatorios poden ser tamén introducidos nos modelos GAMLSS.

Na seguinte sección traballaremos cun exemplo que nos permitirá ir vendo a necesidade de, partindo dun modelo LM, chegar a un modelo GAM. Na Sección 3.7 do Capítulo 3 retomaremos este exemplo para traballar sobre el, utilizando modelos GAMLSS, e así comprobar a necesidade de aplicar estes modelos.

1.3. De LM a GAMLSS mediante un exemplo real.

Nesta sección traballaremos cun exemplo que se expoñerá de forma máis detallada no último capítulo deste traballo. A súa presenza nesta sección vén dada pola necesidade de xustificar a evolución dende os modelos LM ata os modelos GAM (neste caso, incluíndo efectos aleatorios).

Construíronse 4 modelos distintos:

- O modelo mlm é un modelo lineal con variable resposta a variable ganancia e variable explicativa a variable edad utilízase a distribución normal que é a única dispoñible para este modelo.
- O modelo mglm é un modelo lineal xeneralizado con variable resposta a variable ganancia e variable explicativa a variable edad pero, neste caso, a distribución utilizada foi a gaussiana inversa.
- O modelo mgam é un modelo aditivo xeneralizado con variable resposta a variable ganancia e variable explicativa a variable edad pero, neste caso, a distribución utilizada foi a gaussiana inversa e aplicouse suavización sobre a variable edad.
- O modelo mgamm é un modelo mixto aditivo xeneralizado con variable resposta a variable ganancia e variable explicativa a variable edad pero, neste caso, a distribución utilizada foi a gaussiana inversa, aplicouse suavización sobre a variable edad e incluíronse os efectos aleatorios da variable id.

As liñas de comando utilizadas para axustar os modelos descritos anteriormente son as que se poden ver na parte esquerda da Figura 1.1. A partir delas obtívose a táboa que aparece á dereita na mesma figura. Neste táboa pode verse como o valor do AIC decrece, é dicir, se nos guiamos polo criterio AIC o modelo mglm é mellor que o mlm pero, a súa vez, o mgam é mellor que o mlm e o mglm e, por último, o modelo mgamm é mellor que os 3 anteriores. Este exemplo recuperarase ó final do Capítulo 3 para ver como o modelo GAMLSS mellora estes modelos, obténdose un menor valor de AIC.

```
> mlm<-lm(ganancia~edad,data=dath)

> mglm<-glm(ganancia~edad,data=dath,
            family=inverse.gaussian)

> mgam<-gam(ganancia~s(edad),data=dath,
              family=inverse.gaussian)

> mgamm<-gam(ganancia~s(edad),data=dath,
                family=inverse.gaussian,
                     random=list(id=~1))
```

Modelo	AIC
mlm	-91.75
mglm	-98.28
mgam	-105.75
mgamm	-106.53

Figura 1.1: Á esquerda aparecen as liñas de comando utilizadas para axustar os modelos mlm, mglm, mgam e mgamm. Á dereita unha táboa que recolle os valores do AIC destes modelos.

Capítulo 2

GAMLSS

2.1. Introdución

O modelo lineal xeneralizado (GLM) e o modelo aditivo xeneralizado (GAM) ocupan un lugar importante no campo das técnicas estatísticas para a regresión univariante, véxase Nelder e Wedderburn (1972) e Hastie e Tibshirani (1990) respectivamente. Estes dous modelos consideran, para a variable resposta, unha distribución da familia exponencial. A media é modelada a partir das variables explicativas mentres que a varianza vén dada pola seguinte expresión $Var(y) = \phi v(\mu)$, onde ϕ é un parámetro de dispersión constante e, polo tanto, depende da media. Ademais, se se considera unha distribución da familia exponencial, tanto a asimetría como a curtose están en función da media e do parámetro ϕ. Polo tanto, se se traballa con modelos GLM ou GAM nin a varianza, nin a asimetría, nin a curtose son modeladas explicitamente a partir das variables explicativas senón a través da relación das variables independentes coa media e, a desta última, co resto dos parámetros.

Debido a isto, os modelos aditivos xeneralizados de localización, escala e forma (GAMLSS) foron introducidos por Rigby e Stasinopoulos (2001, 2005) e Akantziliotou et al. (2002) como un xeito de solucionar algunhas das limitacións que presentaban os modelos lineais xeneralizados (GLM) e os modelos aditivos xeneralizados (GAM).

Os modelos GAMLSS son unha clase de modelos univariantes. Ademais, os GAMLSS poden considerarse modelos de regresión semi-paramétricos. Son paramétricos no sentido no que necesitan unha distribución paramétrica para a variable resposta e semi no sentido no que o modelado dos parámetros da distribución en función das variables explicativas pode implicar o uso de funcións suavizadas non paramétricas.

Nos modelos GAMLSS a suposición de que a variable resposta $\boldsymbol{y}$ pertence á familia exponencial é flexibilizada, permitindo así traballar con distribucións con moita asimetría ou curtose. Polo tanto, a distribución da variable resposta $\boldsymbol{y}$ pode ser seleccionada de entre unha ampla gama de distribucións, incluídas aquelas con unha alta asimetría ou curtose, tanto continuas como discretas.

Ademais, a parte sistemática do modelo amplíase permitindo modelar non só a media (parámetro de localización) senón tamén a varianza (parámetro de escala), a asimetría ou a curtose (parámetros de forma). Inclúe tanto relacións lineais como non lineais entres os parámetros e as distintas variables explicativas.

Ó longo deste capítulo, máis do seguinte, botaremos man asiduamente das seguintes referencias: Rigby e Stasinopoulos (2005), Rigby e Stasinopoulos (2007) e, fundamentalmente, Stasinopoulos et al. (2015c).

2.2. GAMLSS: o modelo

Se recuperamos a forma do GAM, o modelo máis completo visto ata agora (en conxunto cos modelos mixtos, GAMM),

$$g(\mu) = \boldsymbol{X\beta} + \sum_{j=1}^{J} h_j(x_j)$$

e temos en conta que o modelo GAMLSS modela non só a media senón todos os parámetros da distribución a partir das variables explicativas, tense o modelo

$$g_1(\mu) = \boldsymbol{X_1\beta_1} + \sum_{j=1}^{J_1} h_{j1}(x_{j1})$$

$$g_2(\sigma) = \boldsymbol{X_2\beta_2} + \sum_{j=1}^{J_2} h_{j2}(x_{j2})$$

$$g_3(\nu) = \boldsymbol{X_3\beta_3} + \sum_{j=1}^{J_3} h_{j3}(x_{j3})$$

$$g_4(\tau) = \boldsymbol{X_4\beta_4} + \sum_{j=1}^{J_4} h_{j4}(x_{j4})$$

xa que na maioría dos casos prácticos se teñen como moito 4 parámetros, sendo estes a media (parámetro de localización), a varianza (parámetro de escala), a asimetría e a curtose (parámetros de forma). Estes 4 parámetros denotaranse por μ ó parámetro de localización (media), σ ó parámetro de escala (varianza) e ν e τ ós parámetros de forma (asimetría e curtose, respectivamente).

Se queremos xeneralizar un pouco a notación utilizada podemos escribir:

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

onde $\boldsymbol{\theta_k}$ é o vector de parámetros da distribución.

Non obstante, a notación utilizada será a que se expón a continuación: isto é debido ás numerosas funcións de penalización existentes nos modelos GAMLSS e, tamén, como consecuencia de utilizar un algoritmo backfitting. As funcións $h(x)$ quedan reemplazadas pola expresión $\boldsymbol{Z\gamma}$ ($h(x) = Z\gamma$) onde $\boldsymbol{Z}$ é unha matriz base que depende da variable explicativa x. Finalmente, o modelo GAMLSS quedará descrito do seguinte xeito:

$$\boldsymbol{y} \sim D(\boldsymbol{\theta_k})$$

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{X_k\beta_k} + \sum_{j=1}^{J_k} \boldsymbol{Z_{jk}\gamma_{jk}}$$

onde D é a distribución da variable resposta $\boldsymbol{y}$. $\boldsymbol{X_k}$ e $\boldsymbol{Z_{jk}}$ para $k = 1, 2, 3, 4$ e $j = 1, \ldots, J_k$ son matrices de deseño para os termos lineais e suavizados respectivamente. $\boldsymbol{\beta_k}$ é un vector de parámetros de lonxitude J_k e $\boldsymbol{\gamma_{jk}}$ é unha variable aleatoria q_{jk}-dimensional.

En GAMLSS a estimación lévase a cabo maximizando a seguinte función de verosimilitude penalizada:

$$l_p = l - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \boldsymbol{\lambda_{jk}\gamma'_{jk}G_{jk}\gamma_{jk}}$$

onde $l = \sum_{i=1}^{n} logf(y_i|\theta_i)$ é o logaritmo da función de verosimilitude, $\boldsymbol{\lambda_{jk}}$ son os parámetros de penalización e $\boldsymbol{G_{jk}}$ é unha matriz simétrica que depende dos parámetros $\boldsymbol{\lambda_{jk}}$.

Polo tanto, debemos estimar $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ e $\boldsymbol{\gamma}$.

Para o axuste dos modelos GAMLSS existen 2 algoritmos: o RS que se basea no algoritmo descrito en Rigby e Stasinopoulos (1996a) e o CG que se basea no algoritmo de Cole e Green (1992). Ambos algoritmos describiranse na Subsección 2.2.1. Se se desexa afondar máis pódese acudir a Rigby e Stasinopoulos (2005) ou Stasinopoulos (2015).

Sexa $\boldsymbol{M} = \{\boldsymbol{D}, \boldsymbol{G}, \boldsymbol{T}, \boldsymbol{\Lambda}\}$ un modelo GAMLSS onde

- D especifica a distribución da variable resposta
- G especifica o conxunto de funcións link $(g_1, ..., g_p)$ para os parámetros $(\theta_1, ..., \theta_p)$
- T engloba as variables utilizadas nos preditores correspondentes a μ, σ, ν e τ
- Λ engloba os distintos parámetros de suavizado para as funcións de suavizado

Para un conxunto de datos en concreto, o proceso de selección consiste en comparar moitos modelos construídos combinando os distintos compoñentes de $\boldsymbol{M}$. Falemos agora de cada unha destas compoñentes por separado.

- Compoñente D:

  A selección dunha distribución apropiada conta con 2 pasos. O primeiro paso será axustar diferentes modelos utilizando diferentes distribucións e quedándonos con aquela que proporcione un modelo con menor GAIC. O seguinte paso será validar dita escolla a través dun proceso de diagnose como pode ser o worm plot.

  As familias existentes relacionadas cos modelos GAMLSS son numerosas, no Apéndice A podemos ver no Cadro A.1, as distribucións continuas; no Cadro A.2, as distribucións discretas e no Cadro A.3, as mesturas de distribucións existentes ata o momento. Pero as distribucións GAMLSS non acaban aí. Existen dous paquetes, gamlss.tr e gamlss.cens, que nos permiten construír milleiros de distribucións a partir das xa existentes sen máis que definir estas como distribucións truncadas ou censuradas.

  Se aínda así estas non fosen suficientes, o propio usuario ten a posibilidade de crear unha nova distribución como se describe en Stasinopoulos et al. (2015c).

  NOTA: A única restrición que se lle impón á distribución da variable resposta é que a función $f(\boldsymbol{y}|\boldsymbol{\theta})$ e a súa primeira derivada respecto a cada un dos parámetros de $\boldsymbol{\theta}$ deben poder calcularse.

- Compoñente G:

  A selección da función link ven determinada xeralmente polo rango de valores da variable explicativa.

  Unha boa escolla da función link pode mellorar o axuste do modelo considerablemente; a escolla desta función farase usando o criterio deviance (quedarémonos co de menor deviance).

  Nos Cadros A.1, A.2 e A.3 aparecen, xunto con cada distribución, as funcións link que se utilizan por defecto.

- Compoñente T:

  A selección dos termos aditivos do modelo pode levarse a cabo a través de procesos forward, backward ou stepwise. Ademais, estes procesos poden aplicarse sobre cada parámetro por separado ou sobre todos os parámetros á vez.

  Dependendo da cantidade de datos dispoñibles podemos utilizar distintos métodos de selección. No seguinte capítulo expoñeranse algunhas das funcións utilizadas en R para levar a cabo os propósitos anteriores.

Con respecto aos termos aditivos debemos diferenciar entre os paramétricos e os suavizados.

En GAMLSS o preditor lineal $\boldsymbol{\eta_k}$, para $k = 1, 2, ..., p$, está composto por unha compoñente paramétrica $\boldsymbol{X_k\beta_k}$ e unha compoñente aditiva $\boldsymbol{Z_{jk}\gamma_{jk}}$, para $j = 1, ..., J_k$.

A compoñente paramétrica $\boldsymbol{X_k\beta_k}$ pode incluír termos lineais, interacción entre as variables explicativas (tanto continuas como factores), polinomios, polinomios fraccionados, polinomios a cachos (con nodos fixos). Tamén é posible nalgúns casos incluír relacións non lineais.

Os compoñentes aditivos $\boldsymbol{Z_{jk}\gamma_{jk}}$ do modelo GAMLSS poden modelar unha gran variedade de termos tales como os suavizados ou os efectos aleatorios, así como termos útiles para o análise de series de tempo. Nos modelos GAMLSS poden incluírse distintos termos aditivos, algúns deles (os utilizados ó longo do traballo) son introducidos na Subsección 2.2.2 xunto coa compoñente paramétrica.

- Compoñente Λ:

  Para cada termo suavizado necesitaremos o seu correspondente parámetro de suavizado. Este pode ser previamente fixado ou estimado a partir dos datos.

  Xeralmente cando o que se fai é prefixar o parámetro de suavización, faise a partir dos grados efectivos de liberdade, Hastie e Tibshirani (1990). Non obstante, é máis aconsellable estimar o parámetro de suavización automaticamente. Isto pode facerse a partir dos seguintes métodos:

  - Validación cruzada xeneralizada (GCV).
  - Criterio Akaike xeneralizado (GAIC).
  - Máxima verosimilitude (ML/REML).

  Cada método pode utilizarse:

  - localmente, cando o método é aplicado dentro do algoritmo iterativo
  - globalmente, cando o método é aplicado fora do algoritmo iterativo

  Xeralmente os métodos locais son máis rápidos e producen resultados similares ós métodos globais. Non obstante, os métodos globais poden ser, ás veces, máis fiables. En como estimar os parámetros $\boldsymbol{\lambda}$ tamén se afondará na Subsección 2.2.1.

Unha vez tido en conta todas as compoñentes do modelo $\boldsymbol{M}$ e construídos varios modelos debemos tratar de escoller un entre todos os propostos.

Á hora de escoller o modelo máis axeitado podemos utilizar algún dos seguintes criterios de selección:

- Deviance Global, $GD = -2l(\hat{\boldsymbol{\theta}})$
- Criterio Akaike, $AIC = -2l(\hat{\boldsymbol{\theta}}) + 2q$
- Criterio Bayesiano de Schwarz, $SBC = -2l(\hat{\boldsymbol{\theta}}) + q\log(n)$

onde $-2l(\hat{\boldsymbol{\theta}}) = -2\sum_{i=1}^{n} l(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$ e q son os graos de liberdade utilizados no modelo, tanto en tódolos termos paramétricos lineais como nos non paramétricos.

Se se dispón dunha mostra pequena sóese utilizar o criterio GD para axustar o modelo e os criterios AIC ou SBC para seleccionalo. Cando a base de datos é suficientemente grande podemos dividir dita mostra en dous, mostra de ensaio e mostra de validación. A mostra de ensaio utilízase para axustar o modelo usando o criterio GD e a mostra de validación úsase para seleccionar o modelo utilizando de novo o criterio GD.

Unha vez axustado o modelo GAMLSS, utilizaranse os residuos cuantiles randomizados descritos en Dunn e Smyth (1996) para comprobar a adecuación do modelo e, máis especialmente, a distribución escollida para a variable resposta $\boldsymbol{y}$.

Cando traballamos cun modelo de regresión lineal, os residuos son definidos como a diferenza entre os valores observados e os valores axustados. Non obstante, cando tratamos de traballar con outra distribución distinta da normal, estes residuos xa non son axeitados e defínense os residuos de Pearson ou os deviance residuos. Desafortunadamente estes últimos tipos de residuos non son adecuados para traballar cos modelos gamlss; neste caso debemos traballar cos denominados residuos cuantiles randomizados.

Os residuos cuantiles randomizados veñen dados pola expresión $r_i = \Phi^{-1}(u_i)$ onde Φ^{-1} é a función inversa da distribución acumulada da variable normal estándar e u_i defínese como:

- $F(y_i|\hat{\theta}_i)$ se y_i é continua.
- un valor aleatorio da distribución uniforme no intervalo $[F((y_i-1)|\hat{\theta}_i), F((y_i|\hat{\theta}_i)]$ se y_i é discreta.

  A principal vantaxe destes novos residuos é que sexa cal sexa a distribución da variable resposta, estes sempre teñen unha distribución normal estándar, sempre e cando o modelo sexa correcto.

2.2.1. Algoritmos

Rigby e Stasinopoulos (2005) proporcionan dous algoritmos básicos para maximizar o logaritmo da función da verosimilitude penalizada con respecto a $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ para un $\boldsymbol{\lambda}$ fixado:

- Algoritmo RS:

  O algoritmo RS é unha xeneralización do algoritmo utilizado por Rigby e Stasinopulos (1996a,b) para axustar a media e a dispersión dos modelos aditivos. Este algoritmo non utiliza as derivadas cruzadas da función logaritmo da verosimilitude.

  O algoritmo RS pode describirse utilizando as seguintes 3 partes (anidadas): a iteración exterior, a iteración interior e o algoritmo backfitting (cara atrás).

  A iteración exterior chama repetidamente á iteración interior e esta, á súa vez, chama repetidamente ó algoritmo backfitting. Considérase que o algoritmo converxe cando converxen as 3 partes.

  - A iteración exterior:

    Tras asignarlle uns valores iniciais ós parámetros da distribución $(\mu_0, \sigma_0, \nu_0, \tau_0)$, esta parte do algoritmo procede do seguinte xeito:

    - Axústase o modelo para μ usando a iteración interior a partir das últimas estimacións feitas para $\hat{\sigma}, \hat{\nu}$ e $\hat{\tau}$.
    - Axústase o modelo para σ usando a iteración interior a partir das últimas estimacións feitas para $\hat{\mu}, \hat{\nu}$ e $\hat{\tau}$.
    - Axústase o modelo para ν usando a iteración interior a partir das últimas estimacións feitas para $\hat{\mu}, \hat{\sigma}$ e $\hat{\tau}$.
    - Por último, axústase o modelo para τ usando a iteración interior a partir das últimas estimacións feitas para $\hat{\mu}, \hat{\sigma}$ e $\hat{\mu}$.

    Posteriormente, calcúlase a deviance global. Se esta non cambiou o algoritmo converxeu e detense, se non converxeu repítese o proceso.

  - A iteración interior:

    Para cada un dos axustes feitos, para cada un dos parámetros da distribución, utilízase a iteración interior. A iteración interior é un algoritmo de puntuación local moi similar ao que se utiliza para axustar os modelos lineais xeneralizados (GLM).

    A idea deste algoritmo local de puntuación é repetir axustes ponderados para modificar a variable resposta usando distintos pesos ata que o algoritmo converxe.

    A variable resposta modificada para axustar os distintos parámetros vén dada por

$$\boldsymbol{z_k} = \boldsymbol{\eta_k} + \boldsymbol{w_k}^{-1} \cdot \boldsymbol{u_k}$$

onde $\boldsymbol{z_k}, \boldsymbol{\eta_k}, \boldsymbol{w_k}$ e $\boldsymbol{u_k}$ son vectores de lonxitude n, $\boldsymbol{w_k}^{-1} \cdot \boldsymbol{u_k}$ é o produto de Hadammard elemento a elemento e

$$\boldsymbol{u_k} = \frac{\partial l}{\partial \boldsymbol{\eta_k}} = \left(\frac{\partial l}{\partial \boldsymbol{\theta_k}}\right) \cdot \left(\frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\eta_k}}\right)$$

é a función de puntuación (a primeira derivada do logaritmo da función de verosimilitude con respecto ao preditor lineal).

Os w_k son os pesos utilizados en cada iteración que se poden definir do seguinte xeito:

$$w_k = -f_k \cdot \left(\frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\eta_k}}\right) \cdot \left(\frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\eta_k}}\right)$$

onde f_k pode definirse das 3 seguintes formas dependendo da distribución coa que se estea a traballar:

1. $E\left[\frac{\partial^2 l}{\partial \boldsymbol{\theta_k}^2}\right]$ se se utiliza o algoritmo de puntuación de Fisher.
2. $\frac{\partial^2 l}{\partial \boldsymbol{\theta_k}^2}$ se se utiliza o algoritmo de puntuación estándar de Newton-Raphson.
3. $-\left(\frac{\partial l}{\partial \boldsymbol{\theta_k}}\right) \cdot \left(\frac{\partial l}{\partial \boldsymbol{\theta_k}}\right)$ se se utiliza o algoritmo de puntuación cuasi Newton-Raphson.

Dadas as estimacións actuais para os parámetros da distribución e os pesos da correspondente iteración, realízase de novo o proceso ata que a deviance global non cambia, é dicir, dadas as estimacións actuais dos parámetros calcúlanse $\boldsymbol{z_k}$ e $\boldsymbol{w_k}$, axústanse as variables explicativas utilizando os valores de $\boldsymbol{z_k}$ e $\boldsymbol{w_k}$ a partir do algoritmo backfitting para recalcular as estimacións do preditor lineal e dos parámetros da distribución. Posteriormente calcúlase a deviance global; se esta non se modificou considérase que o algoritmo converxiu e este para. Noutro caso continúase iterando.

- O algoritmo backfitting

  A estimación dos parámetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ lévase a cabo co algoritmo backfitting. Este algoritmo é unha versión do algoritmo Gauss-Seidel presentado en Hastie e Tibshirani (1990).

  A modificación feita é que para moitos dos suavizadores penalizados a matriz de deseño usada para definir as relacións lineais contén a parte lineal da variable explicativa, isto axuda á converxencia do algoritmo.

  Para o bo funcionamento deste algoritmo é necesario un bo algoritmo de pesos para os mínimos cadrados (WLS) e un bo algoritmo de pesos penalizados para mínimos cadrados (WPLS).

  O algoritmo backfitting traballa da seguinte maneira. Deséxase axustar tanto as variables lineais como as suavizadas para $\boldsymbol{z_k}$ utilizando os pesos $\boldsymbol{w_k}$ dentro do algoritmo backfitting (dentro da iteración interior). Para explicar o proceso asumiremos só 2 suavizadores (como se fixo en Stasinopoulos et al. (2015c)) con parámetros $\boldsymbol{\gamma_{k1}}$ e $\boldsymbol{\gamma_{k2}}$ e bases $\boldsymbol{Z_{k1}}$ e $\boldsymbol{Z_{k2}}$ respectivamente.

  Para $\boldsymbol{z_k}$, $\boldsymbol{w_k}$, $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}}$ e $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$ estimadas anteriormente calcúlanse os residuos parciais para $\boldsymbol{\beta_k}$ ($r = \boldsymbol{z_k} - \boldsymbol{Z_{k1}}\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}} - \boldsymbol{Z_{k2}}\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$) e axústanse cun WLS para obter un novo $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$. Despois obtéñense os residuos parciais con respecto a $\boldsymbol{\gamma_{k1}}$ ($r' = \boldsymbol{z_k} - \boldsymbol{X_k}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} - \boldsymbol{Z_{k2}}\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$) e utilízase o algoritmo PWLS para obter unha nova $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}}$. Posteriormente obtéñense os residuos parciais con respecto a $\boldsymbol{\gamma_{k2}}$ ($r'' = \boldsymbol{z_k} - \boldsymbol{X_k}\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} - \boldsymbol{Z_{k1}}\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}}$) e úsase PWLS para obter unha nova estimación de $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$. O proceso repítese ata que $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}}$ e $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$ non cambian.

  A razón de que se utilice un algoritmo backfitting é porque, deste xeito, pódense utilizar outros suavizadores, ademais dos penalizados.

Vexamos os pasos deste algoritmo máis esquematicamente (os pasos en cor vermella pertencen á iteración exterior, os azuis á iteración interior e os laranxa ó algoritmo backfitting):

- PASO 0: Inicializamos.
- PASO 1: Axustamos un modelo para $\boldsymbol{\theta_k}$ dadas as estimacións dos parámetros restantes.
- PASO 2: Calculamos $\boldsymbol{z_k}$ e $\boldsymbol{w_k}$.
- PASO 3: Calculamos r e obtemos $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$.
- PASO 4: Calculamos r' e obtemos $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}}$.
- PASO 5: Calculamos r'' e obtemos $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$.
- PASO 6: Comprobamos se $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k1}}$ e $\hat{\boldsymbol{\gamma}}_{\boldsymbol{k2}}$ cambian. Se cambian volvemos ó PASO 3, se non imos ó PASO 7.
- PASO 7: Recalculamos $\hat{\boldsymbol{\eta}}_{\boldsymbol{k}}$ e $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$. Se hai converxencia imos ó PASO 8, se non a hai volvemos ó PASO 2.
- PASO 8: Se hai converxencia PARAMOS, se non a hai volvemos ó PASO 1.

- Algoritmo CG

  O algoritmo CG é unha xeneralización do que aparece en Cole e Green (1992). Este algoritmo require información sobre a primeira, a segunda e as derivadas cruzadas da función logaritmo da verosimilitude con respecto ós parámetros da distribución.

  O algoritmo CG é un algoritmo de puntuación local formado por unha iteración interior e outra exterior e o algoritmo backfitting. Á diferenza do algoritmo RS, o algoritmo CG sí necesita as derivadas parciais do logaritmo da función de verosimilitude con respecto a cada un dos parámetros da distribución.

  Na iteración exterior a variable resposta e os pesos iterativos para os parámetros da distribución son definidos como:

$$\boldsymbol{z_k} = \boldsymbol{\eta_k} + {\boldsymbol{w_{ks}}}^{-1} \cdot \boldsymbol{u_k}$$

  e

$$\boldsymbol{w_{ks}} = -f_{ks} \cdot \left(\frac{\partial \boldsymbol{\theta_k}}{\partial \boldsymbol{\eta_k}}\right) \cdot \left(\frac{\partial \boldsymbol{\theta_s}}{\partial \boldsymbol{\eta_s}}\right)$$

  onde f_{ks} pode definirse das 3 seguintes formas dependendo da distribución coa que se estea a traballar:

  1. $-E\left[\frac{\partial^2 l}{\partial {\boldsymbol{\theta_k}}^2}\right]$ se se utiliza o algoritmo de puntuación de Fisher.
  2. $\frac{\partial^2 l}{\partial \theta_k^2}$ se se utiliza o algoritmo de puntuación estándar de Newton-Raphson.
  3. $-\left(\frac{\partial l}{\partial \boldsymbol{\theta_k}}\right) \cdot \left(\frac{\partial l}{\partial \boldsymbol{\theta_s}}\right)$ se se utiliza o algoritmo de puntuación cuasi Newton-Raphson.

  Na iteración interior defínese unha nova variable resposta como segue:

$$\boldsymbol{z'_k} = \boldsymbol{z_k} + \boldsymbol{z_k^a}$$

  onde $\boldsymbol{z_k^a}$ representa o seguinte:

$$\mu : \boldsymbol{z_1^a} = -{\boldsymbol{w_{11}}}^{-1} \cdot \left[\boldsymbol{w_{12}} \cdot \left(\boldsymbol{\eta_2} - \boldsymbol{\eta_2^0}\right) + \boldsymbol{w_{13}} \cdot \left(\boldsymbol{\eta_3} - \boldsymbol{\eta_3^0}\right) + \boldsymbol{w_{14}} \cdot \left(\boldsymbol{\eta_4} - \boldsymbol{\eta_4^0}\right)\right]$$

$$\sigma : \boldsymbol{z_2^a} = -{\boldsymbol{w_{22}}}^{-1} \cdot \left[\boldsymbol{w_{12}} \cdot \left(\boldsymbol{\eta_1} - \boldsymbol{\eta_1^0}\right) + \boldsymbol{w_{23}} \cdot \left(\boldsymbol{\eta_3} - \boldsymbol{\eta_3^0}\right) + \boldsymbol{w_{24}} \cdot \left(\boldsymbol{\eta_4} - \boldsymbol{\eta_4^0}\right)\right]$$

$$\nu : \boldsymbol{z_3^a} = -\boldsymbol{w_{33}}^{-1} \cdot \left[\boldsymbol{w_{13}} \cdot \left(\boldsymbol{\eta_1} - \boldsymbol{\eta_1^0}\right) + \boldsymbol{w_{23}} \cdot \left(\boldsymbol{\eta_2} - \boldsymbol{\eta_2^0}\right) + \boldsymbol{w_{34}} \cdot \left(\boldsymbol{\eta_4} - \boldsymbol{\eta_4^0}\right)\right]$$

$$\tau : \boldsymbol{z_4^a} = -\boldsymbol{w_{44}}^{-1} \cdot \left[\boldsymbol{w_{14}} \cdot \left(\boldsymbol{\eta_1} - \boldsymbol{\eta_1^0}\right) + \boldsymbol{w_{24}} \cdot \left(\boldsymbol{\eta_2} - \boldsymbol{\eta_2^0}\right) + \boldsymbol{w_{34}} \cdot \left(\boldsymbol{\eta_3} - \boldsymbol{\eta_3^0}\right)\right]$$

Unha vez definidas estas novas variables resposta axústase cada un dos parámetros da distribución utilizando o algoritmo backfitting. A iteración interior repítese unha e outra vez ata que a deviance global non cambia. Unha vez chegado a este punto, o algoritmo volve a iteración externa e recalcúlanse os valores $\boldsymbol{z_k}, \boldsymbol{w_{ks}}$ e $\boldsymbol{\eta_k^0}$ e comeza unha nova iteración interior. O proceso continúa ata que a deviance global da iteración exterior non cambia. Daquela o algoritmo para.

Vexamos os pasos deste algoritmo máis esquematicamente (os pasos en cor vermella pertencen á iteración exterior, os azuis á iteración interior e os laranxa ó algoritmo backfitting):

- PASO 0: Inicializamos.
- PASO 1: Definimos $\boldsymbol{z_k}$, $\boldsymbol{w_{ks}}$ e $\boldsymbol{\eta_k^0}$ para $k, s = 1, 2, 3, 4$.
- PASO 2: Definimos $\boldsymbol{z_k'}$ para $k = 1, 2, 3, 4$.
- PASO 3: Axustamos os modelos para μ, σ, ν e τ.
- PASO 4: Comprobamos se a deviance global converxeu. Se a resposta é sí imos o PASO 4.1, se a resposta é non imos o PASO 4.2.
- PASO 4.1: Volvemos ó PASO 2.
- PASO 4.2: Comprobamos se a deviance global converxeu. Se a resposta é sí PARAMOS, se a resposta é non volvemos ó PASO 1.

- Estimación de $\boldsymbol{\lambda}$

  Os algoritmos RS e CG estiman os valores de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ unha vez fixado o valor de $\boldsymbol{\lambda}$. Pero previamente deberíamos estimar dito valor. Esta estimación pode ser local, cando o método de estimación é aplicado dentro do algoritmo backfitting dos algoritmos RS ou CG, ou global, cando o método de estimación é aplicado fóra dos algoritmos.

  Para esta estimación pódense utilizar diferentes criterios.

  - Validación cruzada xeneralizada (GCV)
  - Criterio de información de Akaike xeneralizado (GAIC)
  - Métodos baseados en máxima verosimilitude (ML/REML)

  Os métodos locais soen ser máis rápidos e, a miúdo, producen solucións similares aos métodos globais. Ademais, estes métodos asumen que os residuos parciais compórtanse localmente como unha variable normal. Non obstante, os métodos globais son máis fiables.

  - Máxima verosimilitude local

    Na parte do algoritmo backfitting onde se axusta $\boldsymbol{\gamma}$ asúmese un modelo de efectos aleatorios coa intención de estimar o parámetro de suavizado $\boldsymbol{\lambda}$.

    $$\boldsymbol{\epsilon} = \boldsymbol{Z\gamma} + \boldsymbol{e}$$
    $$\boldsymbol{e} \sim N(0, \sigma_e^2 W)$$
    $$\boldsymbol{\gamma} \sim N(0, \sigma_b^2 \boldsymbol{G}^{-1})$$

    sendo $\boldsymbol{\epsilon}$ os residuos parciais dentro do algoritmo backfitting, $\boldsymbol{Z}$ a base para a suavización da variable explicativa correspondente, $\boldsymbol{W}$ a matriz diagonal que contén os valores dos pesos

iterativos e $\boldsymbol{G}$ a matriz de precisión que depende do método de suavizado que se estea a utilizar.

O parámetro de suavizado λ para a suavización da variable explicativa x é un ratio entre dúas varianzas, é dicir,

$$\lambda = \frac{\sigma_e^2}{\sigma_b^2}$$

Os parámetros σ_e^2, σ_b^2 e γ pódense obter utilizando o seguinte algoritmo:

1. Dado o parámetro λ estímase $\boldsymbol{\gamma}$ utilizando un procedemento de mínimos cadrados penalizados:
$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \lambda\boldsymbol{G})^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\epsilon}$$
2. Dada a última estimación de $\hat{\boldsymbol{\gamma}}$ calcúlase $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{S}\boldsymbol{\epsilon}$ onde $\boldsymbol{S} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}+\lambda\boldsymbol{G})^{-1}\boldsymbol{Z}^T\boldsymbol{W}$ e obtense:
$$\sigma_e^2 = \frac{(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}})^T(\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}})}{(n - tr(\boldsymbol{S}))}$$
$$\sigma_b^2 = \frac{\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}}}{tr(\boldsymbol{S})}$$
$$\hat{\lambda} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_b^2}$$
3. Se o valor de λ non cambia, o algoritmo para. Noutro caso caso continúa iterando, volvendo ao paso 1.

- Criterio de información de Akaike xeneralizado local

  Este método tenta minimizar a seguinte expresión:
$$G_{AIC} = ||\sqrt{\boldsymbol{w}} \cdot (\boldsymbol{\epsilon} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})||^2 + k \times tr(\boldsymbol{S})$$
  con respecto a λ para un k de penalización dado onde con $k = 2$ se obtén o AIC local e con $k = log(n)$ se obtén o BIC/SBC local.
- Validación cruzada xeneralizada local

  Este método tenta minimizar a seguinte expresión:
$$V_g = \frac{n||\sqrt{\boldsymbol{w}} \cdot (\boldsymbol{\epsilon} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})||^2}{(n - tr(\boldsymbol{S}))^2}$$
  con respecto a λ.

Nótese que se facemos uso destes algoritmos locais, os algoritmos RS e CG non teñen porque chegar necesariamente ao valor óptimo aínda que na práctica, xeralmente, obtéñense resultados razoables.

2.2.2. Termos Aditivos

Nun GAMLSS podemos incluír termos lineais e interaccións entre as variables independentes na súa parte paramétrica. Tamén poden ser incluídas nesta parte aquelas relacións non lineais que poden ser modeladas utilizando bases lineais:

- Polinomios

  Á hora de traballar con polinomios é obvio que o máis importante é a escolla do grao deste.

  Sabemos que a medida que aumentamos o grao a estimación mellora, pero a costa da aparición de formas irregulares ou picos. Polo tanto, atopar o grao adecuado resulta difícil e, as veces, imposible.

- Polinomios fraccionados

  Denomínase polinomios fraccionados a aqueles polinomios onde para a expresión βx^p o valor de p non é necesariamente un número enteiro positivo; unha pequena cantidade de polinomios fraccionados é suficiente para crear unha base bastante flexible para conseguir axustar unha curva paramétrica para os datos dos que se dispón. Os polinomios franccionados foron introducidos por Royston e Altman (1994).

- Polinomios a cachos

  Un xeito intuitivo de aumentar a flexibilidade da utilización de polinomios é dividir o intervalo total en intervalos máis pequenos e utilizar un polinomio, posiblemente de distinto grao, en cada un dos subintervalos.

  Non obstante, un dos maiores inconvenientes dos polinomios a cachos é que a función estimada non é necesariamente continua no punto onde pasamos dun subintervalo a outro (nodo).

  Cando tratamos de impoñer continuidade nos nodos obtense o que se coñece como splines.

  En xeral, un polinomio a cachos pódese definir da seguinte forma:

$$h(x) = \sum_{j=0}^{D} \beta_{0j} x^j + \sum_{k=1}^{K} \sum_{j=0}^{D} \beta_{kj} (x - b_k)^j H(x > b_k)$$

  onde D é o grao do polinomio en x e K é o número de nodos de b.

  A presencia ou ausencia do termo $\beta_{kj}(x - b_k)^j$ está relacionada coa continuidade ou non da derivada j-ésima do nodo b_k da función, o que implicará seguramente a necesidade de continuidade nas derivadas de menor orde. Isto conseguiríase eliminando todos os termos da forma $\beta_{km}(x - b_k)^m$ para $m = 0, 1, \ldots, k$.

  O nome spline sóese utilizar para polinomios a cachos onde todas as derivadas de orde menor que D son continuas en b_k. Por exemplo,

$$h(x) = \sum_{j=0}^{D} \beta_{0j} x^j + \sum_{k=1}^{K} \beta_k (x - b_k)^D H(x > b_k)$$

  é unha función spline de grao D. Para $D = 3$ teríamos un spline cúbico.

- Bases B-splines

  Un B-spline tamén está formado por polinomios a cachos que cumpren as restricións de continuidade nos nodos.

  Xeralmente, un spline de grao D cumpre as seguinte características:

  - Está formado por $D + 1$ cachos de polinomios de grao D que son continuos nos nodos
  - As derivadas destes polinomios son continuas nos nodos ata a orde $D - 1$
  - Toma valores positivos no dominio expandido por $D + 2$ nodos e 0 no resto
  - $D + 1$ B-splines son non nulos para cada valor de x

Á hora de traballar coa parte suavizada do modelo podemos clasificar os suavizadores principalmente en dous grupos. Os suavizadores penalizados, que usan unha penalización cadrática para controlar a suavización realizada sobre as covariables, e o resto de suavizadores que utilizan outras ideas, Bishop et al. (1995).

- P-splines:

  Dentro dos GAMLSS os suavizadores máis importantes son os penalizados debido a súa flexibilidade e a variedade de casos nos que se poden utilizar. Todos os suavizadores penalizados poden obterse como solución do seguinte problema de mínimos cadrados:

  Sexa $\boldsymbol{Z}$ unha matriz base $n \times p$, $\boldsymbol{\gamma}$ un vector de parámetros $p \times 1$, $\boldsymbol{W}$ unha matriz diagonal de pesos $n \times n$, $\boldsymbol{G}$ unha matriz de penalización $p \times p$, λ o parámetro de penalización e $\boldsymbol{y}$ a variable de interese. Neste contexto, os suavizadores penalizados son os que se obteñen como resultado de minimizar a seguinte expresión con respecto a $\boldsymbol{\gamma}$:

  $$\boldsymbol{Q} = (\boldsymbol{y} - \boldsymbol{Z}\lambda)^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{Z}) + \lambda \boldsymbol{\gamma}^T \boldsymbol{G} \boldsymbol{\gamma}$$

  A solución do anterior problema vén dada pola expresión:

  $$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \lambda \boldsymbol{G})^{-1} \boldsymbol{Z}^{\boldsymbol{T}} \boldsymbol{W} \boldsymbol{y}$$

Capítulo 3

GAMLSS en R

3.1. Introdución

Tras describir o modelo GAMLSS dun xeito teórico no capítulo anterior, neste capítulo farase unha pequena recompilación de como se pode traballar en R con estes modelos. Nas seguintes seccións describiranse os paquetes e as funcións de R utilizadas ó longo dos exemplos do seguinte capítulo e, posiblemente, algunhas máis que se consideraron de interese. Este capítulo é tamén importante porque se explicará como interpretar as saídas obtidas, para así poder comprender as conclusión que se extraerán no seguinte capítulo.

Unha descrición máis ampla destas e doutras funcións relacionadas cos GAMLSS poden atoparse no xa mencionado Stasinopoulos et al. (2015c).

3.2. GAMLSS: os seus paquetes

Os paquetes principais relacionados cos modelos GAMLSS son os paquetes:

- gamlss que contén, entre outras, a función principal gamlss(), Stasinopoulos et al. (2016c).
- gamlss.dist que contén as principais distribucións utilizadas polos modelos GAMLSS, Stasinopoulos et al. (2016a).
- gamlss.data que contén as principais bases de datos para as que se utilizou GAMLSS, Stasinopoulos e Rigby (2016a).

Ademais destes 3, existen outros paquetes necesarios para levar a cabo determinadas accións. Cando do que se trata é de traballar coas distribucións podemos precisar os paquetes:

- gamlss.tr que permite crear distribucións truncadas (á esquerda, á dereita ou por ambos lados) a partir das distribución xa dispoñibles, Stasinopoulos e Rigby (2016b).
- gamlss.cens que permite crear distribucións censuradas (á esquerda, á dereita ou por ambos lados) a partir das distribución xa dispoñibles, Stasinopoulos et al. (2016b).
- gamlss.mx que permite crear mesturas de distribucións, Stasinopoulos e Rigby (2016c).
- gamlss.demo que permite ver exemplos de distribucións para uns determinados parámetros dados, Stasinopoulos et al. (2015a).

Outros paquetes dispoñibles son:

- gamlss.nl que permite axustar modelos paramétricos non lineais, Stasinopoulos et al. (2015b).
- gamlss.spatial que permite axustar modelos espaciais, De Bastiani et al. (2016).

3.3. GAMLSS: o axuste

Cando nos dispoñemos a traballar con modelos GAMLSS son moitas as funcións que están dispoñibles en R. Nesta sección citamos, explicamos e interpretamos as saídas das que consideramos as principais funcións relacionadas co axuste dos modelos GAMLSS.

Seguramente a función gamlss() sexa a máis importante pero, por orde de utilización, posiblemente deberíamos mencionar antes as funcións histDist() e fitDist(). Estas funcións están dispoñibles en R cos seguintes argumentos principais:

- histDist(y, family = NO, freq = NULL, density = FALSE, nbins = 10, xlim = NULL, ylim = NULL,

  main = NULL, xlab = NULL, ylab = NULL, data = NULL, ...)

- fitDist(y, k = 2, type = c("realAll", "realline", "realplus", "real0to1", "counts", "binom"),

  try.gamlss = FALSE, extra = NULL, data = NULL, ...)

  Neste caso seguramente o argumento máis relevante sexa o argumento type=" " que indica o tipo de distribucións que lle queremos axustar ós nosos datos:

  - realline, esta opción recolle aquelas distribucións continuas que están definidas en toda a recta real (GU, RG, ST1,...).
  - realplus, esta opción recolle aquelas distribucións continuas que están definidas soamente na parte positiva da recta real (EXP, GG, BCTo,...).
  - realAll, esta opción recolle todas as distribucións das opcións realline e realplus.
  - real0to1, esta opción recolle aquelas distribucións continuas que están definidas entre 0 e 1 (BE, BEZI, BEINF,...).
  - counts, esta opción recolle aquelas distribucións que se poden considerar de conteo (PIG, ZIP, ZAP,...)
  - binom, esta opción recolle aquelas distribucións que se poden considerar de tipo binomial (BI, BB,ZIBI,...).

Estas dúas funcións xogan un papel importante á hora de escoller a distribución máis axeitada para un conxunto de datos. A primeira delas apórtanos saídas como a que se pode ver na Figura 3.1, neste caso a liña de comando utilizada foi:

```
> mNO<-histDist(vsg,density=T,xlab="VSG,mm",ylab="",main="NO",nbins=30)
```

Non se lle especificou ningunha familia co argumento family =" ", polo tanto a distribución utilizada foi a distribución normal que é a que trae por defecto a función histDist().

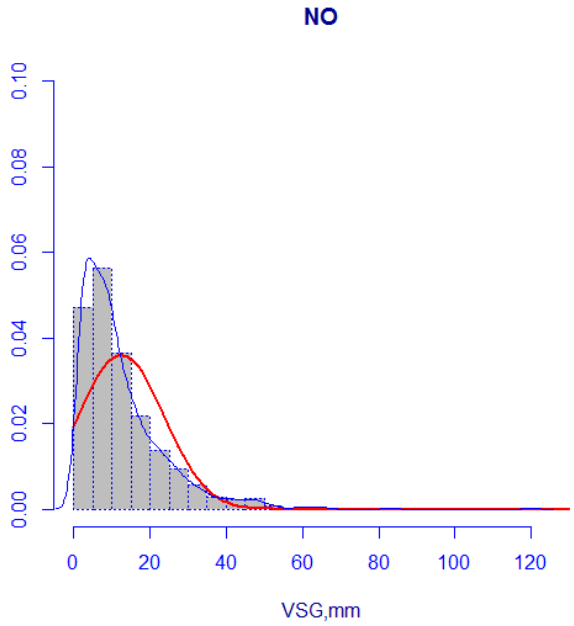

Figura 3.1: Saída obtida a partir da función histDist().

Seguramente este sexa o primeiro paso que deberíamos dar cando un conxunto de datos chega as nosas mans. O primeiro que xeralmente nos preguntamos é que distribución seguirán e, posteriormente, se esa distribución pode ser a distribución normal.

Xeralmente a distribución normal non será a máis axeitada para os nosos datos, pero ver a densidade dos datos, xunto coa densidade da distribución normal, como ocorre na Figura 3.1, pode axudarnos a ver características dos datos como pode ser unha posible asimetría.

Dependendo do coñecemento que cada usuario teña das distintas distribucións, considerarán a utilización ou non da función fitDist(). Dita función axusta as distribucións dispoñibles en R para o argumento seleccionado en type =" " e ordénaas utilizando o criterio AIC, de xeito que nos indicará como mellor axuste aquela distribución con menor AIC. Esta función pode ser realmente útil cando o usuario non é capaz por si mesmo de ver ningunha opción clara para o axuste dos datos. Non obstante, hai que ter coidado coas saídas obtidas a partir desta función xa que, como moitas outras relacionadas con modelos GAMLSS, sofre de fortes problemas de converxencia.

Se se decide utilizar a función fitDist() unha posible saída sería a seguinte:

```
> fitvsg <- fitDist(vsg,type="realplus")
> fitvsg$fit
```

```
  exGAUS      GIG    BCPEo       GG    BCCGo     BCTo    LOGNO       IG
10096.64 10139.74 10142.20 10172.73 10173.65 10175.65 10180.91 10184.01
      GA     WEI3   IGAMMA      EXP  PARETO2
10229.94 10272.20 10365.33 10370.68 10372.68
```

Esta indicaríanos que as distribucións exGAUS, GIG ou BCPEo son as máis axeitadas para o axuste dos datos tendo en conta o criterio AIC.

Con esta información poderíamos utilizar de novo da función histDist(), esta vez utilizando o argumento family =“ ” para introducir as distribucións obtidas coa función fitDist(), para ver os resultados dun xeito máis visual, Figura 3.2.

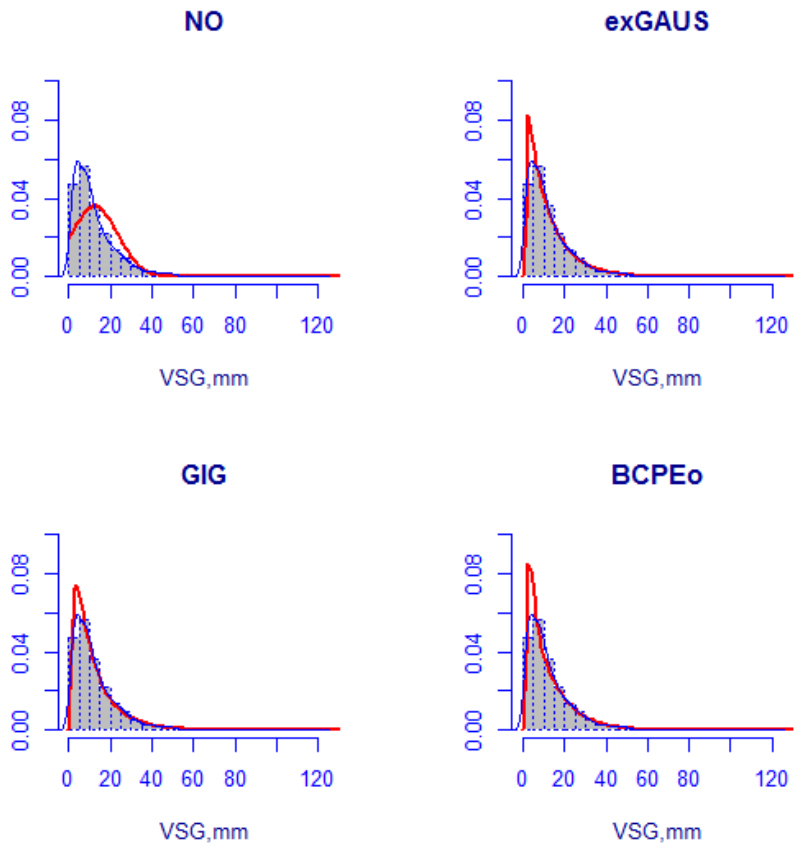

Figura 3.2: Saída obtida a partir da función histDist().

Unha vez escollida a distribución para realizar o axuste sí que debemos botar man da función gamlss() para a realización do mesmo:

- gamlss(formula = formula(data), sigma.formula = 1, nu.formula = 1, tau.formula = 1, family = NO(), data = sys.parent(), weights = NULL, contrasts = NULL, method = RS(), start.from = NULL, mu.start = NULL, sigma.start = NULL, nu.start = NULL, tau.start = NULL, mu.fix = FALSE, sigma.fix = FALSE, nu.fix = FALSE, tau.fix = FALSE, control = gamlss.control(...), i.control = glim.control(...), ...)

O primeiro que deberíamos facer é axustar o modelo sen variables independentes para coñecer a estimación dos parámetros da distribución escollida. Unha vez feito isto podemos proceder a incluír as variables explicativas que consideremos pertinentes.

Esta función leva asociadas a maioría das funcións que xa existían para os modelos GLM ou GAM como poden ser a summary() ou predit(), obviamente as saídas obtidas cando o que se lles proporciona é un modelo GAMLSS son diferentes a cando o modelo é un GLM ou GAM. Ó longo dos exemplos

do seguinte capítulo iremos vendo esas diferencias e interpretando correctamente as saídas obtidas. Así como coñecendo funcións que foron lixeiramente modificadas na forma, pero non no fondo, como pode ser a función term.plot() análoga á termplot() dos modelos GAM.

Á hora de axustar o modelo GAMLSS tamén debemos ter en conta a posible interacción entre as variables, ou a súa suavización. Neste traballo, xeralmente, non consideraremos a interacción entre as distintas variables explicativas e utilizaremos a función pb() para realizar a suavización daquelas variables continuas que consideremos necesario suavizar pero, para afondar no tema, podemos recorrer de novo a Stasinopoulos et al. (2015c).

Tras construír varios modelos, ou simplemente aquel que nos pareza o máis completo, debemos comezar cun traballo de selección do modelo. Algunhas das funcións das que podemos botar man nesta parte do proceso veñen explicadas na seguinte sección.

3.4. GAMLSS: a selección do modelo

Unha vez metidos na parte de axuste do modelo debemos facer unha boa selección de certos elementos: a distribución, as covariables, a suavización ...

Nesta sección describiremos algunhas das funcións que nos poden resultar de axuda para unha boa selección do modelo.

En primeiro lugar debemos determinar cal vai ser o noso plan de actuación dependendo, por exemplo, do tamaño da mostra que temos.

Se dispoñemos dunha mostra suficientemente grande podemos considerar dividir dita mostra en dous e utilizar a primeira parte para axustar o modelo e a segunda parte para validar dito modelo. Se pola contra o noso tamaño de mostra é máis ven pequeno deberíamos traballar con toda a mostra para levar a cabo o axuste.

Vexamos que funcións podemos utilizar en cada caso:

- Non dispoñemos de base de datos de validación:
  - Traballamos con todos os datos á vez:
    - drop1()
      Permítenos saber cal das variables do modelo se pode eliminar atendendo a significación que se pode ver na saída da función. Un exemplo de saída da función drop1() é a seguinte:

```
> drop1(mhes,what="sigma")

Single term deletions for
sigma

Model:
~pb(edad) + sexo
             Df     AIC    LRT Pr(Chi)
<none>           -114.52
pb(edad) 0.69374 -116.80 -0.900 1.00000
sexo     3.95674 -110.27 12.164 0.01564 *
```

      A saída anterior indícanos que á hora de axustar o parámetro de escala poderíamos prescindir da variable edad. Non obstante, non poderíamos prescindir da variable sexo.

      NOTA: Se soamente existen variables continuas sen suavizar no modelo, esta saída é a mesma ca que se pode observar coa función summary(), pero no momento en que se inclúan variables suavizadas ou variables factor a fiabilidade da función summary() pérdese mentres que a da drop1() non.

- add1()
  Permítenos saber que interaccións ou suavizacións engadir atendendo á significación que se pode ver na saída da función.
- stepGAIC()
  Permítenos, utilizando o criterio AIC, escoller aquel modelo máis axeitado tendo en conta dito criterio. Esta función é análoga á función stepAIC() da libraría MASS, só que admite novos argumentos como o referente ó parámetro para o que se leva a cabo a selección.
  Se o que pretendemos é realizar unha selección tendo en conta todos os parámetros da distribución á vez, as funcións que debemos utilizar son as dúas que se expoñen a continuación:
- stepGAICAll.A()
  A estratexia utilizada por esta función para unha distribución dada é a seguinte:
  1. Utilizar o criterio GAIC cara adiante para seleccionar un modelo apropiado para μ, fixando σ, ν e τ.
  2. Dado o modelo para μ obtido en 1 e para ν e τ fixas, úsase un proceso de selección cara adiante para escoller un modelo apropiado para σ.
  3. Dados os modelos para μ e σ obtidos en 1 e 2 respectivamente e con τ fixa, utilizamos un proceso de selección cara adiante para escoller un modelo apropiado para ν.
  4. Dados os modelos para μ, σ e ν obtidos en 1, 2 e 3 respectivamente, utilizamos un proceso de selección cara adiante para escoller un modelo apropiado para τ.
  5. Dados os modelos para μ, σ e τ obtidos en 1, 2 e 4 respectivamente, utilizamos un proceso de selección cara atrás para escoller un modelo apropiado para ν.
  6. Dados os modelos para μ, ν e τ obtidos en 1, 5 e 4 respectivamente, utilizamos un proceso de selección cara atrás para escoller un modelo apropiado para σ.
  7. Dados os modelos para σ, ν e τ obtidos en 6, 5 e 4 respectivamente, usamos un proceso de selección cara atrás para escoller un modelo apropiado para μ e finalizamos o proceso.

  A continuación móstrase parte da saída que se obtén utilizando a seguinte liña de comando (omítense partes xa que a saída é moi longa):

```
> mauccovA<-gamlss(auc~pb(glu)+age+sex+bmi+ipq+tab,
sigma.formula = ~pb(age)+sex,data=dat0,family=BCTo)
> stepGAICAll.A(mauccovA)

-----------------------------------------------------
Distribution parameter:  mu
Start:  AIC= 4070.7
 auc ~ pb(glu) + age + sex + bmi + ipq + tab

-----------------------------------------------------
Distribution parameter:  sigma
Start:  AIC= 4070.7
 ~pb(age) + sex

-----------------------------------------------------
Distribution parameter:  nu
Start:  AIC= 4070.7
 ~1

-----------------------------------------------------
Distribution parameter:  tau
```

```
Start:  AIC= 4070.7
 ~1

-----------------------------------------------------

...

-----------------------------------------------------
Distribution parameter:  sigma

...

Step:  AIC= 4068.71
 ~pb(age)

              Df    AIC
<none>           4068.7
- pb(age) 1.6104 4084.7
-----------------------------------------------------
Distribution parameter:  mu

...

Step:  AIC= 4066.27
 auc ~ pb(glu) + age + ipq + tab

               Df    AIC
<none>            4066.3
- ipq     1.94480 4067.5
- tab     2.01248 4068.7
- age     0.84154 4082.1
- pb(glu) 5.32450 4288.4
-----------------------------------------------------

Family:  c("BCTo", "Box-Cox-t-orig.")
Fitting method: RS()

Call:
gamlss(formula = auc ~ pb(glu) + age + ipq + tab, sigma.formula = ~pb(age),
    nu.formula = ~1, tau.formula = ~1, family = BCTo, data = dat0,
    trace = FALSE)

...
```

- stepGAICAll.B()
  Neste caso, a estratexia B, a diferenza da estratexia A, obriga a que sexan escollidos os mesmos termos para todos os parámetros da distribución. Un exemplo de saída sería a seguinte:

```
> mauccovB<-gamlss(auc~pb(glu)+age+sex+bmi+cal+fibra,sigma.formula =
 ~pb(glu)+age+sex+bmi+cal+fibra,data=dat0,family=BCTo)
> stepGAICAll.B(mauccovB)
```

```
Start:  AIC= 4067.45
 auc ~ pb(glu) + age + sex + bmi + cal + fibra

              Df    AIC
- sex     1.7794 4063.7
- bmi     2.3956 4066.2
- cal     1.8718 4066.5
<none>           4067.5
- fibra   1.7734 4070.9
- age     2.2990 4089.5
- pb(glu) 8.1214 4277.9

...

Step:  AIC= 4060.69
 auc ~ pb(glu) + age + fibra

              Df    AIC
<none>           4060.7
- fibra   1.5583 4062.7
- age     2.0219 4084.7
- pb(glu) 8.0069 4282.4

Family:  c("BCTo", "Box-Cox-t-orig.")
Fitting method: RS()

Call:
gamlss(formula = auc ~ pb(glu) + age + fibra, sigma.formula = ~pb(glu) +
    age + fibra, nu.formula = ~1, tau.formula = ~1, family = BCTo,
    data = dat0, trace = FALSE)

...
```

- Utilizamos validación cruzada:
  - gamlssCV()
    Utilízase para axustar o modelo, comparte varios argumentos coa función gamlss():
    - gamlssCV(formula = NULL, sigma.formula = 1, nu.formula = 1, tau.formula = 1, data = NULL, family = NO, control = gamlss.control(trace = FALSE), K.fold = 10, set.seed = 123, rand = NULL, parallel = c("no", "multicore", "snow"), ncpus = 1L, cl = NULL, ...)

    Cabe destacar a función do argumento K.fold =" " que nos permite indicarlle o número de partes nas que queremos romper a mostra. Tamén podemos definir unha variable categórica para indicarlle cada un dos grupos e incluíla no argumento rand =" ".
  - CV() utilízase para comparar os modelos axustados coa función gamlssCV().

- Dispoñemos de base de datos de validación:
  - gamlssVGD()
    Utilízase para axustar o modelo na base de datos de ensaio e validalo na base de datos de validación; igual que a función gamlssCV(), comparte varios argumentos coa función gamlss():

- gamlssVGD(formula = NULL, sigma.formula = 1, nu.formula = 1, tau.formula = 1, data = NULL, family = NO, control = gamlss.control(trace = FALSE), rand = NULL, newdata = NULL, ...)

Cabe destacar a función do argumento rand =“ ” e a do argumento newdata =“ ”. Se temos dúas bases de datos diferenciadas como base de ensaio e base de validación debemos darlle o nome da base de ensaio ó argumento data=“ ” e o nome da base de validación ó argumento newdata =“ ”. Se pola contra só dispoñemos dunha base de datos podemos utilizar o argumento rand =“ ” para especificar como queremos dividir a nosa base en dous.

- VGD() utilízase para comparar os modelos axustados coa función gamlssVGD()
- getTGD() utilízase case para o mesmo que se utiliza a función gamlssVGD(). Tamén se utiliza para axustar un modelo na base de ensaio pero, neste caso, sen ter en conta a base de validación.
- TGD() utilízase para comparar o ben que se adaptan os modelos axustados coa función getTGD() á base de validación.
- add1TGD() e drop1TGD() teñen unha función análoga ás funcións add1() e drop1().
- stepTGD() ten unha función análoga á función stepGAIC(), só que neste caso o criterio de selección é a deviance global da base de validación e non o GAIC. Ata o momento esta función só é capaz de traballar cun parámetro á vez.

3.5. GAMLSS: a diagnose do modelo

Unha vez construído o modelo que consideramos máis adecuado, tras levar a cabo a selección do modelo, tanto coa axuda das funcións da sección anterior como tendo en conta consideracións feitas pola natureza do problema, como pode ser a importancia de conservar unha variable aínda que esta non sexa significativa; debemos considerar levar a cabo unha última parte de diagnose do modelo.

Ó igual que ocorre cos modelos xa coñecidos, á hora de validar os modelos GAMLSS, existen diversas funcións en R que traballan cos “residuos”do modelo. A diferenza cando traballamos con modelos GAMLSS está na definición do que denominamos aquí residuos, referímonos á definición dada no Capítulo 2.

As funcións das que dispón R, ou polo menos as máis importante, son as seguintes:

- plot.gamlss() (ou simplemente plot() se o primeiro argumento é un modelo GAMLSS)

  A función plot() ten como saída 4 gráficos que representan o seguinte:

  - residuos fronte ós valores axustados para a media
  - residuos fronte a algunha das variables explicativas do modelo (se esta non é especificada simplemente se toman no eixo x os valores $1 : N$ sendo N o número total de observacións)
  - estimación Kernel da densidade dos residuos
  - QQ-plot dos residuos

  Un exemplo de saída da función plot() é a que se pode ver na Figura 3.3.

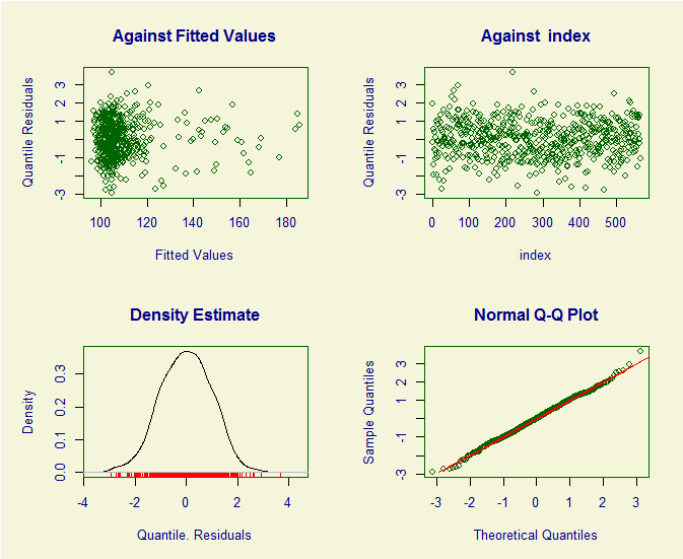

Figura 3.3: Saída obtida a partir da función plot().

Dado que o gráfico de arriba á esquerda representa os residuos fronte aos valores axustados para a media; os puntos que aparecen neste gráfico non debería ter ningún patrón visible, como ocorre neste caso.

O segundo gráfico pode ter unha interpretación diferente segundo o que se estea a representar nel, dado que podemos escoller a opción de que se pinten os residuos fronte a algunha das variables explicativas do modelo.

Os dous gráficos de abaixo móstrannos a posible normalidade dos residuos. O gráfico da esquerda debería semellarse á densidade dunha distribución normal, mentres que o da dereita se trata dun QQ-plot; polo tanto, todos os puntos deberían caer sobre a liña vermella (neste caso non todos o fan pero aínda así son poucos os que caen fóra).

Ademais, xunto co gráfico da Figura 3.3, obtense unha saída co seguinte formato:

```
> plot(mvsgTGD)


******************************************************************
              Summary of the Quantile Residuals
                           mean   =  -6.013267e-05
                       variance   =  1.000188
                coef. of skewness  =  0.009045191
                coef. of kurtosis  =  3.279199
Filliben correlation coefficient  =  0.9986419
******************************************************************
```

Os parámetros da saída anterior deben ser semellantes ós dunha normal para que poidamos considerar que estamos ante un modelo axeitado.

NOTA: Se estamos traballando con unha serie de tempo a función plot() ten a opción de darlle como argumento ts = T e deste xeito substitúense os dous gráficos de arriba polos gráficos das ACF e PACF, como se pode ver na Figura 3.4 tomada de Stasinopoulos et al. (2015c).

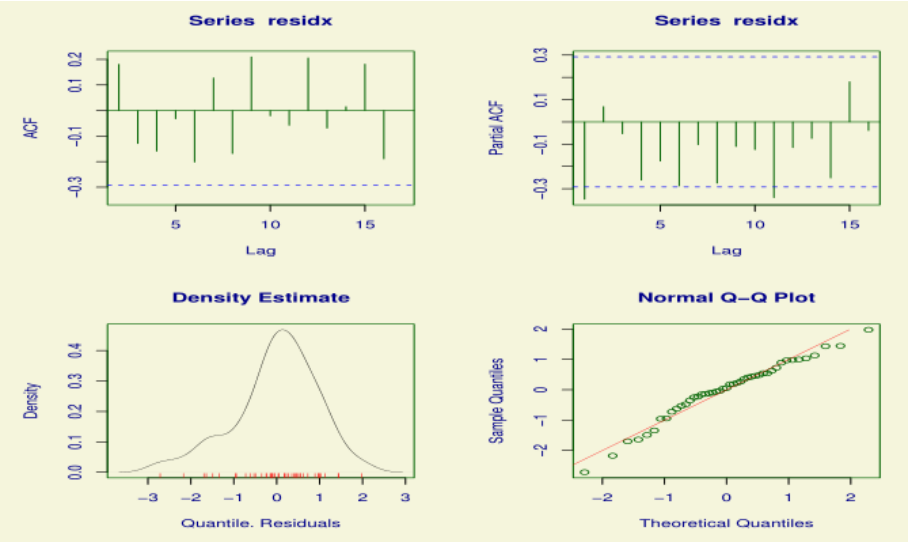

Figura 3.4: Saída obtida a partir da función plot() tomada de Stasinopoulos et al. (2015c).

Cando se traballa con distribucións discretas é mellor botar man da función rqres.plot(), que tamén permite obter worm plots e QQ-plots. Para ver un exemplo disto recorremos de novo a Stasinopoulos et al. (2015c).

- wp()

Os worm plot foron introducidos por van Buuren e Fredrils (2001) coa intención de identificar as rexións dunha variable explicativa onde o modelo non se axusta adecuadamente ós datos.

A función wp() pode ter como saída un único gráfico se deixamos o argumento xvar = NULL que trae a propia función por defecto, ou varios gráficos se lle damos a ese argumento unha das variables explicativas. Neste último caso obteremos unha gráfica por cada un dos intervalos que lle indiquemos no argumento n.iner =“ ” (4 é o que trae por defecto).

Un par de exemplos de saída para a función wp() cun único gráfico é a que se pode ver na Figura 3.5 onde:

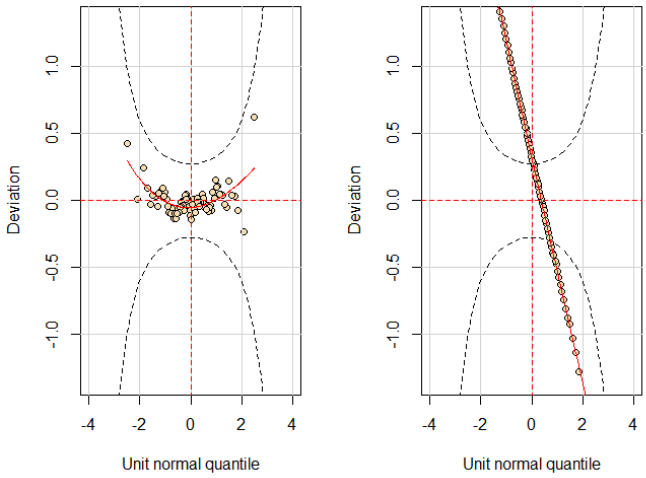

Figura 3.5: Saída obtida a partir da función wp().

- os puntos representan os residuos e a liña discontinua vermella (horizontal) os seus valores esperados, polo tanto fixándonos nisto podemos ver que tan lonxe están uns dos outros.
- as liñas discontinuas negras marcan o intervalo de confianza do 95 %, polo tanto para que poidamos considerar que o modelo é correcto, só un 5 % dos puntos poderían quedar fóra deste intervalo. Se isto non é así debemos considerar o noso modelo inadecuado para explicar a variable resposta.
- a curva continua vermella é un axuste cúbico dos puntos e pode indicarnos diferentes problemas no modelo como poden ser os que aparecen no Cadro 3.1. Un exemplo de cada un destes problemas pode verse na Figura 3.6 tomada de Stasinopoulos et al. (2015c).

Atendendo aos criterios anteriores, na Figura 3.5 deberíamos considerar un bo modelo o correspondente ao worm plot da esquerda, pero un modelo inadecuado ao modelo asociado ao worm plot da dereita.

Forma do wp	**Residuos**	**Variable resposta**
Datos por encima de 0	media alta	parámetro de localización baixo
Datos por debaixo de 0	media baixa	parámetro de localización alto
Pendente positiva	varianza alta	parámetro de escala baixo
Pendente negativa	varianza baixa	parámetro de escala alto
uy Forma de U	asimetría positiva	asimetría baixa
Forma de U inversa	asimetría negativa	asimetría alta
Forma de S xirada á esquerda	leptocúrtica	curtosis baixa
Forma de S xirada á esquerda e dada a volta	platicúrticos	curtosis alta

Cadro 3.1: Problemas detectados a través do worm plot.

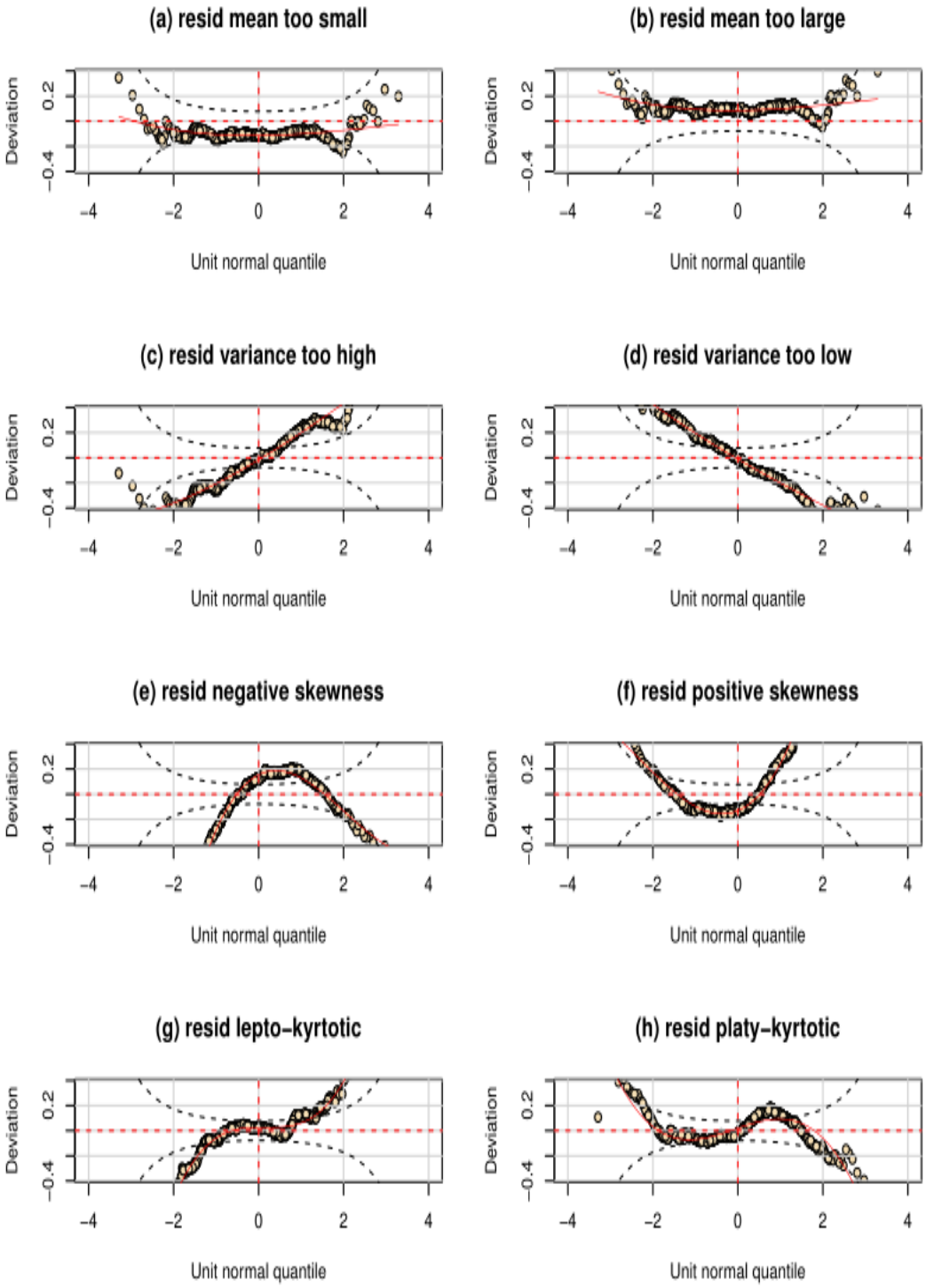

Figura 3.6: Imaxe tomada de Stasinopoulos et al. (2015c) onde se pode ver un exemplo dos problemas expostos no Cadro 3.1. As gráficas da primeira fila están relacionadas coa media dos residuos (demasiado baixa no da esquerda, demasiado alta no da dereita), as da segunda fila coa varianza dos residuos (demasiado baixa no da esquerda, demasiado alta no da dereita), as da terceira fila coa asimetría dos residuos (negativa a da esquerda, positiva a da dereita) e as da cuarta fila coa curtose dos residuos (leptocúrtica a da esquerda, platicúrtica a da dereita).

Un exemplo de saída para a función wp() con varios gráficos é a que se pode ver na Figura 3.7. Neste caso indicouse o argumento xvar = age. Pero dado que non se lle especificou o argumento n.iner =“ ”, a variable idade foi dividida en 4 grupos. Eses 4 grupos son os que están indicados na parte superior do gráfico.

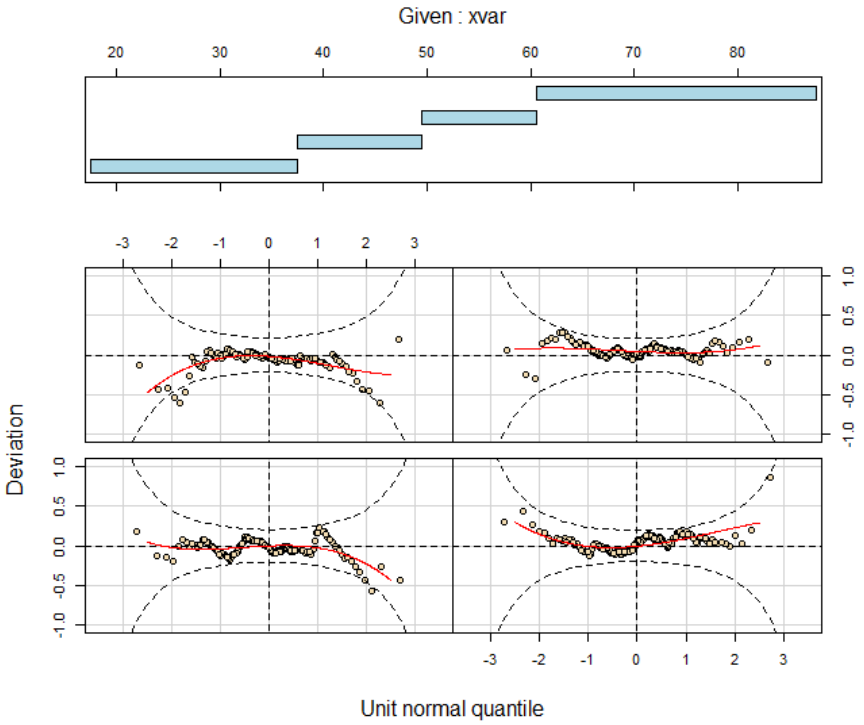

Figura 3.7: Saída obtida a partir da función wp().

- Q.stats()

  A función Q.stats() intenta detectar aqueles residuos nos que os seus parámetros son significativamente distintos ós dunha distribución normal, atendendo ós rangos dalgunha das variables explicativas.

  A función Q.stats() ten como saída unha parte numérica e outra gráfica, a numérica é a que se pode ver a continuación e a gráfica é a que se pode ver na Figura 3.8:

```
> mauc1ageNO<-gamlss(dat1$auc~dat1$age,family=NO)
> mauc1age<-gamlss(dat1$auc~dat1$age,family=LO)
> Q.stats(mauc1ageNO,xvar=age,n.inter=4)
```

```
                          Z1         Z2           Z3           Z4   AgostinoK2   N
17.5 to 37.5      -0.6109911 -0.5756907 -1.029780977  1.020469820 2.101808e+00 117
37.5 to 48.5       0.8249188 -0.4574230 -1.957661902  1.815633226 7.128964e+00  90
48.5 to 60.5      -0.2447343 -0.4461494 -0.421592889  2.090189003 4.546631e+00 110
60.5 to 87.5       0.1269793  1.4138633  3.086472047  2.422722802 1.539590e+01 113
TOTAL Q stats      1.1298198  2.7387142 14.596939246 14.576358511 2.917330e+01 430
df for Q stats     2.0000000  3.0000000  4.000000000  4.000000000 8.000000e+00   0
p-val for Q stats  0.5684114  0.4336883  0.005614524  0.005665563 2.956395e-04   0
```

```
> Q.stats(mauc1age,xvar=age,n.inter=4)

                            Z1         Z2         Z3         Z4 AgostinoK2   N
17.5 to 37.5      -0.473332066 -0.3985895 -0.5505386 -0.6656418  0.7461718 117
37.5 to 48.5       1.032433323 -0.4537090 -1.6036437  0.5930581  2.9233912  90
48.5 to 60.5      -0.189351839 -0.4112335  0.4123755  0.1567641  0.1946286 110
60.5 to 87.5       0.001682078  1.0433728  1.9489265  0.3531354  3.9230190 113
TOTAL Q stats      1.325818759  1.6224652  6.8431340  0.9440765  7.7872105 430
df for Q stats     2.000000000  3.0000000  4.0000000  4.0000000  8.0000000   0
p-val for Q stats  0.515349803  0.6543068  0.1444137  0.9181539  0.4545274   0
```

Na saída numérica debemos intentar detectar valores maiores de 1.96 (en valor absoluto) que serían os que non se poden aceptar para unha distribución normal. Na saída gráfica estes valores son marcados con un recadro dentro do círculo correspondente.

Para tratar de exemplificar como traballa a función Q.stats() axustáronse dous modelos, un deles prevendo que non fose axeitado (mauc1ageNO) e outro que sí o fose (mauc1age). De acordo con isto, na primeira saída numérica atopamos valores maiores que 1.96 mentres que na segunda non. O mesmo ocorre cos gráficos, o da esquerda correspondente ó modelo mauc1ageNO presenta recadros dentro dos círculos mentres que o da dereita, correspondente ó modelo mauc1age, non.

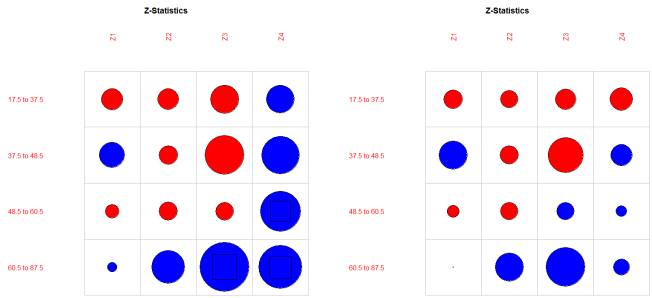

Figura 3.8: Saída obtida a partir da función Q.stats(). A gráfica da esquerda correspóndelle ó modelo mauc1ageNO e a da dereita ó modelo mauc1age. Os recadros dentro dos círculos indican que ese valor non se axusta a unha normal nese parámetro e para esa franxa da variable explicativa escollida. As cores azul e vermella indican valores positivos e negativos respectivamente.

3.6. GAMLSS: os centiles

Nesta sección centrarémonos nas funcións relacionadas cos valores que toma a nosa variable resposta nos distintos percentís atendendo a algunha das variables explicativas.

Xeralmente este tipo de funcións úsanse cando a distribución da variable resposta é continua e utilízase unha única covariable. Cando se pretende incluír tamén na análise unha variable categórica, como pode ser o sexo, normalmente sepáranse os datos atendendo a dita variable e faise a análise en cada grupo por separado.

Non obstante, existen publicacións onde se utilizou máis dunha variable explicativa, idade e altura, Cole et al. (2009) ou Quanjer et al. (2012).

A función principal deste grupo de funcións é a función centiles(). Esta función ten como saída un gráfico como o que aparece na Figura 3.9, onde os puntos representan un diagrama de dispersión

das variables resposta e explicativa e as liñas móstrannos como se van movendo os valores da variable resposta (nos percentís indicados) en función dos valores da variable explicativa. Estas liñas tamén nos dan unha idea do tipo de relación (lineal ou non) que hai entre as dúas variables, neste sentido as liñas adoptarán unha forma similar a do term.plot() obtido se axustamos un modelo só con estas dúas variables. Ademais da parte gráfica, a saída de R proporciónanos como dato a porcentaxe de datos que quedan por debaixo de cada un dos centiles representados.

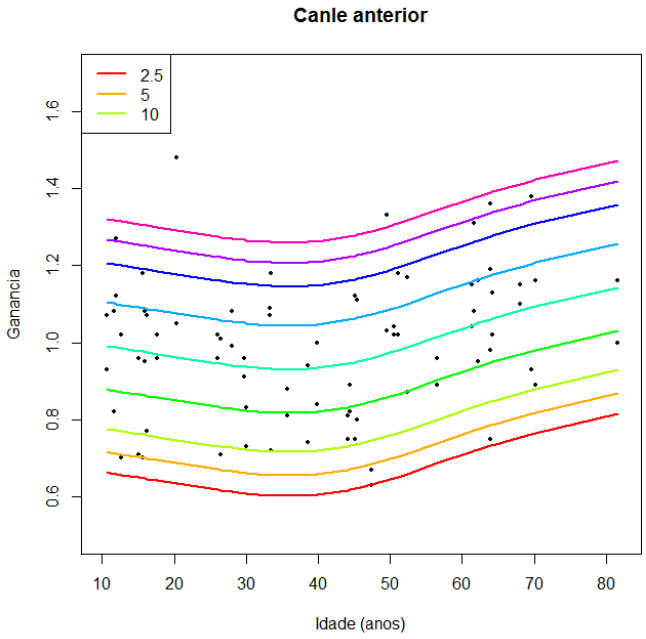

Figura 3.9: Saída obtida a partir da función centiles().

Un resultado similar obtense utilizando a función centiles.split(), só que neste caso os valores da variable explicativa son divididos en intervalos e obteremos un gráfico para cada un destes intervalos como se pode observar na Figura 3.10.

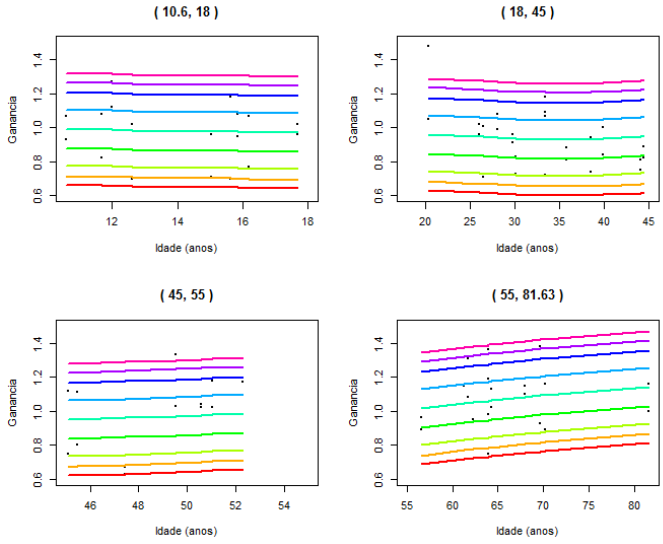

Figura 3.10: Saída obtida a partir da función centiles.split().

Se o que desexamos é obter nunha mesma gráfica a saída de dous modelos distintos, podemos utilizar a función centiles.com(). Na Figura 3.11 podemos ver unha saída obtida a partir desta función; neste caso os modelos comparados foron o modelo $y \sim NO\{\mu = edad, \log(\sigma) = 1\}$ (sen suavización) e o modelo $y \sim NO\{\mu = pb(edad), \log(\sigma) = 1\}$ (con suavización).

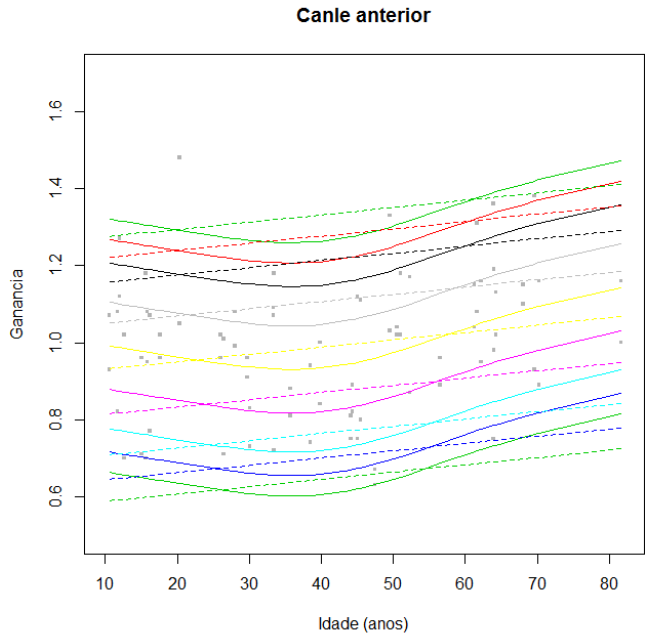

Figura 3.11: Saída obtida a partir da función centiles.com().

Por último, se o que queremos é predicir o valor da variable resposta (para distintos centiles) para novos valores da variable explicativa, debemos utilizar a función centiles.pred(). Se ademais dos valores desexamos tamén unha saída gráfica será suficiente con utilizar o argumento plot = T dentro da función centiles.pred(), obtendo así unha saída como a que aparece na Figura 3.12.

edad	C2.5	C5	C10
10	0.6645827	0.7174106	0.7783177
15	0.6495894	0.7024173	0.7633245
20	0.6346522	0.6874801	0.7483873
25	0.6203291	0.6731570	0.7340642
30	0.6083948	0.6612227	0.7221299
35	0.6023603	0.6551882	0.7160953
40	0.6056883	0.6585162	0.7194233
45	0.6203221	0.6731500	0.7340572
50	0.6453393	0.6981672	0.7590744
55	0.6763610	0.7291888	0.7900960
60	0.7084990	0.7613269	0.8222341
65	0.7383496	0.7911775	0.8520847
70	0.7645867	0.8174146	0.8783217
75	0.7874834	0.8403112	0.9012184
80	0.8083624	0.8611903	0.9220975

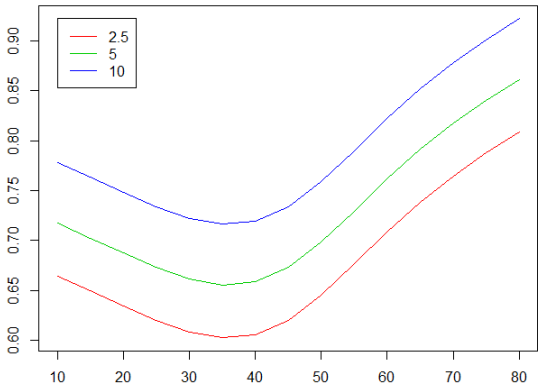

Figura 3.12: Á esquerda aparece a saída numérica obtida ó utilizar a función centiles.pred(), á dereita aparece a saída gráfica obtida a partir da mesma función.

Estas funcións son moi útiles cando se intentan dar valores de referencia como se verá no último capítulo deste traballo.

3.7. De LM a GAMLSS mediante un exemplo real

Retomamos o exemplo da Sección 1.3 do Capítulo 1 e incluímos na táboa da Figura 1.1 o modelo GAMLSS (con e sen efectos aleatorios).

```
> mgamlss<-gamlss(ganancia~pb(edad),data=dath,
          sigma.formula = ~ sexo,family="IG")

> mgamlssm<-gamlss(ganancia~re(random=~1|id)
                        +pb(edad),data=dath,
                    sigma.formula = ~ sexo,
                              family="IG")
```

Modelo	AIC
mlm	-91.75
mglm	-98.28
mgam	-105.75
mgamm	-106.53
mgamlss	-110.35
mgamlssm	-115.28

Figura 3.13: Á esquerda aparecen as liñas de comando utilizadas para axustar os modelos mgamlss e mgamlssm. Á dereita unha táboa que recolle os valores do AIC destes modelos.

Polo tanto, completamos a táboa da Figura 1.1 con estes 2 novos modelos obtendo así a táboa da Figura 3.13 onde se pode ver como o valor do AIC destes dous últimos modelos incluídos é menor que os que xa formaban parte da táboa da Figura 1.1. Entón sería axeitado utilizar o modelo GAMLSS (con efectos aleatorios) para este exemplo e, para moitos máis, como veremos no seguinte capítulo.

Nótese que a modificación feita nos modelos mgamlss e mgamlssm con respecto ós modelos mgam e mgamm foi modelar a varianza con respecto ó sexo. Non obstante, poderíase seguir mellorando o modelo entrando tamén na selección dunha nova familia.

Capítulo 4

Aplicación dos GAMLSS en exemplos reais

4.1. Introdución

Neste último capítulo levarase a cabo a parte máis práctica do traballo realizado.

A modalidade deste Traballo Fin de Máster inclúe un período de prácticas, neste caso na Unidade de Epidemioloxía Clínica do Complexo Hospitalario Universitario de Santiago de Compostela (CHUS). En dita unidade trátase de dar apoio estadístico non só ó persoal do hospital senón tamén a aqueles estudantes que están realizando os seus Traballo Fin de Grao (TFG) ou Traballo Fin de Máster (TFM), ou incluso as súas teses doutorais (PhD).

Os casos prácticos que se recollen nas seguintes seccións forman parte do traballo realizado na Unidade de Epidemioloxía durante a realización das prácticas. O que se presenta no resto do capítulo só é unha parte dalgún dos casos cos que se traballou. Estes foron escollidos tanto pola relevancia clínica que presentan como pola variedade de situacións que están recollidas neles para poder así mostrar como os GAMLSS son capaces de abarcar distintas situacións. Obviamente ó longo das prácticas os casos tratados foron moitos máis, pero neste TFM só se recollen aqueles máis relevantes dos que están relacionados co tema deste traballo, os GAMLSS.

Na Sección 4.2 preséntanse un par de casos que se poden englobar no proxecto AEGIS. Este proxecto será presentado con máis detalle en dita sección así como cada un dos dous casos. Na Sección 4.3 trátase un exemplo que xa non está relacionado co proxecto da Estrada, pero que se considerou de interese para mostrar como se poden combinar modelos mixtos con GAMLSS.

4.2. AEGIS

Nesta sección trataremos dous casos prácticos relacionados co proxecto A Estrada Glycation and Inflammation Study (AEGIS), polo tanto comezaremos esta sección describindo brevemente dito proxecto para, posteriormente, analizar con máis detalle os dous casos cos que se traballou en maior profundidade.

AEGIS é un estudio transversal que conta con 1516 participantes e que foi levado a cabo para coñecer, dende unha perspectiva poboacional, cales son aqueles aspectos que inflúen nos procesos de inflamación e glicación.

AEGIS está dirixido polos Dres Arturo González-Quintela (Medicina Interna) e Francisco Gude (Epidemioloxía Clínica) e conta coa participación activa dun amplo grupo de profesionais pertencentes a diferentes ámbitos.

Foi levado a cabo no municipio de A Estrada (Pontevedra) que ten unha poboación maior de idade de 18744 habitantes e unha extensión de 282 km^2. A súa poboación é tanto rural como urbana cunha

proporción 1-4.

A selección dos participante levouse a cabo do seguinte xeito. Dado que a partir da tarxeta sanitaria se cobre máis do 95 % da poboación, tomouse a partir desta unha mostra estratificada por idade ([18, 30), [30, 40), [40, 50), [50, 60), [60, 70), [80, +80)). A mostra xerada foi de 500 persoas por cada grupo de idade, obtendo polo tanto un total de 3500 persoas. Deste total 428 non puideron ser incluídos no estudo porque este finalizou antes de que se intentase contactar con eles; ademais 84 morreran, 211 non responderon, 134 xa non residían no municipio e 19 non tiñan asistencia sanitaria. A estes casos houbo que sumarlles 394 persoas que non cumprían os criterios de inclusión por presentar demencia, atraso mental, enfermidades cerebrovascular graves, cáncer, enfermidade terminal ou incapacidade para comunicarse. Finalmente, daqueles cos que se contactou e que cumprían os criterios de inclusión, aceptaron participar no proxecto 1516 persoas.

A estas 1516 persoas realizóuselles unha entrevista clínica no centro de saúde de A Estrada entre novembro de 2012 e xuño de 2015. Ademais, recolléronse por cada participarte un cuestionario con datos demográficos e antropométricos; un rexistro do estilo de vida que incluía actividade física, alimentación e consumo de tabaco e alcohol; varios test psicolóxicos; un exame periodental; probas alérxicas e unha mostra de sangue.

Ademais, un subgrupo participou na parte de glicación do proxecto que incluía a monitorización continua da glicosa. Para esta parte do proxecto tiveron que ser descartadas 451 persoas tras incluír algún criterio máis de exclusión como foron a incapacidade para cumprimentar debidamente o protocolo, comer fóra asiduamente, ter alerxia ós adhesivos ou calquera condición médica que puidese afectar o bo funcionamento do dispositivo. Das 1065 persoas que podían participar nesta parte aceptaron facelo 622. Destas 622 persoas que aceptaron participar tiveron que ser descartadas 41 tras a monitorización, por non ser capaces de levala a cabo, quedado finalmente 581 participantes válidos.

Este estudio foi levado a cabo de acordo cos principios da Declaración de Helsinki e coa lexislación vixente. Foi aprobado polo Comité Ético de Investigación Clínica de Galicia, Santiago de Compostela, España (referencias 2010-315 y 2012-025).

Este proxecto conta con axudas da Xunta de Galicia (10CSA918028PR: Marcadores de inflamación y su relación con enfermedades frecuentes en la población general adulta. Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica. 2010-12), do Fondo de Investigacións Sanitarias (PI11/02219: Niveles de hemoglobina glicosilada y gap de glicación en relación con estilos de vida y las enfermedades prevalentes en la población general adulta. Fondo de Investigaciones Sanitarias. 2012 - 2014; RD12/0005/0007: Red de investigación en actividades preventivas y promoción de la salud en Atención Primaria (REDIAP). Fondo de Investigaciones Sanitarias. 2013 - 2015) e de Medtronic Inc.

A base de datos construída a partir dos datos recollidos nos cuestionarios e nas probas realizadas ten unha cantidade de variables bastantes extensa ademais, a partir destes datos base fóronse construíndo outras variables, ou ben categorizando algunhas variables continuas; ou ben construíndo variables resumo como pode ser a variable ipq (actividade física). Tamén se obtiveron novas variables, como por exemplo os índices de variabilidade (dos que se fala no primeiro exemplo) a partir das variables primarias.

Na seguinte lista recóllense tanto as variables explicativas utilizadas no exemplo dos índices de variabilidade como as que se usarán para o exemplo da VSG:

- age: esta variable recolle a idade de cada un dos participantes medida en anos e en valores enteiros.
- sex: esta variable recolle o sexo de cada un dos participantes, é dicir, home ou muller.
- bmi: esta variable recolle o índice de masa corporal. Este dato non estaba incluído na base inicial senón que foi calculado a partir do peso e da altura de cada un dos participantes, variables que sí formaban parte da base inicial.
- ipq: esta variable refírese á actividade física realizada por cada un dos participantes. Trátase dunha variable categórica con 3 categorías (baixa, moderada e alta). Esta é unha variable resumo obtida a partir de varias das preguntas recollidas na base inicial.

- tab012, tab: esta variable refírese ao consumo de tabaco de cada un dos participantes. Trátase dunha variable categórica con 3 categorías (non fumador, exfumador e fumador). Considerouse exfumador a todo aquel que levase máis dun ano sen fumar.

- oh3, oh4: estas variables refírense ao consumo de alcohol de cada un dos participantes. Trátase de variables categóricas con 3 (0-9, 10-139, 140-140+) e 4 (0-9, 10-139, 140-279, 280-280+) categorías respectivamente. A utilización destas 2 variables depende do criterio que desexe utilizar o profesional e construíronse a partir da variable grs da base inicial que recollía os gramos de alcohol consumidos por cada participante.

- sm: esta variable recolle o feito de se o participante ten ou non síndrome metabólico. Trátase, polo tanto, dunha variable categórica. Non é unha variable da base inicial senón que se construíu a posteriori a partir das variables dm (diabético si ou non), waist (medida da cintura en cm), tri (triglicéridos medidos en mg/dL), hta (presión arterial mmHg) e chdl (colesterol medido en mg/dL), se polo menos 3 destas variables se consideran anómalas considerouse que o participante tiña síndrome metabólico.

- glu: esta variable recolle a glicosa de cada un dos participantes en mg/dL.

- cal: esta variable recolle as calorías inxeridas por cada un dos participantes en Kcal.

- hc: esta variable recolle a porcentaxe de hidratos de carbono inxeridos por cada participante.

- lip: esta variable recolle a porcentaxe de lípidos inxeridos por cada participante.

- pro: esta variable recolle a porcentaxe de proteínas inxeridas por cada participante.

- fibra: esta variable recolle a porcentaxe de fibra inxerida por cada participante.

Unha vez contextualizados os datos e explicadas as variables que se utilizarán, imos presentar os dous casos para os que se levou a cabo un estudo máis detallado.

4.2.1. Índices de variabilidade

O primeiro dos exemplos que se expoñerán neste traballo para deixar patente a importancia dos GAMLSS esta relacionado co proxecto AEGIS e, máis concretamente, cos índices de variabilidade.

Dado que os índices de variabilidade están relacionados coa glicosa, a submostra que se utilizará para levar a cabo este estudo é a dos 581 individuos que participaron na parte de glicación do proxecto, onde accederon a levar un sensor durante 6 días que recollía os seus niveis de glicosa cada 5 minutos.

Tras un estudo e unha recompilación daqueles índices que resultaron máis interesantes, o seguinte paso foi calcular ditos índices a partir dos datos dispoñibles de glicosa. Tanto o código R para levar a cabo a construción dos índices, como unha explicación máis detallada de cada un deles (xunto con diversas referencias) poden atoparse en Rúa (2015). Neste traballo simplemente se presenta o Cadro 4.1, onde se recolle a fórmula utilizada para a construción de destes índices.

Índice	Fórmula	Variable
Área baixo a curva (AUC)	$\frac{1}{2}\sum_{i=1}^{N}(t_{i+1}-t_i)(BG_{i-1}+BG_i)$	auc
Glicosa media (MG)	MBG	mgi
M-value (M)	$\sum_{i=1}^{N}\frac{\left\lvert 10\log_{10}\left(\frac{BG}{IGV}\right)\right\rvert^3}{N}+\frac{\text{máx}(BG)-\text{mín}(BG)}{20}$	m
Average Daily Risk Ratio (ADRR)	$\frac{1}{m}\sum_{i=1}^{m}(LR+HR)$	adrr
Glycemic Risk Assessment in Diabetes Equation (GRADE)	$median(425(\log_{10}(\log_{10}(BG/18)+0.16)^2)$	grade
High Blood Glucose Index (HBGI)	$\frac{1}{N}\sum_{i=1}^{N} rh(BG_i)$	hbgi
Hiperglicemia	$100\frac{Nhyper}{N}$	hyper
Low Blood Glucose Index (LBGI)	$\frac{1}{N}\sum_{i=1}^{N} rl(BG_i)$	lbgi
Hipoglicemia	$100\frac{Nhypo}{N}$	hypo
Desviación estándar (SD)	$\sqrt{\frac{\sum_{i=1}^{N}(BG_i-MBG)^2}{N}}$	sd
Coeficiente de variación (CV)	$100\frac{SD}{MBG}$	cv
IQR	Q_3-Q_1	iqr
Lability Index (LI)	$\sum_{i=1}^{N-1}\frac{(BG_i-BG_{i+1})^2}{(t_{i+1}-t_i)}$	li
Mean Absolute Glucose (MAG)	$\frac{1}{T}\sum_{i=1}^{N-1}\lvert BG_i-BG_{i+1}\rvert$	mag
Mean Amplitude of Glycemic Excursions (MAGE)	$\sum\frac{\lambda}{x}$ se $\lambda > SD$	mage
Continuous Overlapping Net Glycemic Action (CONGA)	$\sqrt{\frac{\sum_{i=n+1}^{N}(D_i-\overline{D})^2}{N-1}}$ con $D_i=BG_i-BG_{i-n}$	conga
Mean Of Daily Differences (MODD)	$\frac{1}{K(m-1)}\sum_{i=1}^{N-K}\lvert BG_i-BG_{K+i}\rvert$	modd
J-Index (J)	$0.001(MBG+SD)^2$	ji

Cadro 4.1: Índices de variabilidade. $BG \equiv$ glicosa en sangre. $MBG \equiv$ media das medicións de glicosa. $N \equiv$ número de medicións. $Nhypo \equiv$ número de medicións para os que a glicosa estaba por debaixo de 70. $Nhype \equiv$ número de medicións para os que a glicosa estaba por encima de 140. $m \equiv$ número de días. $K \equiv$ medicións en 24 horas. $T \equiv$ tempo total (en horas). $\lambda \equiv$ diferencias de pico a val. $x \equiv$ número de diferencias válidas. $IGV \equiv$ valor ideal de glicosa.

Este estudo formará parte dunha tese doutoral na que se fará un estudo detallado de cada un dos índices recollidos no Cadro 4.1. Non obstante, a finalidade deste Traballo Fin de Máster non require o estudo de todos estes índices e dado que en total se recollen 18, utilizaremos compoñentes principais co método sparse para quedarnos cun número menor de índices nos que afondar dun xeito máis detallado.

Tras a utilización de compoñentes principais co método sparse, decidimos quedarnos cos seguintes índices: AUC, MG, IQR, MAGE. Dado que a saída obtida en R foi a seguinte:

```
> library(elasticnet)
> sparse<-spca(ind,2,para=rep(5,length=2),sparse="varnum",type="predictor")
> sum(sparse$pev)

[1] 0.7427751

> spcavec<-sparse$loadings
> spcavec

              PC1        PC2
auc    -0.2887347 -0.2899809
mgi    -0.4284621 -0.6009651
m       0.0000000  0.0000000
adrr    0.0000000  0.0000000
grade   0.0000000  0.0000000
hbgi    0.0000000  0.0000000
hyper   0.0000000  0.0000000
lbgi    0.0000000  0.0000000
hypo    0.0000000  0.0000000
sd      0.0000000  0.0000000
cv      0.0000000  0.0000000
iqr    -0.5282546  0.4785220
li      0.0000000  0.0000000
mag     0.0000000  0.0000000
mage   -0.5347506  0.5519826
conga   0.0000000  0.0000000
modd    0.0000000  0.0000000
ji      0.0000000  0.0000000
```

Polo tanto estes 4 índices explican o 74 % da variabilidade total.

O obxectivo da tese mencionada é estudar os valores de cada un dos 18 índices en función das covariables age, sex, bmi, ipq, tab, oh3, glu, cal, hc, lip, pro e fibra para a poboación total e tamén dar valores de referencia para ditos índices, centrándose xa na poboación normoglucémica. O obxectivo neste traballo será o mesmo, só que para os índices escollidos previamente.

Antes de comezar coa análise, considero necesario explicar que a poboación total tamén se dividiu en tres grupos: normoglucémicos, prediabéticos e diabéticos. Esta división foi feita a partir das porcentaxes de hemoglobina glicada de tal xeito que se clasificou como individuos normoglucémicos, prediabéticos e diabéticos aqueles que tiñan a porcentaxe de hemoglobina glicada menor que 5.7 %, entre 5.7 e 6.4 % e maior de 6.4 % respectivamente, para intentar ver un certo patrón nos valores dos índices de cada subgrupo.

Neste exemplo traballaremos coa poboación total á hora de ver a relación de cada un dos índices coas variables explicativas e centrarémonos na poboación normoglucémica cando o propósito sexa dar valores de referencia.

Comezaremos polo estudo que atinxe á poboación total. O primeiro paso será asignarlle unha distribución axeitada a cada un dos índices para, posteriormente, axustar o modelo GAMLSS coas covariables mencionadas e tendo en conta a distribución escollida.

Para cada un dos índices faremos 4 propostas de distribución a primeira será a distribución normal que a teremos sempre de referencia e as outras 3 serán escollidas coa axuda da función fitDist() que utiliza o criterio AIC para escoller as distribucións que mellor se adaptan ós datos. Para este exemplo

utilizaremos o argumento type = "realAll" da función fitDist(), xa que os valores dos índices poden estar ó longo de toda a recta real (ou ser positivos dependendo do caso).

Nun primeiro momento optaremos por utilizar aquela distribución que obteña un menor AIC, pero non desbotaremos as outras opcións dados os problemas de converxencia cos que nos fomos atopando ó longo do estudo.

Aínda que nun principio unha das distribucións resulte máis axeitada polo valor do AIC, ó intentar axustar un modelo con demasiadas variables, ou con variables suavizadas, é dicir, un modelo máis complexo, pode ser que aparezan problemas de converxencia e nos teñamos que decantar por utilizar unha distribución que a priori tivo peores resultados utilizando o criterio AIC, pero que se comporta mellor con respecto á converxencia.

As distribucións propostas para cada un dos índices son as que aparecen na Figuras 4.1 e 4.2. Finalmente, tras ter en conta o valor do AIC e os posibles problemas de converxencia, as distribucións escollidas para cada un dos índices foron as que se presentan na Figura 4.3. No Cadro 4.2 están recollidos os parámetros de cada unha delas.

Índice (Distribución)	μ (log)	σ (log)	ν	τ (log)
AUC (BCTo)	4.558049	-2.38047	-5.126	1.022
MG (BCTo)	4.669867	-2.34312	-5.4707	1.3806
IQR (BCPEo)	2.89995	-0.94782	-0.7571	0.51899
MAGE (BCPEo)	3.31474	-0.97492	-0.62057	-0.49136

Cadro 4.2: Parámetros para cada unha das distribucións escollidas para cada un dos 4 índices (AUC, MG, IQR, MAGE) para a poboación total.

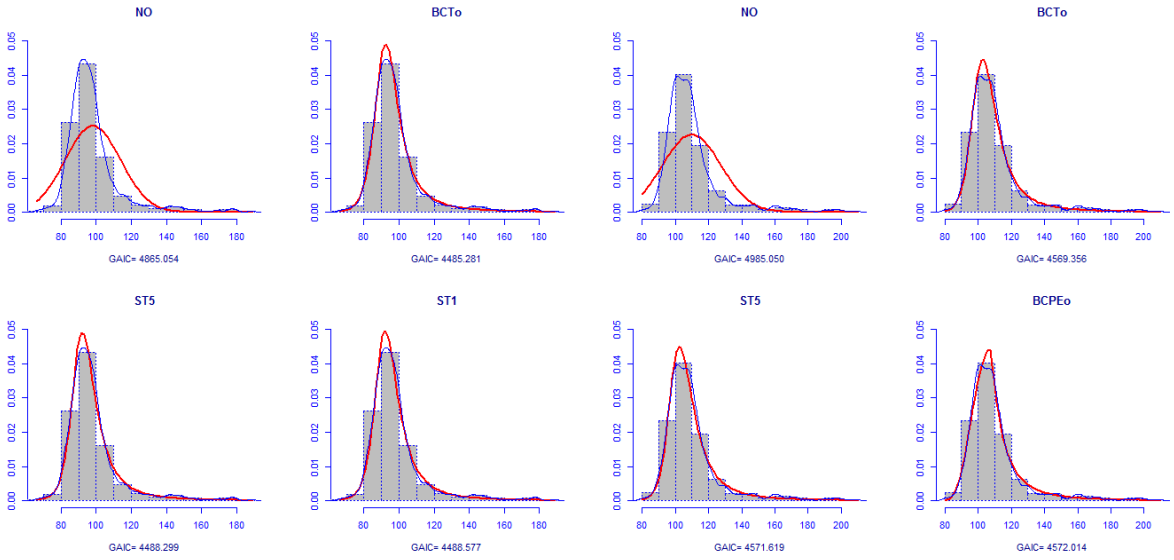

Figura 4.1: Opcións de distribución para a poboación total para os índices AUC e MG, esquerda e dereita respectivamente.

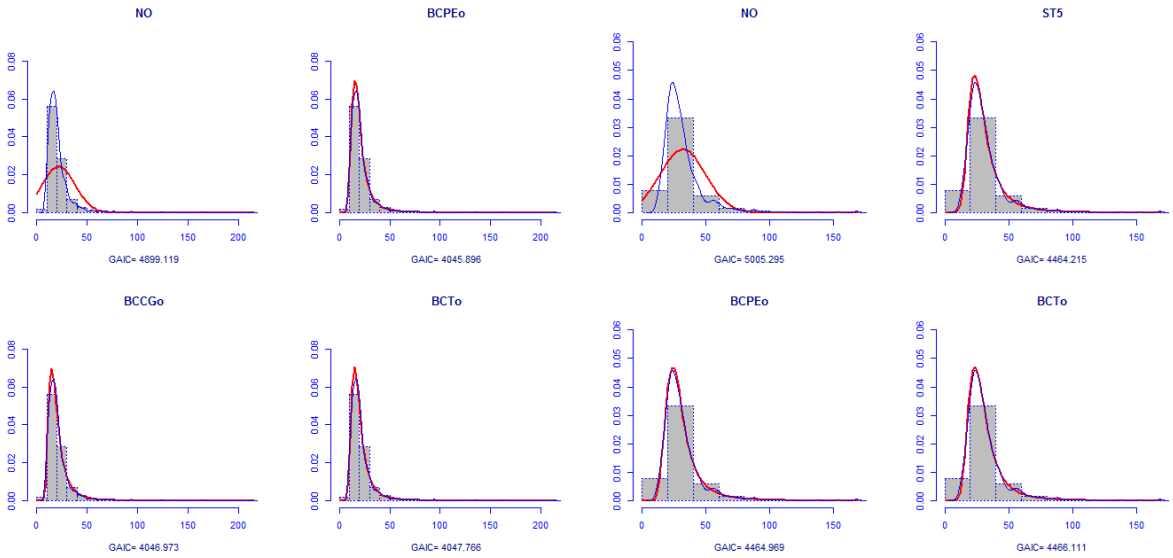

Figura 4.2: Opcións de distribución para a poboación total para os índices IQR e MAGE, esquerda e dereita respectivamente.

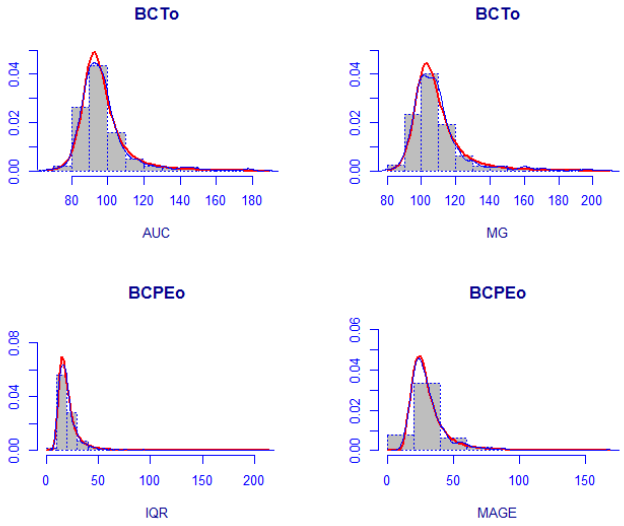

Figura 4.3: Distribucións escollidas para a poboación total para os 4 índices (AUC, MG, IQR, MAGE).

Levaremos a cabo agora unha regresión múltiple para cada un dos índices coas covariables mencionadas ó comezo e tendo en conta a distribución escollida para cada un deles. Pero previamente presentamos un resumo dos valores das covariables obtidos na poboación total.

Nesta submostra de 581 participantes perdéronse 16 casos por faltar algún dos datos, polo tanto téñense 565 persoas das cales 355 (63 %) son mulleres e 210 (37 %) homes. Ademais 205 (36 %) facían pouca actividade física, 216 (38 %) moderada e 144 (26 %) moita. Hai 305 (54 %) persoas que non fuman, 150 (27 %) exfumadores e 110 (19 %) fumadores. Con respecto ó consumo de alcohol téñense

226 (40 %) que apenas beben, 271 (48 %) que teñen un consumo moderado e 68 (12 %) que teñen un consumo alto. A media (xunto co máximo e o mínimo) das variables continuas aparecen no Cadro 4.3.

Variable	Mínimo	Media	Máximo
age	18	48.25	87
bmi	17.36	28.2	52.54
glu	63	93.38	254
cal	715	2064	4358
pro	10	18.18	29
lip	19	34.04	49
hc	32	47.76	66
fibra	3	19.66	52

Cadro 4.3: Datos resumo das covariables continuas da poboación total para os índices de variabilidade.

Vexamos que resultados se obteñen para cada un dos 4 índices:

- AUC:

  O primeiro modelo que se executou foi:

$$y \sim BCTo\{\log(\mu) = pb(glu) + pb(age) + sex + pb(bmi) + ipq + tab + oh3 + pb(cal) + pb(hc)$$

$$+pb(lip) + pb(pro) + pb(fibra), \log(\sigma) = pb(age) + sex, \nu = 1, \log(\tau) = 1\}$$

  pero tras comprobar cales das variables non necesitaban suavización quedámonos co modelo

$$y \sim BCTo\{\log(\mu) = pb(glu) + age + sex + bmi + ipq + tab + oh3 + cal + hc$$

$$+lip + pro + fibra, \log(\sigma) = pb(age) + sex, \nu = 1, \log(\tau) = 1\}$$

  Este é o modelo de interese para o clínico. Tanto a saída da función summary() como a saída da función term.plot() que se pode ver nas Figuras 4.4 e 4.5 indican que, para o parámetro de localización, a relación entre a AUC e a idade, o índice de masa corporal, o consumo de tabaco e as calorías é crecente (aínda que no bmi non significativamente) mentres que é decrecente para o consumo de alcohol, os hidratos de carbono, os lípidos, as proteínas e a fibra (significativamente en todas elas). A relación que presenta a AUC coa glicosa non é lineal (pero sí significativa), é crecente na parte central dos datos pero parece estancarse tanto nos datos máis altos como nos máis baixos. O sexo e a actividade física non resultan significativos.

  Para o parámetro de escala a idade tamén resulta significativa, neste caso a relación non é completamente lineal pero pódese dicir que é crecente. Non obstante, o sexo non é significativo.

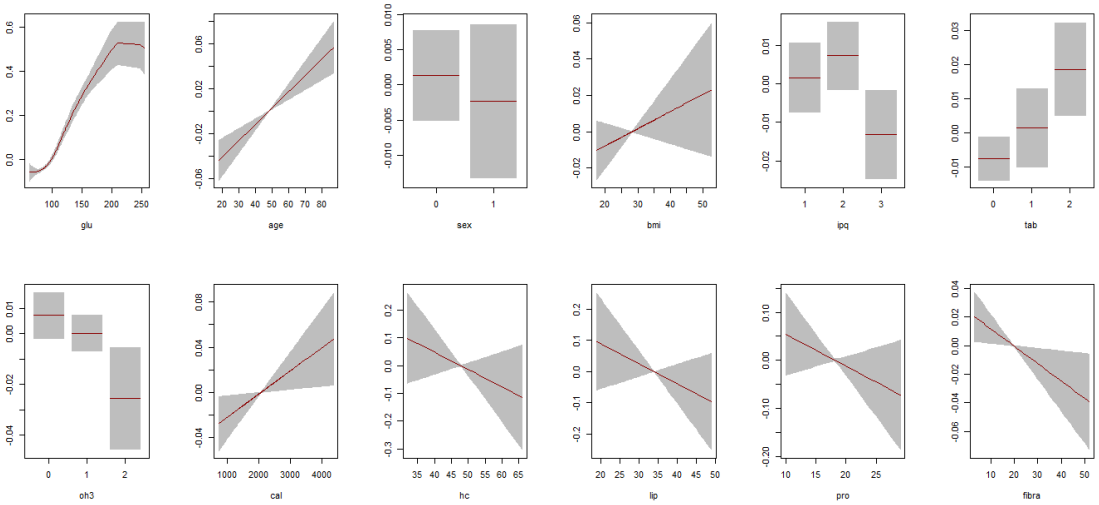

Figura 4.4: Saída da función term.plot() para o parámetro μ para a AUC.

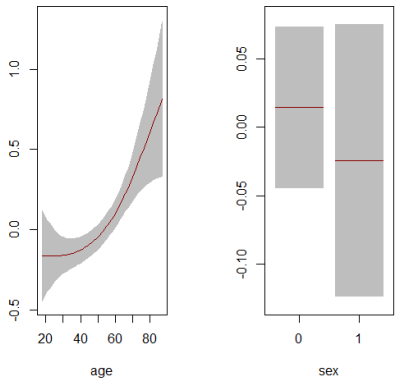

Figura 4.5: Saída da función term.plot() para o parámetro σ para a AUC.

Ó último modelo aplicóuselle a función stepGAICAll.A() para intentar simplificalo e obtívose a partir desta o modelo

$$y \sim BCTo\{\log(\mu) = pb(glu)+age+ipq+tab+oh3+cal+fibra, \log(\sigma) = pb(age), \nu = 1, \log(\tau) = 1\}$$

onde se descartaron as variables sex, bmi, hc, lip e pro para o parámetro μ e a variable sex para o parámetro σ. Para este modelo todas as variables, excepto a ipq, son significativas (para as categóricas polo menos unha das categorías).

Este modelo, ademais, foi validado utilizando as funcións plot() e wp(), Figura 4.6.

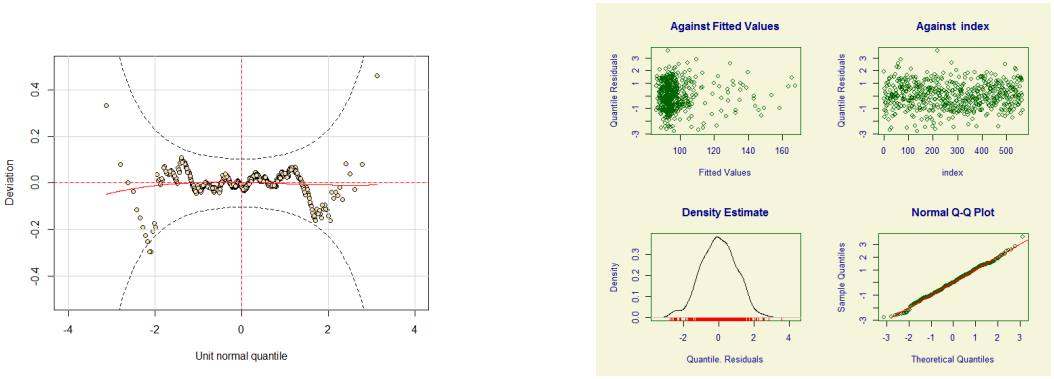

Figura 4.6: Gráficos de validación para o modelo axustado para o índice AUC.

Para os 3 índices restantes farase unha análise análoga á que se fixo para a AUC.

- MG:

  O primeiro modelo que se executou foi:

$$y \sim BCTo\{\log(\mu) = pb(glu) + pb(age) + sex + pb(bmi) + ipq + tab + oh3 + pb(cal) + pb(hc)$$

$$+pb(lip) + pb(pro) + pb(fibra), \log(\sigma) = pb(age) + sex, \nu = 1, \log(\tau) = 1\}$$

  pero tras comprobar cales das variables non necesitaban suavización quedámonos co modelo

$$y \sim BCTo\{\log(\mu) = pb(glu) + age + sex + bmi + ipq + tab + oh3 + cal + hc$$

$$+lip + pro + fibra, \log(\sigma) = pb(age) + sex, \nu = 1, \log(\tau) = 1\}$$

  Este é o modelo de interese para o clínico. Tanto a saída da función summary() como a saída da función term.plot() que se pode ver nas Figuras 4.7 e 4.8 indican que, para o parámetro de localización, a relación entre a MG e a idade, o índice de masa corporal, o consumo de tabaco e as calorías é crecente (aínda que no bmi non significativamente) mentres que é decrecente para o consumo de alcohol, os hidratos de carbono, os lípidos, as proteínas e a fibra (significativamente en todas elas). A relación que presenta a MG coa glicosa non é lineal (pero sí significativa), é crecente na parte central dos datos pero parece estancarse tanto nos datos máis altos como nos máis baixos. O sexo e a actividade física non resultan significativos.

  Para o parámetro de escala a idade tamén resulta significativa mentres que o sexo non. A relación entre σ e a idade é crecente e practicamente lineal aínda que parece estancarse nas idades máis baixas como ocorría coa AUC.

  Os resultados obtidos para a MG son moi similares aos obtidos para a AUC.

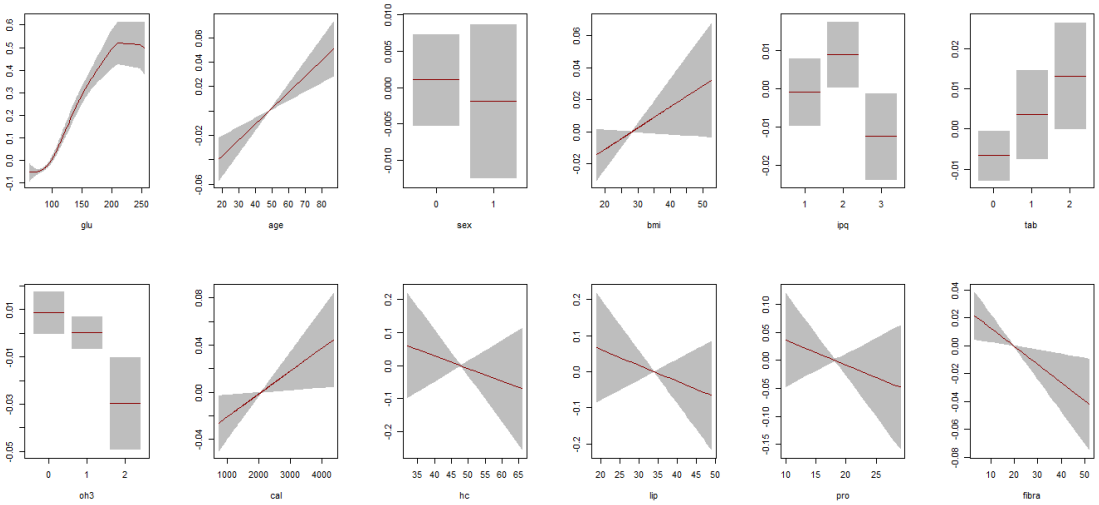

Figura 4.7: Saída da función term.plot() para o parámetro μ para a MG.

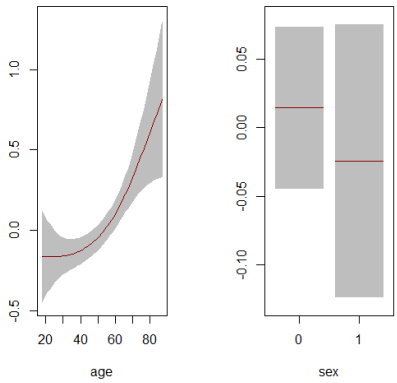

Figura 4.8: Saída da función term.plot() para o parámetro σ para a MG.

Ó último modelo aplicóuselle a función stepGAICAll.A() obténdose a partir desta o modelo

$$y \sim BCTo\{\log(\mu) = pb(glu) + age + bmi + ipq + tab + oh3 + cal + fibra,$$

$$\log(\sigma) = pb(age), \nu = 1, \log(\tau) = 1\}$$

onde se descartaron as variables sex, hc, lip e pro para o parámetro μ e a variable sex para o parámetro σ. Para este modelo todas as variables, excepto o bmi e a ipq, son significativas (para as categóricas polo menos unha das categorías).

Este modelo, ademais, foi validado utilizando as funcións plot() e wp(), Figura 4.9.

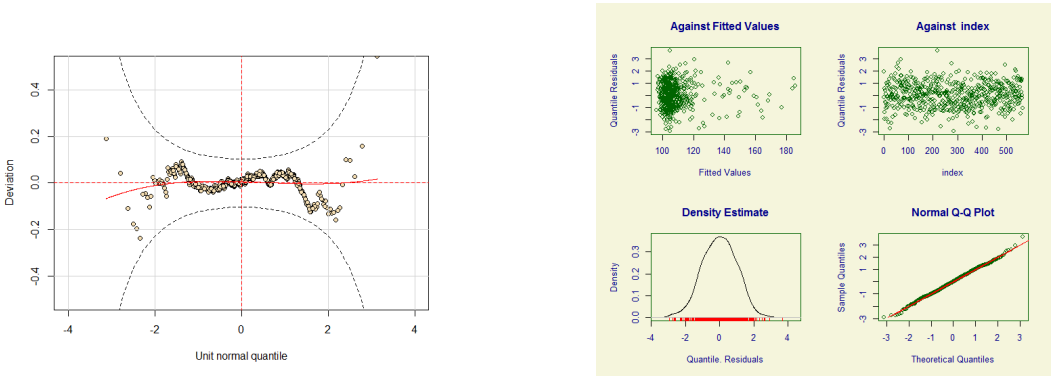

Figura 4.9: Gráficos de validación para o modelo axustado para o índice MG.

- IQR:

  O primeiro modelo que se executou foi:

$$y \sim BCPEo\{\log(\mu) = pb(glu) + pb(age) + sex + pb(bmi) + ipq + tab + oh3 + pb(cal) + pb(hc)$$

$$+pb(lip) + pb(pro) + pb(fibra), \log(\sigma) = pb(age) + sex, \nu = 1, \log(\tau) = 1\}$$

  pero tras comprobar cales das variables non necesitaban suavización quedámonos co modelo

$$y \sim BCPEo\{\log(\mu) = pb(glu) + age + sex + bmi + ipq + tab + oh3 + cal + hc$$

$$+lip + pro + fibra, \log(\sigma) = age + sex, \nu = 1, \log(\tau) = 1\}$$

  Este é o modelo de interese para o clínico. Tanto a saída da función summary() como a saída da función term.plot() que se pode ver nas Figuras 4.7 e 4.8 indican que, para o parámetro de localización, a relación entre o IQR e a idade é crecente (significativamente) mentres que é decrecente para o índice de masa corporal, o consumo de alcohol, as calorías, os hidratos de carbono, os lípidos, as proteínas e a fibra (o bmi, cal, hc, lip, pro e fibra non significativamente). A relación que presenta o IQR coa glicosa non é lineal (pero sí significativamente), é crecente na parte central dos datos pero parece decrecer tanto nos datos máis altos como nos máis baixos. O sexo e a actividade física tampouco resultan significativos. O consumo de tabaco si sae significativo obténdose valores máis altos de IQR para fumadores e non fumadores (lixeiramente maiores, aínda que non significativamente, para fumadores).

  Para o parámetro de escala a idade tamén resulta significativa, a relación é crecente. O sexo non é significativo.

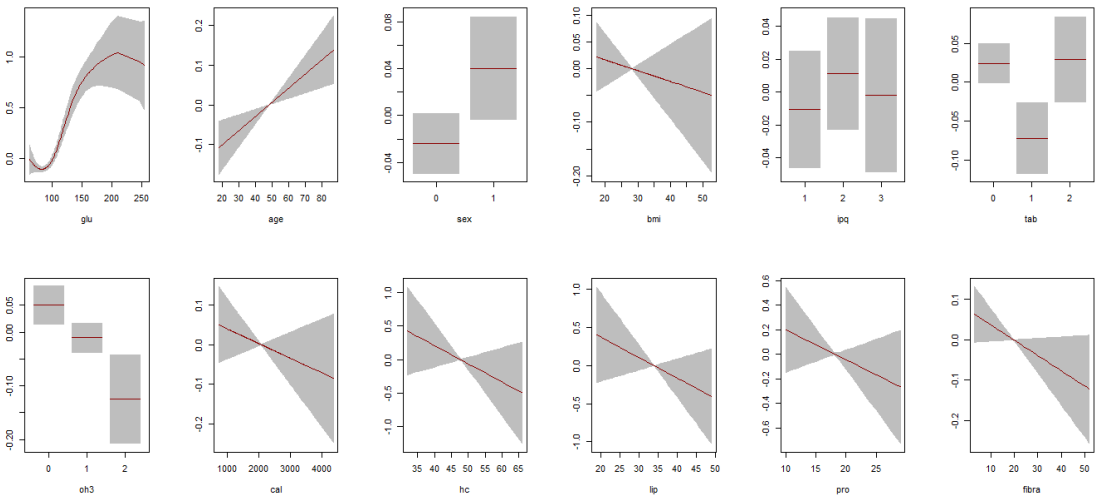

Figura 4.10: Saída da función term.plot() para o parámetro μ para o IQR.

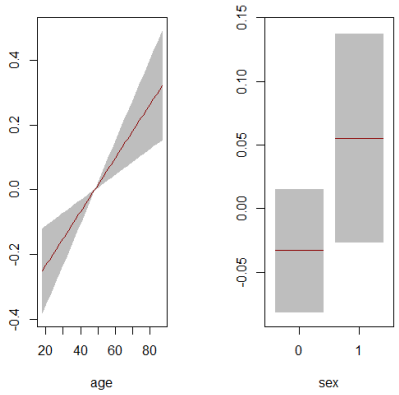

Figura 4.11: Saída da función term.plot() para o parámetro σ para o IQR.

Ó último modelo aplicóuselle a función stepGAICAll.A() obténdose a partir desta o modelo

$$y \sim BCPEo\{\log(\mu) = pb(glu) + age + tab + oh3 + fibra, \log(\sigma) = age, \nu = 1, \log(\tau) = 1\}$$

onde se descartaron as variables sex, bmi, ipq, cal, hc, lip e pro para o parámetro μ e a variable sex para o parámetro σ. Para este modelo todas as variables son significativas (para as categóricas polo menos unha das categorías).

Este modelo, ademais, foi validado utilizando as funcións plot() e wp(), Figura 4.12.

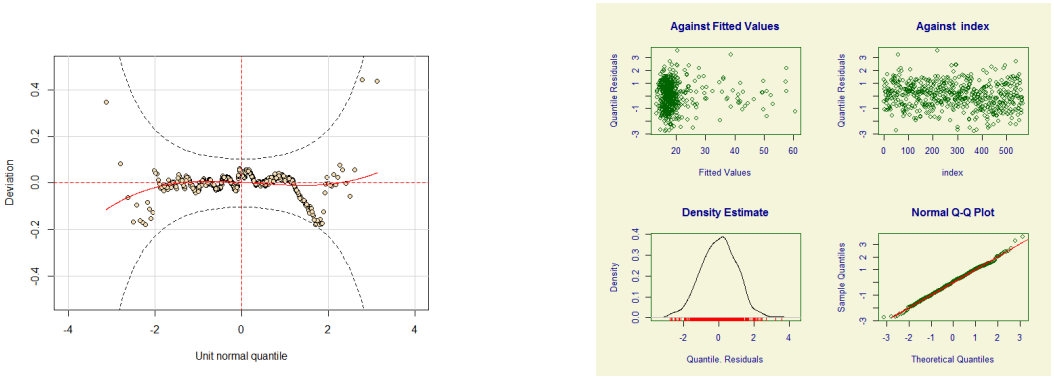

Figura 4.12: Gráficos de validación para o modelo axustado para o índice IQR.

- MAGE:

  O primeiro modelo que se executou foi:

$$y \sim BCPEo\{\log(\mu) = pb(glu) + pb(age) + sex + pb(bmi) + ipq + tab + oh3 + pb(cal) + pb(hc)$$

$$+pb(lip) + pb(pro) + pb(fibra), \log(\sigma) = pb(age) + sex, \nu = 1, \log(\tau) = 1\}$$

  pero tras comprobar cales das variables non necesitaban suavización quedámonos co modelo

$$y \sim BPEo\{\log(\mu) = pb(glu) + age + sex + bmi + ipq + tab + oh3 + cal + hc$$

$$+lip + pro + fibra, \log(\sigma) = age + sex, \nu = 1, \log(\tau) = 1\}$$

  Este é o modelo de interese para o clínico. Tanto a saída da función summary() como a saída da función term.plot() que se pode ver nas Figuras 4.7 e 4.8 indican que, para o parámetro de localización, a relación entre a MAGE e a idade, o índice de masa corporal, o consumo de tabaco e as calorías (o bmi non significativamente) é crecente mentres que é decrecente para o consumo de alcohol, os hidratos de carbono, os lípidos, as proteínas e a fibra (os hc, lip e pro non significativamente). A relación que presenta a MAGE coa glicosa non é lineal (pero sí significativamente), é crecente na parte central dos datos pero parece decrecer nos datos máis altos e estancarse nos máis baixos. O sexo non resulta significativo e a actividade física sí que o é nalgunha das categorías; para unha actividade física alta o valor da MAGE baixa.

  Para o parámetro de escala a idade tamén resulta significativa, a relación é crecente. O sexo non é significativo.

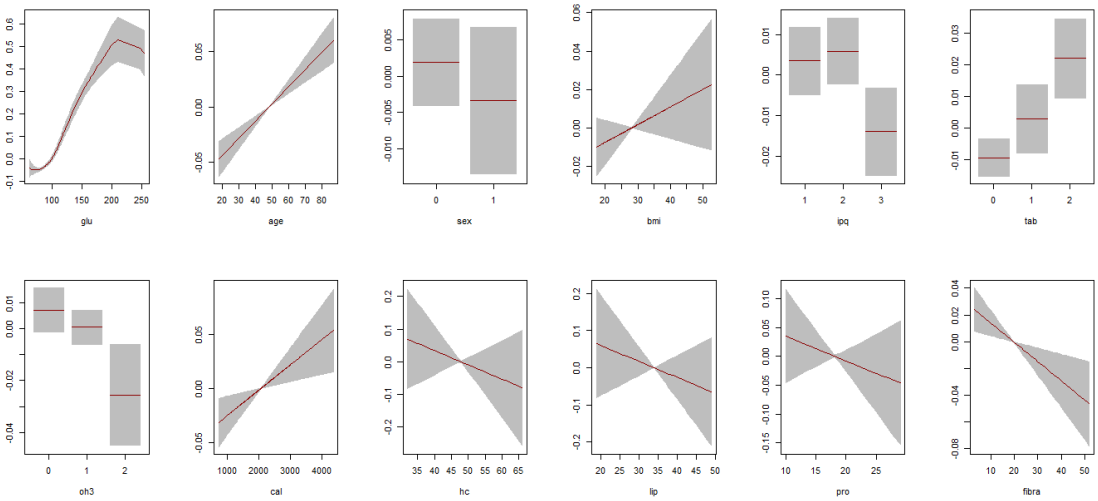

Figura 4.13: Saída da función term.plot() para o parámetro μ para a MAGE.

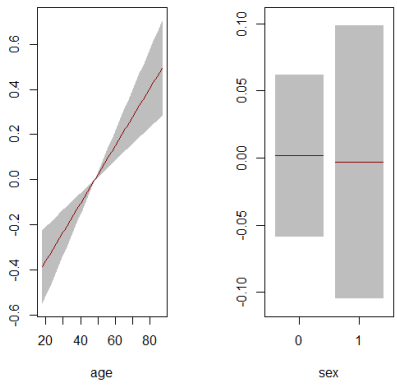

Figura 4.14: Saída da función term.plot() para o parámetro σ para a MAGE.

Ó último modelo aplicóuselle a función stepGAICAll.A() obténdose a partir desta o modelo

$$y \sim BCPEo\{\log(\mu) = pb(glu)+age+ipq+tab+oh3+cal+fibra, \log(\sigma) = age, \nu = 1, \log(\tau) = 1\}$$

onde se descartaron as variables sex, bmi, hc, lip e pro para o parámetro μ e a variable sex. Para este modelo todas as variables son significativas, excepto a ipq; para as categóricas polo menos unha das categorías.

Este modelo, ademais, foi validado utilizando as funcións plot() e wp(), Figura 4.15.

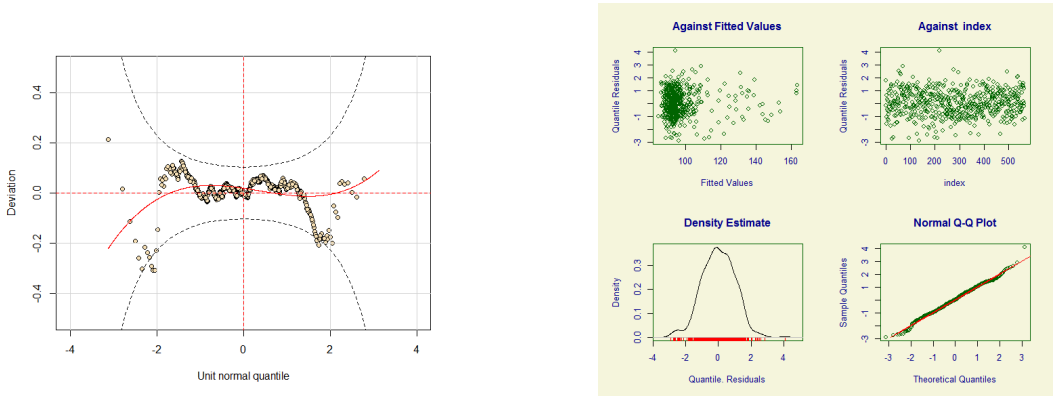

Figura 4.15: Gráficos de validación para o modelo axustado para o índice MAGE.

En conclusión, o estudo destes 4 índices dá lugar a resultados similares. Se nos centramos no parámetro μ:

- O sexo non influe en ningún deles nin está presente no modelo final.

- Hai variables que son significativas para os 4 índices. A glicosa (obviamente), a idade, o consumo de tabaco ou de alcohol forman parte do modelo final dos 4 índices e son tamén significativas todas elas nos 4 casos.

Se nos referimos ó parámetro σ a situación é análoga. A idade é significativa nos 4 casos mentres que o sexo non influe en ningún.

Unha análise máis correcta e completa sería axustar os 4 parámetros das distribucións (BCTo, BCPEo) a partir de todas as covariables de interese, pero isto daría lugar a un modelo moi complexo en varios sentidos. Primeiramente, utilizar a función stepGAICAll.A() requiriría un tempo considerable. Tamén provocaría problemas de converxencia xa que, incluso nestes modelos máis sinxelos, se tivo que estar xogando tanto co número de iteracións como coa mestura dos 2 algoritmos dispoñibles.

Ademais, a interpretación do modelo obtido, se finalmente os 4 parámetros son axustados por algunha das covariables, sería difícil de interpretar. Por isto, para a finalidade, tanto do tesis doutoral na que irá incluída esta análise como para este Traballo Fin de Máster considerouse suficiente traballar co modelo inicial que se indicou.

Agora daremos valores de referencia para cada un dos índices, centrándonos soamente na poboación normoglucémica. Dado que cambiamos da poboación total á poboación normoglucémica, os datos cambian. A poboación normoglucémica esta composta por 430 persoas das cales se tiveron que desbotar 13 casos por ter datos faltantes. Dos 417 que quedaron 265 (64 %) son mulleres e 152 (36 %) homes. 150 (36 %) facían pouca actividade física, 155 (37 %) moderada e 112 (27 %) moita. Con respecto ó consumo de tabaco 217 (52 %) son non fumadores, 107 (26 %) exfumadores e 93 (22 %) fumadores. Ademais, 159 (38 %) apenas toman alcohol, 211 (50 %) tómano dun xeito moderado e 47 (12 %) excédense no seu consumo. A media (xunto co máximo e o mínimo) das variables continuas aparecen no Cadro 4.4. Na distribución dentro das variables categóricas non hai apenas cambios ao pasar da poboación total á normoglucémica. Non obstante, nas medias das variables continuas si que se poden apreciar diferencias. Obviamente, a media da glicosa descende considerablemente, pero tamén o fai a media da idade e o índice de masa corporal, aínda que non tanto. A media das calorías aumenta lixeiramente, pero a variacións nas porcentaxes de proteínas, lípidos, hidratos de carbono e fibra apenas se nota.

Variable	Mínimo	Media	Máximo
age	18	44.36	81
bmi	17.36	27.08	52.54
glu	63	85.9	118
cal	907	2120	4358
pro	10	17.64	29
lip	19	34.2	49
hc	32	48.13	66
fibra	5	19.27	51

Cadro 4.4: Datos resumo das covariables continuas da poboación normoglucémica para os índices de variabilidade.

O que si que cambiou seguramente é a distribución dos datos dos 4 índices. Polo tanto debemos realizar de novo unha busca da distribución máis axeitada para cada índice na poboación normoglucémica. Ó igual que fixemos para a poboación total, tomaremos 4 opcións para cada índice, mantendo a distribución normal como referencia e as 3 primeiras que nos proporcione a función fitDist() co argumento type = "realAll".

Neste caso as distribucións propostas son as que aparecen nas Figuras 4.16 e 4.17 e as que se escolleron as que aparecen na Figura 4.18. No Cadro 4.5 aparecen recollidos os parámetros das distribucións escollidas.

Índice (Distribución)	μ	σ	ν	τ
AUC (LO)	93.0692	exp(1.43099)	-	-
MG (BCTo)	exp(4.643825)	exp(-2.63772)	-0.1337	exp(2.6829)
IQR (IG)	exp(2.88409)	exp(-2.6237)	-	-
MAGE (RG)	23.1107	exp(1.8854)	-	-

Cadro 4.5: Parámetros para cada unha das distribución escollidas para cada un dos 4 índices (AUC, MG, IQR, MAGE) para a poboación normoglucémica.

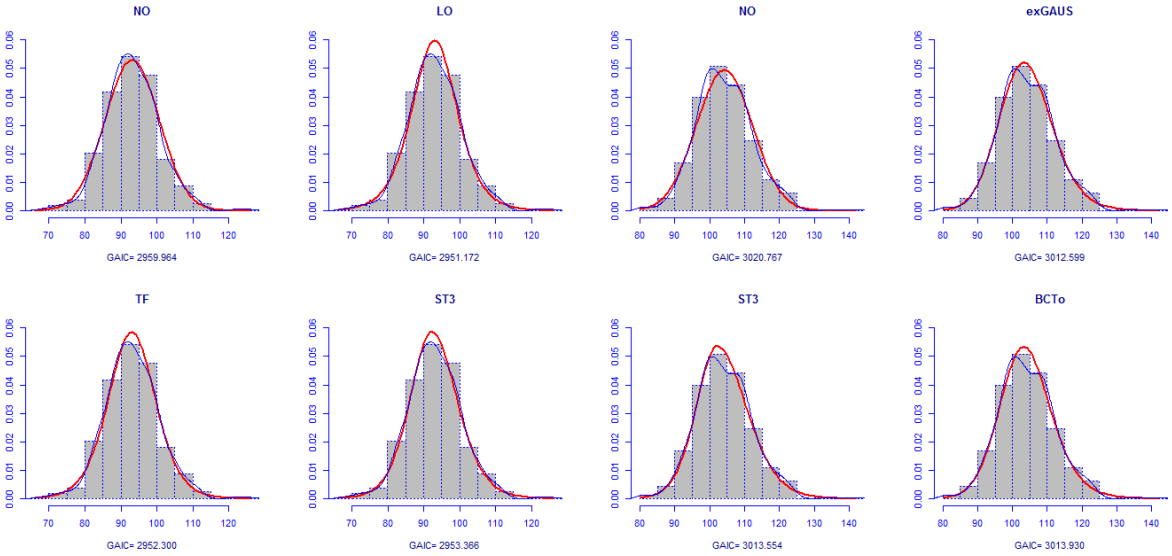

Figura 4.16: Opcións de distribución para a poboación normoglucémica para os índices AUC e MG, esquerda e dereita respectivamente.

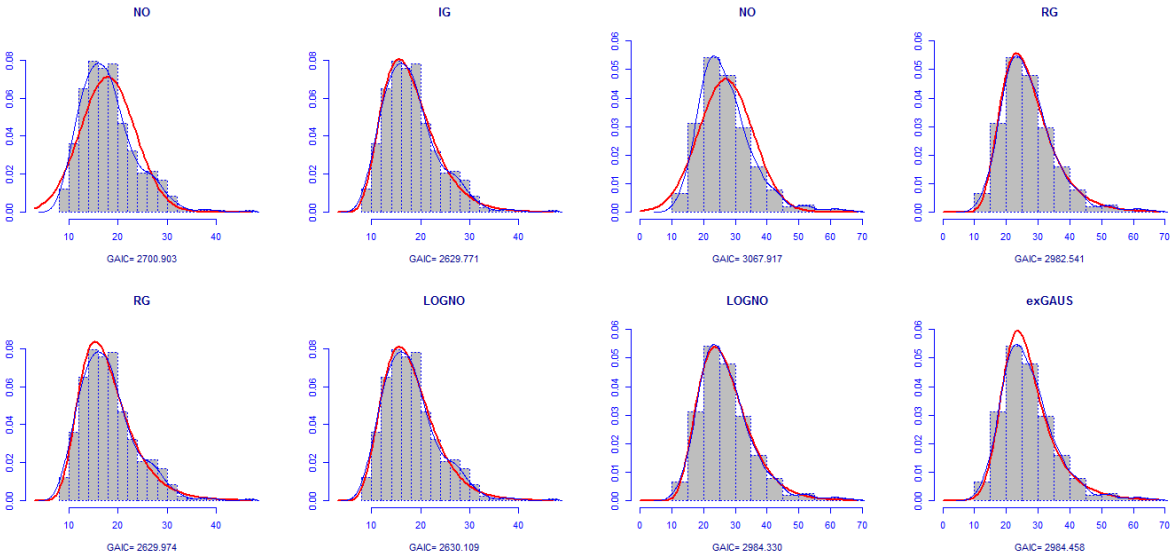

Figura 4.17: Opcións de distribución para a poboación normoglucémica para os índices IQR e MAGE, esquerda e dereita respectivamente.

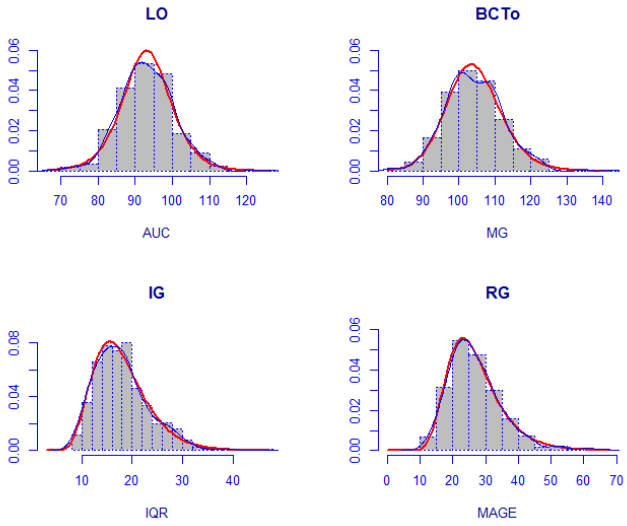

Figura 4.18: Distribucións escollidas para a poboación normoglucémica para os 4 índices (AUC, MG, IQR, MAGE).

Unha vez que temos as distribucións escollidas, botaremos man das funcións relacionadas cos centiles para levar a cabo un estudo sobre os valores de referencia de cada un dos índices atendendo á idade e ó sexo.

Para cada un dos 4 índices comezaremos axustando o seguinte modelo:

$$y \sim D\{g(\mu) = pb(age) + sex, g(\sigma) = pb(age) + sex, g(\nu) = 1, g(\tau) = 1\}$$

onde, como xa se definiu máis veces ao longo do traballo, D denota a distribución usada e $g(.)$ as funcións link para cada un dos parámetros.

Posteriormente utilizouse a función drop1() para seleccionar o modelo máis axeitado, obténdose os seguintes resultados:

- AUC:
$$y \sim LO\{\mu = age, \log(\sigma) = 1\}$$
- MG:
$$y \sim BCTo\{\log(\mu) = age, \log(\sigma) = 1, \nu = 1, \log(\tau) = 1\}$$
- IQR:
$$y \sim IG\{\log(\mu) = 1, \log(\sigma) = 1\}$$
- MAGE:
$$y \sim RG\{\mu = age, \log(\sigma) = age + sex\}$$

Polo tanto, a AUC e a MG non dependen do sexo aínda que sí da idade (non suavizada) como se pode ver na Figura 4.19.

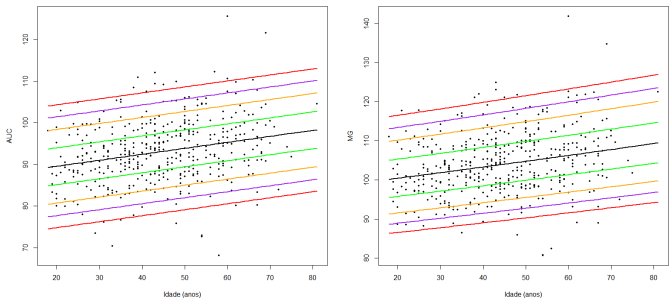

Figura 4.19: Valores de referencia para os índices AUC e MG con respecto á idade.

Non obstante, para o IQR nada influe, o mellor AIC obtense co modelo sen idade nin sexo (o modelo sen covariables). Isto provocará que á hora de obter os valores de referencia con respecto á idade estes sexan (casi) constantes para cada percentil como se observa na Figura 4.20.

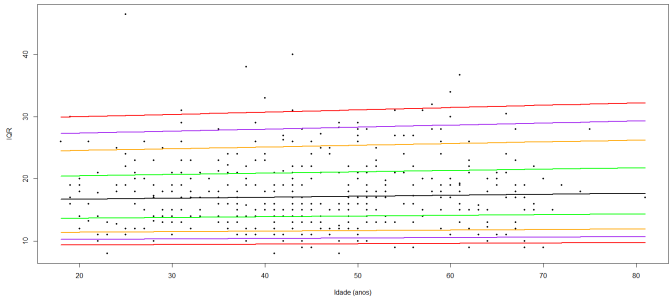

Figura 4.20: Valores de referencia para o índice IQR con respecto á idade.

Por último, a MAGE depende tanto da idade como do sexo aínda que o sexo só influe no parámetro de escala, isto pode verse na Figura 4.21, onde o rango de valores é lixeiramente diferente para homes e mulleres.

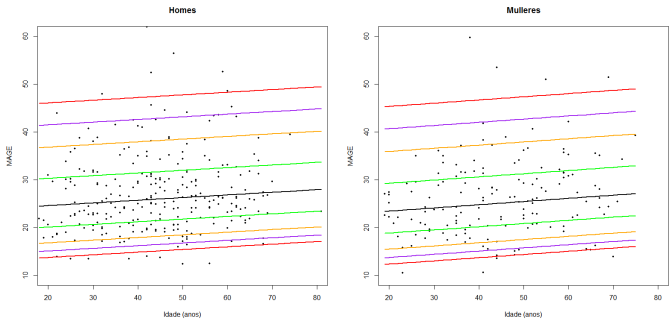

Figura 4.21: Valores de referencia para o índice MAGE con respecto á idade e diferenciando por sexo.

4.2.2. Factores de inflamación

Neste segundo exemplo cambiamos os índices de variabilidade polos factores de inflamación e os 581 participantes polos 1516. Neste caso seguiremos traballando cos datos do proxecto AEGIS, pero para este exemplo si que podemos dispoñer do total dos participantes. Agora centraremos a nosa atención nos factores de inflamación, máis concretamente na velocidade de sedimentación globular (VSG).

A VSG é a precipitación dos glóbulos vermellos (eritrocitos) nun tempo determinado, que se relaciona directamente coa tendencia dos glóbulos vermellos cara a formación de acúmulos. Este factor utilízase para detectar procesos inflamatorios ou infecciosos.

De novo, este exemplo será incluído nunha tese doutoral, onde ademais da VSG se estudarán en detalle algunhas citoquinas como a interleuquina 8 (IL-8), o factor de necrose tumoral (TNF) ou a reacción de cadea da polimerasa (PCR). O obxectivo principal de dita tese é axustar os valores destas citoquinas atendendo as covariables age, sex, bmi, ipq, tab012, oh4 e sm.

Neste traballo o obxectivo será o mesmo, pero só presentaremos o estudo detallado da VSG.

Dado que agora traballamos coa mostra total, as características desta cambian. En primeiro lugar tivéronse que descartar 17 persoas por ter enfermidades que podían influír nos valores da VSG. Ademais, hai 27 valores perdidos da VSG; polo tanto quedámonos cun tamaño mostral final de 1472 persoas das cales 817 (56 %) son mulleres e 655 (44 %) homes. Hai 573 (39 %) persoas que fan pouca actividade física, 538 (36 %) que fan unha actividade física moderada e 361 (25 %) que fan moita actividade física. Ademais, 808 (55 %) son non fumadores, 382 (26 %) exfumadores e 282 (19 %) fumadores. Con respecto ó consumo de alcohol hai 534 (36 %) persoas que apenas beben, 577 (39 %) que teñen un consumo moderado, 233 (16 %) que beben moito e 128 (9 %) que beben excesivamente. Dentro dos 1472, 1172 (80 %) non teñen síndrome metabólico e 300 (20 %) si. A idade media é de 52.51, tendo a persoa máis nova 18 anos e a máis vella 91. Atendendo ó índice de masa corporal a media é de 28.22, sendo o valor máis baixo 17.36 e o máis alto 50.7.

A primeira parte do estudo é similar para todos os exemplos, consiste en atopar unha distribución adecuada para a variable resposta, neste caso a VSG.

A única diferenza co exemplo dos índices de variabilidade é que neste caso sabemos que os valores da VSG son estritamente positivos, polo tanto escolleremos como distribucións a comparar: a distribución normal, unha das distribucións obtidas co argumento type = "realplus" da función fitDist(), unha terceira obtida co argumento type = "realline" da mesma función e, por último, esta terceira distribución truncada en 0.

As distribucións escollidas para a VSG foron as que aparecen na Figura 4.22, destas 4 decidimos quedarnos coa GIG, fundamentalmente por problemas de converxencia. As estimacións dos parámetros da distribución escollida son: 2.52196 para μ (en escala logarítmica), 0.05186 para σ (en escala logarítmica) e 0.4221 para ν.

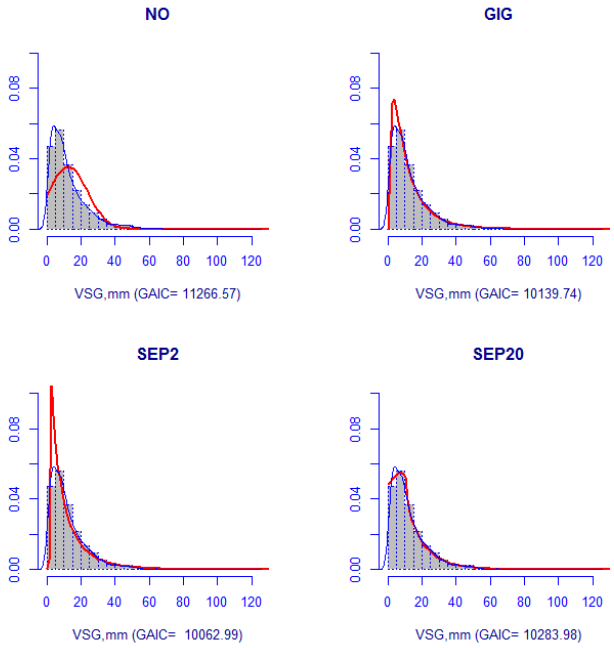

Figura 4.22: Distribucións escollidas para axustar a variable vsg.

Para continuar coa análise, realizaremos regresións simples con cada unha das variables de interese, para posteriormente incluír todas estas nunha regresión múltiple, onde compararemos as significacións e os resultados obtidos coas regresións simples.

Previamente, dado que dispoñemos dun número de casos considerable, decidimos dividir a mostra en dous e utilizar a primeira parte para axustar o modelo e a segunda para comprobar a validez deste (isto só se fará para o modelo de regresión múltiple).

Na primeira parte (mostra de ensaio) están contidos o 70 % dos casos, polo tanto temos un tamaño de 1044 persoas das cales 579 (56 %) son mulleres e 465 (44 %) son homes. Á categoría de actividade física baixa pertencen 399 (38 %), á moderada 383 (37 %) e á alta 262 (25 %). Hai 572 (55 %) non fumadores, 272 (26 %) fumadores e 200 (19 %) fumadores. 373 (36 %) persoas apenas beben, 418 (40 %) fano moderadamente, 160 (15 %) beben moito e 93 (9 %) excesivamente. Ademais, 208 (20 %) teñen síndrome metabólico e 836 (80 %) non. A idade media é de 52.36 tendo a persoa máis nova 18 anos e a máis vella 91. Para o índice de masa corporal a media é de 28.30, sendo o valor máis baixo 17.36 e o máis alto 50.70. Polo tanto, este 70 % ten as mesmas características que a mostra total.

Para todas as regresión simples obtívose significación na variable explicativa involucrada e, cando esta se trata dunha variable categórica, a significación deuse en todas e cada unha das categorías. A relación entre a VSG e a idade e o índice de masa corporal é crecente mentres que coa actividade física e o consumo de alcohol e tabaco a relación é inversa. En relación ó sexo, o valor nas mulleres é significativamente maior que nos homes e tamén naquelas persoas que teñen síndrome metabólico fronte ás que non o teñen.

Os resultados obtidos tras realizar a regresión múltiple modifícanse. O consumo de tabaco, o síndrome metabólico e algunha das categorías do consumo de alcohol perden a significación. Esta perda de significación posiblemente ten que ver coa relación existente entre as variables. Cando estas se estudan de 1 en 1 pode ser que todas aporten información, pero no momento no que involucramos máis variables no modelo algunha delas pode perder importancia debido a que a información que aportaba por separado xa non é necesaria se forman parte do modelo outras variables relacionadas con ela.

O resto das variables seguen a ser significativas e manteñen a mesma relación coa variable resposta

que na regresión simple.

A Figura 4.23 móstranos o term.plot() para os modelos de regresión simple mentres que a Figura 4.24 nos mostra o term.plot() para o modelo de regresión múltiple. Nelas pódense ver as relacións descritas nos parágrafos anteriores.

Para a finalidade clínica o análise remata co resultado desta regresión múltiple polo interese de cada unha das variables que interveñen en dito modelo. Pero seguiremos traballando sobre este modelo dende un punto de vista máis matemático, buscando simplificalo.

Utilizamos entón a función stepTGD() para levar a cabo dita selección. A indicación desta función é que eliminemos as variables bmi, oh4 e ipq. Se o facemos as variables que aínda quedan no modelo (age, sex, tab012 e sm), excepto o tabaco, son todas significativas. O síndrome metabólico recupera a significación obténdose valores máis altos para aquelas persoas con síndrome metabólico como ocorría na regresión simple.

Verifiquemos coas funcións wp() e plot() que realmente se trata dun modelo axeitado.

Na Figura 4.25, podemos ver que ambas saídas corresponden a un modelo adecuado.

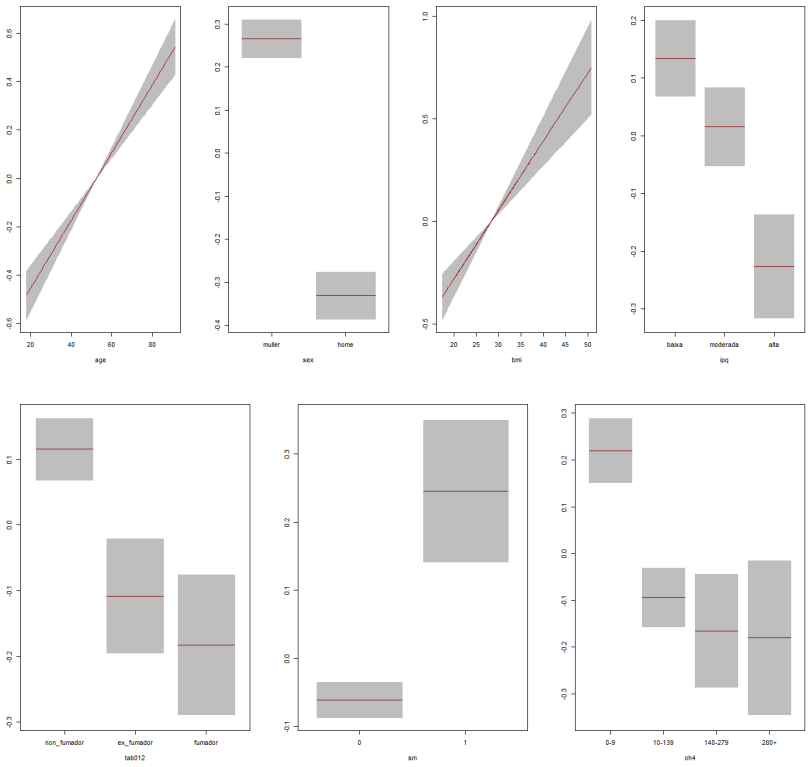

Figura 4.23: temp.plot() para os modelos de regresión simple.

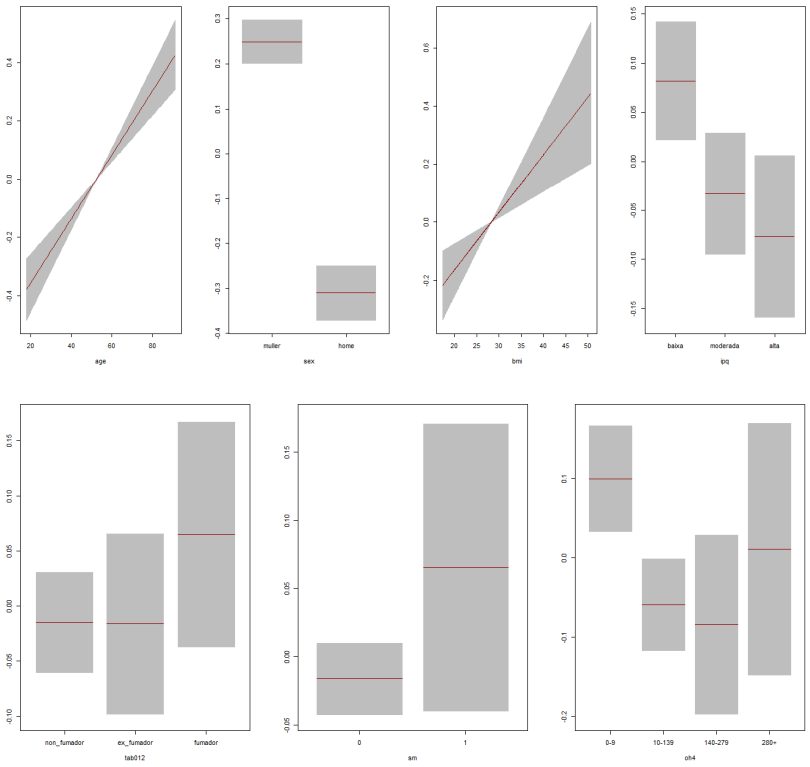

Figura 4.24: temp.plot() para o modelo de regresión múltiple.

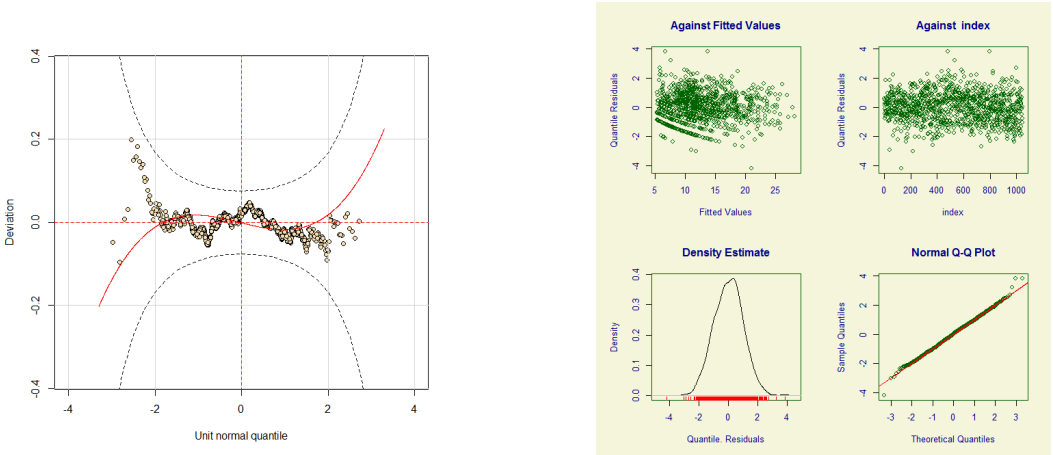

Figura 4.25: Gráficos de validación para o modelo axustado para a VSG.

4.3. Modelos mixtos con GAMLSS

No primeiro capítulo falouse brevemente dos modelos mixtos e da súa gran importancia. Ó igual que ocorre cando traballamos cos modelos lineais xeneralizados (GLM) ou cos modelos aditivos xene-

ralizados (GAM), cando se traballa cos modelos aditivos xeneralizados de localización, escala e forma (GAMLSS) tamén podemos falar de modelos que combinan modelos mixtos e GAMLSS. O caso que se presenta nesta sección é exemplo disto.

A base de datos coa que se traballa neste caso está composta por 40 persoas. Estes participantes son persoas sans elixidas ó azar entre pacientes/acompañantes que acudían ás consultas de otorrino e non tiñan antecedentes de vertixe. Foron escollidos 5 homes e 5 mulleres dentro das seguintes franxas de idade: [0, 20], [21, 40], [41, 60], [61, 80]. Para cada un destes suxeitos están recollidos, entre outros datos, o sexo, a idade e as ganancias obtidas para cada unha das canles na proba vHIT:

- h.1: canle horizontal do oído dereito.
- h.2: canle horizontal do oído esquerdo.
- p.1: canle posterior do oído dereito.
- p.2: canle posterior do oído esquerdo.
- a.1: canle anterior do oído dereito.
- a.2: canle anterior do oído esquerdo.

O vHIT ou proba de impulso cefálico cuantificada, permite realizar a proba clínica do impulso cefálico mediante unha micro cámara de vídeo de alta velocidade e coa axuda de xiroscopios integrados e un sistema de rexistro comparar a velocidade do movemento ocular co da cabeza. Isto permite cuantificar a relación entre ambos movementos detectando diferenzas anormais e, ademais, rexistrar movementos oculares anómalos que aparecen en pacientes con lesións vestibulares.

Suponse que os valores normais son aqueles que están próximos a 1. Non obstante, ata agora considéranse anormais só aqueles valores menores de 0.8. Un dos obxectivos que se pretende na tese na que irá incluído este estudo é buscar os valores de referencia para a proba do vHIT, tentando ver se a idade e o sexo inflúen nestes valores ou se é suficiente ter uns únicos valores de referencia sen ter en conta ningún destes dous factores.

Este estudo levouse a cabo entre febreiro de 2013 e xuño de 2015 (ambos incluídos) na Unidade de Otoneuroloxía do Servizo de Otorrinolaringoloxía do Complexo Hospitalario de Santiago de Compostela (CHUS).

Os valores medios, xunto co máximo e o mínimo obtido para cada unha das 6 canles están recollidos no Cadro 4.6.

Canle	**Mínimo**	**Media**	**Máximo**
h.1	0.840	1.078	1.580
h.2	0.7600	0.9778	1.3200
p.1	0.630	1.044	1.630
a.2	0.6300	0.9962	1.4800
a.1	0.6700	0.9828	1.3600
p.2	0.7800	1.0785	1.4600

Cadro 4.6: Medidas resumo para as 6 canles.

Nel vese reflectido como as medias das ganancias das 6 canles están próximas a 1, isto concorda co que se espera clinicamente.

Polo tanto para levar a cabo o obxectivo proposto debemos buscar a distribución que máis se adapte ós datos recollidos, ó igual que ocorría nos exemplos anteriores, para así poder construír ditos valores

de referencia, determinando a necesidade ou non de separar estes valores tendo en conta cada unha das canles e dentro destas considerando as covariables idade e sexo.

O estudo levouse a cabo tendo en conta, en primeiro lugar, cada unha das 6 canles por separado pero, posteriormente, considerouse a posibilidade de agrupalas por pares. Atendendo a esta intención, o primeiro que se fixo foi estudar a correlación entre cada unha das 6 canles. A correlación existente entre os valores obtidos para cada unha das canles é a que se pode ver no Cadro 4.7.

Canle	h.1	h.2	p.1	a.2	a.1	p.2
h.1	1.00000000	0.7174014	0.05058330	0.30858352	0.08996576	0.09903489
h.2	0.71740139	1.0000000	-0.12311639	0.15131441	0.24936639	0.14468963
p.1	0.05058330	-0.1231164	1.00000000	0.35608652	-0.08752669	-0.07438088
a.2	0.30858352	0.1513144	0.35608652	1.00000000	0.03188292	-0.29725190
a.1	0.08996576	0.2493664	-0.08752669	0.03188292	1.00000000	0.14661513
p.2	0.09903489	0.1446896	-0.07438088	-0.29725190	0.14661513	1.00000000

Cadro 4.7: Correlación entre as distintas canles.

Esta correlación tamén se pode ver graficamente na Figura 4.26. Á vista destes resultados non parece demasiado evidente que se poida facer a agrupación que se pretende, pero viuse no transcurso da análise que realmente si que se poden agrupar as canles por pares, é dicir, podemos xuntar h.1 con h.2, p.1 con p.2 e a.1 con a.2.

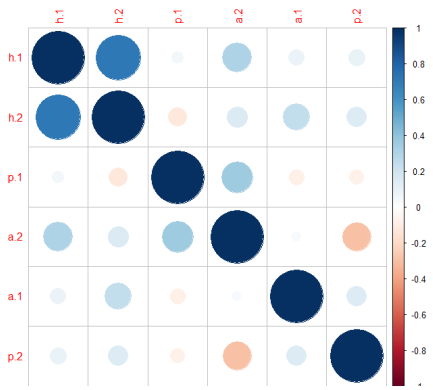

Figura 4.26: Correlación entre as distintas canles.

Polo tanto omitirase a primeira parte do estudo onde se levou a cabo a análise por separado de cada unha das 6 canles e presentarase o estudo realizado para os pares de canles.

Na parte omitida obtívose que tanto a canle horizontal esquerda como a dereita podían ser axustadas por unha distribución LOGNO. Ademais, as canles posteriores (esquerda e dereita) e as anteriores (esquerda e dereita) admiten a distribución NO como unha distribución axeitada. Dados estes resultados decidiuse traballar coas canles por parellas e ver se estas distribucións escollidas por separado seguían a ser válidas para os pares de canles. Veremos, a continuación, que estas distribucións sí resultaron axeitadas.

Debemos ter en conta que para levar a cabo esta parte da análise temos que utilizar modelos mixtos, xa que ó considerar ambos oídos á vez estamos tendo dous datos pertencentes ó mesmo individuo.

As distribucións obtidas para o axuste da canle horizontal, posterior e anterior a partir da función fitDist() son as que se presentan na Figura 4.27.

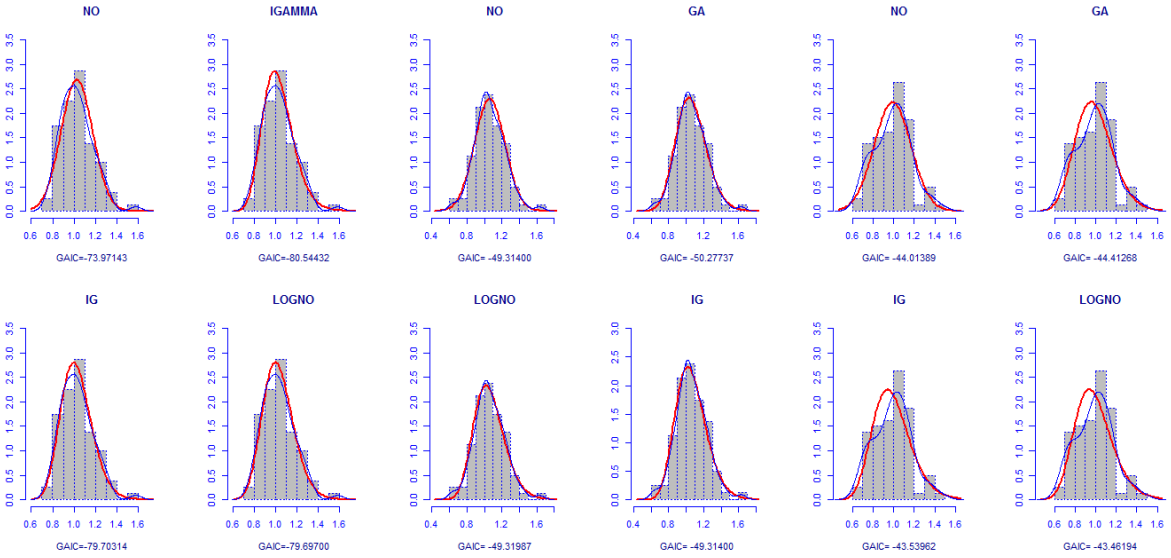

Figura 4.27: Distribucións escollidas pola función fitDist() para o axuste dos datos obtidos para a canle horizontal, anterior e posterior de esquerda a dereita (xunto coa distribución normal utilizada de referencia).

- Canle horizontal:

  Atendendo ao valor do GAIC, non deberíamos tomar como primeira opción a distribución LOGNO, pero dado que a diferencia entre o valor do GAIC da distribución LOGNO e o valor do GAIC da distribución IGAMMA (que sería a de menor GAIC) é menor que 2, podemos optar por utilizar esta distribución que era a que nos servía tanto para a canle horizontal esquerda como para a canle horizontal dereita.

- Canle posterior:

  De novo atendendo ao valor de GAIC, non deberíamos tomar como primeira opción a distribución NO pero, ao igual que antes, a diferencia no valor do GAIC entre a distribución NO e a distribución GAMMA (que sería a de menor GAIC) nin sequera é de 1 punto.

- Canle anterior:

  Para a canle anterior obtivemos uns resultados similares aos da canle posterior, polo tanto aceptaremos novamente a distribución normal como unha distribución válida para o axuste dos datos.

Para comprobar que realmente podemos aceptar as distribucións mencionadas como distribucións axeitadas, realizaremos unha tarefa de diagnose coa función wp(). Os resultados obtidos con esta función son os que se poden ver na Figura 4.28. Dado que os puntos caen dentro do intervalo formado polas liñas negras discontinuas, podemos asegurar que se escolleu unha distribución axeitada. Os parámetros para as distribucións escollidas son os que se poden ver no Cadro 4.8.

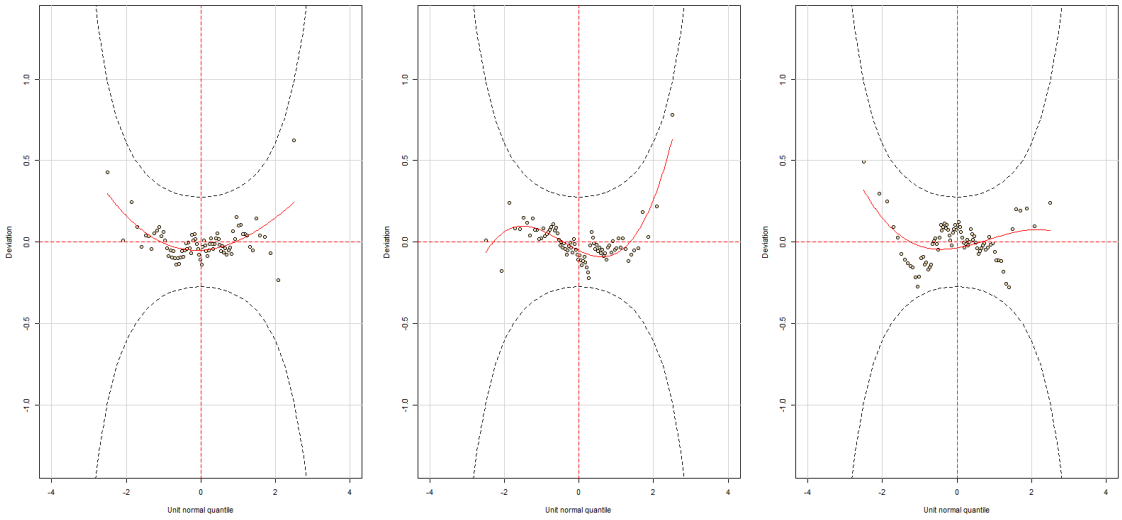

Figura 4.28: Validación das distribucións escollidas para as distintas canles (de esquerda a dereita: canle horizontal, canle anterior, canle posterior).

Canle	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$ (log)
Horizontal	0.01730	-1.95935
Posterior	1.06150	-1.75215
Anterior	0.9895	-1.71903

Cadro 4.8: Parámetros para as distribucións escollidas para as distintas canles.

Unha vez escollida a distribución para cada unha das canles, estamos en disposición de calcular os valores de referencia. Para levar a cabo esta tarefa botaremos man das funcións centiles() e centiles.pred(), a primeira delas para obter resultados gráficos e a segunda numéricos. Os percentís de interese son o 2.5, 5 e 10 %, polo tanto será neses nos que fagamos máis fincapé.

Nesta parte da análise axustaremos cada par de canles por idade e sexo, dado que os GAMLSS nos permiten axustar todos os parámetros da distribución utilizada (nesta caso traballaremos coas distribucións log-normal e normal, polo tanto traballaremos unicamente con dous parámetros) a partir das covariables; axustaremos media e varianza por idade e sexo. Tras dito axuste consideraremos se é necesario dar valores de referencia con respecto á idade e ó sexo (ou só dalgún deles) ou se pola contra ditos valores non dependen de ningún destes dous factores.

Nos tres casos comezaremos axustando o modelo que consideramos máis complexo:

$$y \sim D\{\mu = pb(age) + sex, log(\sigma) = pb(age) + sex\}$$

onde D denota a distribución NO ou LOGNO dependendo da canle coa que esteamos a traballar.

Trátase do modelo onde, ademais de axustar tanto media como varianza por idade e sexo, suavizamos a idade en ambos lados.

Unha vez axustado este modelo botaremos man da función term.plot() para ver se a suavización feita é necesaria. Posteriormente utilizaremos a función drop1() para ver que variables do modelo son prescindibles. Unha vez feito isto, se o modelo final depende do sexo, faremos os valores de referencia

para ambos sexos, se ese non é o caso só daremos uns valores de referencia en conxunto. Vexamos o que ocorre en cada par de canles:

- Canle horizontal:

  Tras axustar o modelo

  $$y \sim LOGNO\{\mu = pb(age) + sex, log(\sigma) = pb(age) + sex\}$$

  o term.plot() obtido móstranos que a relación entre σ e a idade é lineal; non está tan claro para a relación entre o parámetro μ e a idade. Debido a isto eliminaremos a suavización da variable idade cando esta se utiliza para axustar σ, pero non cando se utiliza para axustar μ.

  Tras aplicarlle a función drop1() a ambos parámetros obtívose unha saída que nos indicaba que podemos prescindir da idade á hora de axustar o parámetro de escala.

  Polo tanto, axustamos varios modelos a partir destas indicacións e o de menor GAIC foi $y \sim LOGNO\{\mu = pb(age) + sex, log(\sigma) = sex\}$, para o cal tanto a idade como o sexo resultan ser significativas para realizar o axuste do parámetro μ e o sexo tamén o é para o axuste do parámetro σ.

  Dada a importancia do sexo para ambos parámetros decidiuse dar os valores de referencia atendendo ó sexo. Pero antes debemos comprobar que o modelo axustado é adecuado. Para isto volvemos utilizar a función wp(), Figura 4.32.

  Á vista da Figura 4.32, podemos considerar o noso modelo un modelo adecuado. Polo tanto realizamos a representación dos valores de referencia coa axuda da función centiles() (separando por sexo), Figura 4.29.

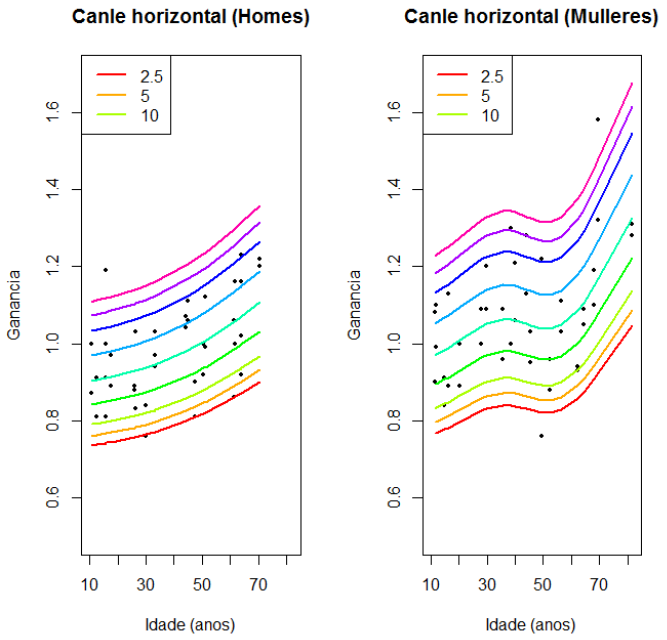

Figura 4.29: Gráfico para os valores de referencia da ganancia da canle horizontal.

Como se pode comprobar na Figura 4.29, a forma que adoptan as curvas son distintas entre homes e mulleres. Ademais destas gráficas, utilizouse a función centiles.pred() para obter o Cadro 4.9, onde se mostran os valores de referencia para os percentís de interese (2.5, 5, 10 %).

Sexo	**Homes**			**Mulleres**		
Idade	**Percentil 2.5**	**Percentil 5**	**Percentil 10**	**Percentil 2.5**	**Percentil 5**	**Percentil 10**
10	0.7351751	0.7598063	0.7892305	0.7614251	0.7908146	0.8261095
15	0.7412410	0.7660755	0.7957424	0.7773711	0.8073762	0.8434102
20	0.7475992	0.7726466	0.8025681	0.7956177	0.8263270	0.8632068
25	0.7548050	0.7800939	0.8103037	0.8146684	0.8461130	0.8838759
30	0.7635217	0.7891027	0.8196613	0.8305939	0.8626532	0.9011543
35	0.7741664	0.8001039	0.8310887	0.8386563	0.8710268	0.9099016
40	0.7866831	0.8130400	0.8445257	0.8369437	0.8692481	0.9080436
45	0.8009413	0.8277760	0.8598323	0.8283840	0.8603581	0.8987567
50	0.8169796	0.8443516	0.8770499	0.8213071	0.8530080	0.8910786
55	0.8349948	0.8629704	0.8963896	0.8247961	0.8566317	0.8948640
60	0.8550698	0.8837179	0.9179407	0.8445601	0.8771585	0.9163069
65	0.8770430	0.9064273	0.9415295	0.8806301	0.9146207	0.9554411
70	0.9004550	0.9306237	0.9666629	0.9271286	0.9629139	1.0058898
75	0.9248599	0.9558463	0.9928622	0.9778109	1.0155525	1.0608776
80	0.9499803	0.9818083	1.0198297	1.0293355	1.0690659	1.1167794

Cadro 4.9: Valores de referencia para os percentís 2.5, 5, 10 % para a canle horizontal.

Á vista dos resultados obtidos no Cadro 4.9 e na Figura 4.29, é claro que para os homes a medida que a idade aumenta os valores de referencia tamén o fan (aínda que dun xeito non lineal). Isto indícanos que canto maior sexa o home que entre na consulta menor marxe debemos deixar por debaixo de 1 para consideralo un valor anómalo. Se nos fixamos no percentíl 5 dun neno de 10 anos, o valor obtido é sobre 0.76 mentres que se nos fixamos no mesmo percentíl para un home de 75 anos é sobre 0.96.

Para as mulleres a progresión non é tan clara, os resultados obtidos indícannos que os valores de referencia comezan subindo a canda a idade, pero para a franxa de idade 40-60 os valores descenden para logo volver subir.

- Canle anterior:

  Tanto na canle anterior como na posterior procederemos de xeito análogo a como o fixemos para a canle horizontal, comentando os diferentes resultados que se van obtendo.

  Tras axustar o modelo

  $$y \sim NO\{\mu = pb(age) + sex, log(\sigma) = pb(age) + sex\}$$

  o term.plot() obtido indícanos de novo que a relación entre σ e a idade é lineal, mentres que para a relación entre o parámetro μ e a idade non está claro que se poda considerar lineal. Debido a isto eliminaremos a suavización da idade cando esta se utiliza para axustar σ, pero non cando se utiliza para axustar μ.

  Aplicamos a función drop1() a ambos parámetros. A saída que se obtivo indícanos que podemos prescindir do sexo á hora de axustar o parámetro de localización e tanto da idade como do sexo á hora de axustar o parámetro de escala.

Atendendo a isto, fomos eliminando estas variables do modelo, empezando polas que se refiren a σ e comprobando en todo momento o valor do GAIC xa que foi este criterio o que utilizamos para escoller o modelo final.

Neste caso o modelo de menor GAIC é $y \sim NO\{\mu = pb(age), log(\sigma) = 1\}$.

Dado que para esta canle o sexo non é relevante para ningún dos parámetros da distribución, consideramos dar uns únicos valores de referencia atendendo soamente á idade, que sí resulta significativa para o axuste do parámetro de localización. Ó igual que antes, previamente ó calculo destes valores de referencia, levaremos a cabo a validación do modelo escollido utilizando a función wp(), Figura 4.32.

Á vista da Figura 4.32, corroboramos que o modelo que estamos a considerar é un modelo adecuado. Polo tanto realizamos a representación dos valores de referencia, Figura 4.30.

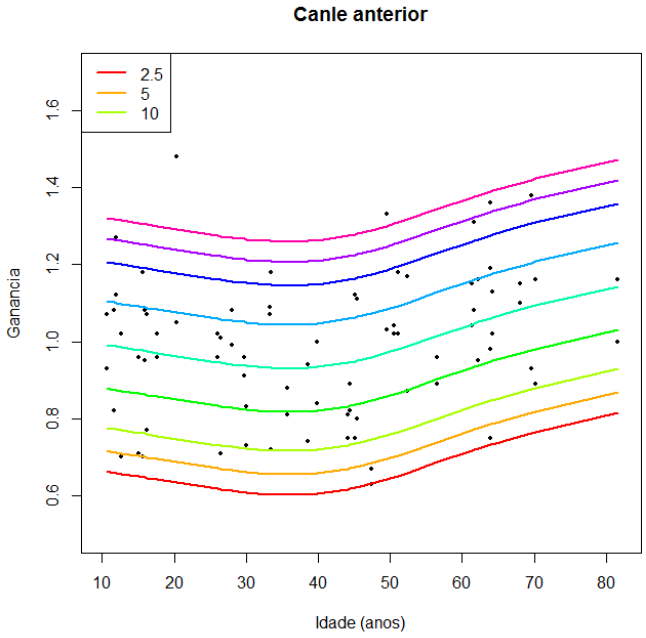

Figura 4.30: Gráfico para os valores de referencia da ganancia da canle anterior.

Ademais desta gráfica, obtívose o Cadro 4.10 onde se mostran os valores de referencia para os percentís de interese (2.5, 5, 10 %).

Idade	Percentil 2.5	Percentil 5	Percentil 10
10	0.6645827	0.7174106	0.7783177
15	0.6495894	0.7024173	0.7633245
20	0.6346522	0.6874801	0.7483873
25	0.6203291	0.6731570	0.7340642
30	0.6083948	0.6612227	0.7221299
35	0.6023603	0.6551882	0.7160953
40	0.6056883	0.6585162	0.7194233
45	0.6203221	0.6731500	0.7340572
50	0.6453393	0.6981672	0.7590744
55	0.6763610	0.7291888	0.7900960
60	0.7084990	0.7613269	0.8222341
65	0.7383496	0.7911775	0.8520847
70	0.7645867	0.8174146	0.8783217
75	0.7874834	0.8403112	0.9012184
80	0.8083624	0.8611903	0.9220975

Cadro 4.10: Valores de referencia para os percentís 2.5, 5, 10 % para a canle anterior.

Neste caso, a diferencia do que ocorría na canle horizontal, non é necesario facer a separación por sexos. Non obstante, a idade resulta significativa. Atendendo á forma que presenta a gráfica da Figura 4.30 e ós resultados do Cadro 4.10 pódese dicir que os valores de referencia baixan lixeiramente no primeiro tramo, ata os 35 anos máis ou menos, para logo crecer.

- Canle posterior:

  Tras axustar o modelo

  $$y \sim NO\{\mu = pb(age) + sex, log(\sigma) = pb(age) + sex\}$$

  o term.plot() obtido para esta última canle indícanos que tanto a relación entre σ e a idade, como a relación ente μ e a idade son lineais. Polo tanto eliminaremos a suavización da variable idade no axuste de ambos parámetros.

  A saída obtida coa función drop1() indícanos que podemos prescindir tanto da idade como do sexo á hora de axustar o parámetro de localización e da idade á hora de axustar o parámetro de escala.

  Atendendo a isto, fomos eliminando estas variables do modelo ata quedarnos co modelo de menor AIC.

  Neste caso o modelo de menor GAIC é $y \sim NO\{\mu = 1, log(\sigma) = sexo\}$.

  Dada a significación da variable sexo para o parámetro de escala decidiuse dar os valores de referencia atendendo ó sexo. Pero antes debemos comprobar que o modelo axustado é adecuado. Para isto volvemos utilizar a función wp(), Figura 4.32.

  Á vista da Figura 4.32, podemos considerar o noso modelo un modelo adecuado. Polo tanto realizamos a representación dos valores de referencia (separando por sexo), Figura 4.31.

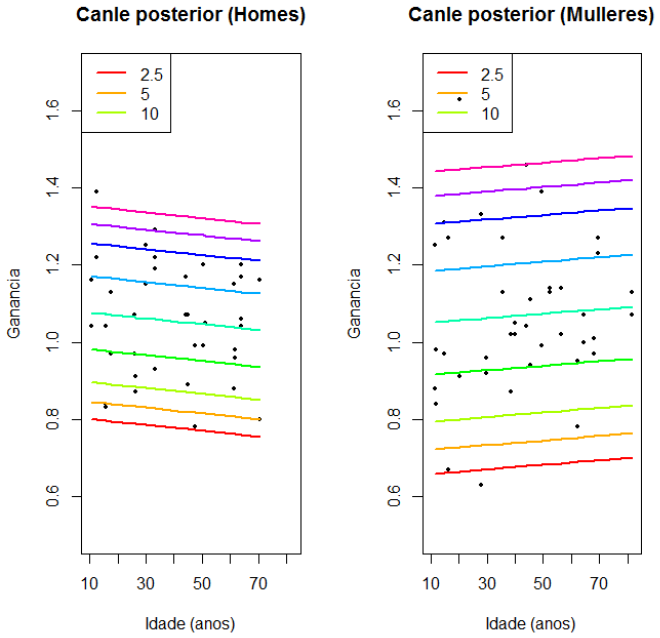

Figura 4.31: Gráfico para os valores de referencia da ganancia da canle posterior.

Como se pode comprobar na Figura 4.31, as gráficas obtidas son distintas entre homes e mulleres. Ademais destas gráficas, obtívose o Cadro 4.11 onde se mostran os valores de referencia para os percentís de interese (2.5, 5, 10 %).

Para as canles posteriores os resultados obtidos no Cadro 4.11 e na gráfica da Figura 4.31 indícannos que a relación entre a idade e os valores de referencia son lineais, tanto para homes como para mulleres. Pero o comportamento entre ambos sexos é diferente, mentres que para os homes os valores de referencia diminúen coa idade, para as mulleres aumentan. Ademais, o rango de valores das mulleres é considerablemente maior que o dos homes, de aí a significación da idade con respecto ó parámetro de escala.

Sexo	Homes			Mulleres		
Idade	**Percentil 2.5**	**Percentil 5**	**Percentil 10**	**Percentil 2.5**	**Percentil 5**	**Percentil 10**
10	0.8007945	0.8450681	0.8961128	0.6581014	0.7210756	0.7936807
15	0.7970221	0.8412958	0.8923405	0.6610600	0.7240341	0.7966393
20	0.7932501	0.8375237	0.8885684	0.6640185	0.7269926	0.7995977
25	0.7894784	0.8337521	0.8847968	0.6669768	0.7299509	0.8025560
30	0.7857070	0.8299807	0.8810254	0.6699349	0.7329090	0.8055141
35	0.7819355	0.8262091	0.8772538	0.6728928	0.7358669	0.8084721
40	0.7781637	0.8224373	0.8734820	0.6758505	0.7388247	0.8114298
45	0.7743917	0.8186654	0.8697101	0.6788079	0.7417820	0.8143872
50	0.7706198	0.8148935	0.8659382	0.6817649	0.7447390	0.8173442
55	0.7668481	0.8111217	0.8621664	0.6847215	0.7476956	0.8203008
60	0.7630764	0.8073500	0.8583947	0.6876780	0.7506521	0.8232572
65	0.7593047	0.8035783	0.8546230	0.6906345	0.7536086	0.8262137
70	0.7555329	0.7998065	0.8508512	0.6935911	0.7565652	0.8291703
75	0.7517610	0.7960347	0.8470794	0.6965477	0.7595219	0.8321270
80	0.7479892	0.7922628	0.8433075	0.6995045	0.7624786	0.8350837

Cadro 4.11: Valores de referencia para os percentís 2.5, 5, 10 % para a canle posterior.

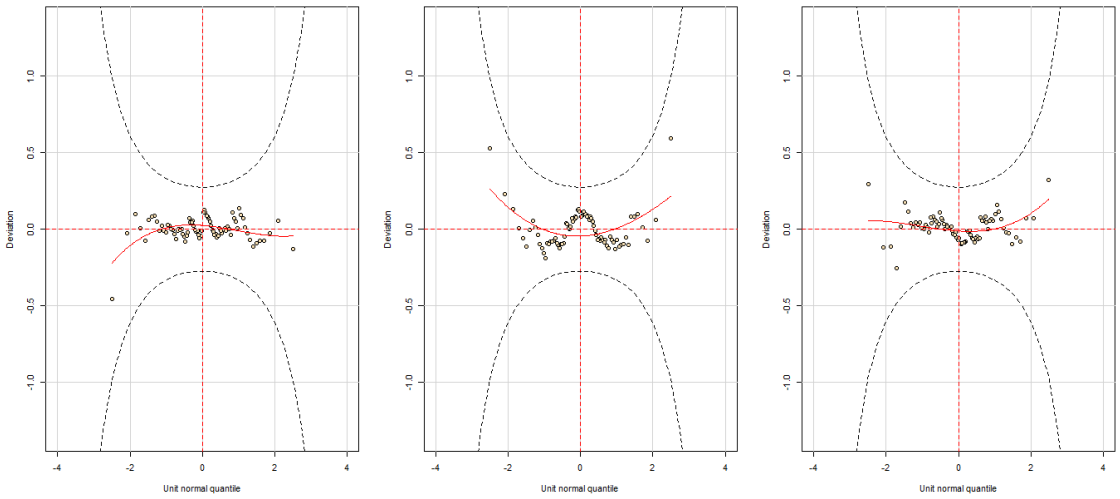

Figura 4.32: Gráficos de validación para os modelos axustados para os valores da ganancia das canles horizontal, anterior e posterior, de esquerda a dereita respectivamente.

Os resultados obtidos para cada par de canles foron moi diferentes, para os 3 casos a idade resultou significativa pero o sexo só o é para as canles horizontais e posteriores. Ademais só as relacións das canles posteriores resultaron lineais.

Con respecto ós valores obtidos, cabe destacar que os das canles anterior e posterior poden ser realistas. Non obstante, os das canles horizontais non tanto xa que se toman valores superiores a 1 en varios dos percentís e debemos ter en conta que os percentís mostrados foron os 2.5, 5 e 10 %. Este feito pode ter que ver co tamaño da mostra, tan só se dispoñía de 40 casos en total, 80 ó estudar ambas canles á vez, pero de novo 40 cando se separou por sexo. Polo tanto sería recomendable recompilar máis casos e realizar de novo á análise levada a cabo.

4.4. Conclusións

Os exemplos presentados ó longo do capítulo forman parte de traballos clínicos, como tales todos eles teñen obxectivos clínicos. Neste caso eses obxectivos poden reducirse a dous:

- Dar valores de referencia
- Buscar a relación existente entre dúas variables

Dado que os resultados obtidos debían ter sentido clínico á hora de dar valores de referencia traballouse soamente coas variables referentes á idade e ó sexo xa que son aquelas máis comúns e das que é fácil dispoñer cando un paciente chega á consulta.

Ademais utilizáronse modelos que, dentro de ser válidos, fosen sinxelos de interpretar. Por esta razón nunca se incluíron variables explicativas para modelar a asimetría e a curtose e só se utilizou a idade e o sexo para modelar a varianza. Esta decisión non foi só por razóns clínicas senón tamén matemáticas; a utilización dun modelo complexo producía problemas de converxencia ou un tempo de execución considerable.

Se nos centramos na parte dos exemplos onde o obxectivo era ver a relación entre variable resposta e variable explicativa, o obxectivo clínico que requirían os profesionais era simplemente ver os resultados obtidos para as variables de interese. Non obstante, tanto por mostrar técnicas matemáticas como por considerarse de interese clínico para futuros estudios, realizouse unha selección do modelo para simplificalo e quedarnos con aquelas variables que si influían na variable resposta.

Apéndice A

Familias en GAMLSS

Distribución	**Nome en R**	μ	σ	ν	τ
Beta	BE()	logit	logit	-	-
Box-Cox Cole e Green	BCCG()	identity	log	identity	-
Box-Cox power exponential	BCPE()	identity	log	identity	log
Box-Cox-t	BCT()	identity	log	identity	log
Exponencial	EXP()	log	-	-	-
Expoñencial Gaussiana	exGAUS()	identity	log	log	-
Exonencial xeneralizada beta tipo 2	EGB2()	identity	identity	log	log
Gamma	GA()	log	log	-	-
Beta xeneralizada tipo 1	GB1()	logit	logit	log	log
Beta xeneralizada tipo 2	GB2()	log	identity	log	log
Gamma xeneralizada	GG()	log	log	identity	-
Gaussiana inversa xeneralizada	GIG()	log	log	identity	-
t xeneralizada	GT()	identity	log	log	log
Gumbel	GU()	identity	log	-	-
Gamma inversa	IGAMMA()	log	log	-	-
Gaussiana Inversa	IG()	log	log	-	-
Johnson's SU reparametrizada	JSU()	identity	log	identity	log
Jonhson's orixinal SU	JSUo()	identity	log	identity	log
Loxística	LO()	identity	log	-	-
Logit Normal	LOGITNO()	logit	log	-	-
Log normal	LOGNO()	identity	log	-	-
Log normal 2	LOGNO2()	log	log	-	-
Log normal (Box-Cox)	LNO()	identity	log	fixado	-
Normal Exponencial t	NET()	identity	log	fixado	fixado
Normal	NO()	identity	log	-	-

Normal reparametrizada	NO2()	identity	log	-	-
Familia normal	NOF()	identity	log	identity	-
Pareto 2 orixinal	PARETO2o()	log	log	-	-
Pareto 2	PARETO2()	log	log	-	-
Pareto 2 reparametrizada	GP()	log	log	-	-
Power exponencial	PE()	identity	log	log	-
Reverse Gumbel	RG()	identity	log	-	-
Reverse xeneralizada extrema	RGE()	identity	log	log	-
Skew normal tipo 1	SN1()	identity	log	identity	-
Skew normal tipo 2	SN2()	identity	log	identity	-
Skew Power exponential	SEP()	identity	log	identity	log
Skew power exponential tipo 1	SEP1()	identity	log	identity	log
Skew power exponential tipo 2	SEP2()	identity	log	identity	log
Skew power exponential tipo 3	SEP3()	identity	log	log	log
Skew power exponential tipo 4	SEP4()	identity	log	log	log
Sinh-Arcsinh orixinal	SHASHo()	identity	log	identity	log
Sinh-Arcsinh orixinal 2	SHASHo2()	identity	log	identity	log
Sinh-Arcsinh	SHASH()	identity	log	log	log
Skew t tipo 1	ST1()	identity	log	identity	log
Skew t tipo 2	ST2()	identity	log	identity	log
Skew t tipo 3	ST3()	identity	log	log	log
Skew t tipo 3	ST3C()	identity	log	log	log
Skew t tipo 3 reparametrizado	SST()	identity	log	log	logshifto2
Skew t tipo 4	ST4()	identity	log	log	log
Skew t tipo 5	ST5()	identity	log	identity	log
Familia t	TF()	identity	log	log	-
Familia t reparametrizada	TF2()	identity	log	logshifto2	-
Weibull	WEI()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-
Weibull (μ a media)	WEI3()	log	log	-	-
Multinomial (3 categorías)	MN3	log	log	-	-
Multinomial (4 categorías)	MN4	log	log	log	-
Multinomial (5 categorías)	MN5	log	log	log	log

Cadro A.1: Distribucións continuas dispoñibles para os GAMLSS (xunto coas súas funcións link).

Distribución	**Nome en R**	μ	σ	ν
Beta binomial	BB()	logit	log	-
Binomial	BI()	logit	-	-
Xeométrica	GEOM()	log	-	-
Logarítmica	LG()	logit	-	-
Delaporte	DEL()	log	log	logit
Binomial negativa tipo I	NBI()	log	log	-
Binomial negativa tipo II	NBII()	log	log	-
Poisson	PO()	log	-	-
Dobre Poisson	DPO()	log	log	-
Gaussiana inversa Poisson	PIG()	log	log	-
Sichel	SI()	log	log	identity
Sichel(μ a media)	SICHEL()	log	log	identity
Waring (μ a media)	WARING()	log	log	-
Yule (μ a media)	YULE()	log	-	-
Beta binomial cero alterada	ZABB()	logit	log	logit
Binomial cero alterada	ZABI()	logit	logit	-
Logarítmica cero alterada	ZALG()	logit	logit	-
Binomial negativa cero alterada	ZANBI()	log	log	logit
Poisson cero alterada	ZAP()	log	logit	-
Beta binomial cero inflada	ZIBB()	logig	log	logit
Binomial cero inflada	ZIBI()	logig	logit	-
Binomial negativa cero inflada	ZINBI()	log	log	logit
Poisson cero inflada	ZIP()	log	logit	-
Poisson cero inflada (μ a media)	ZIP2()	log	logit	-
Poisson Gaussiana inversa cero inflada	ZIPIG()	log	log	logit

Cadro A.2: Distribucións discretas dispoñibles para os GAMLSS (xunto coas súas funcións link).

Distribución	**Nome en R**	μ	σ	ν	τ
Beta inflada (en 0)	BEOI()	logit	log	logit	-
Beta inflada (en 0)	BEINF0()	logit	logit	log	-
Beta inflada (en 1)	BEZI()	logit	log	logit	-
Beta inflada (en 1)	BEINF1()	logit	logit	log	-
Beta inflada (en 0 e 1)	BEINF()	logit	logit	log	log
Gamma cero axustada	ZAGA()	log	log	logit	-

Gausiana inversa cero axustada	ZAIG()	log	log	logit	-

Cadro A.3: Mesturas de distribucións dispoñibles para os GAMLSS (xunto coas súas funcións link).

NOTA: Posiblemente non estean recollidas todas as distribucións existentes hoxe en día. Debido á posibilidade de implementar novas distribucións estes cadros poden seguir medrando. Intentouse realizar unha recompilación utilizando os paquetes gamlss.family e gamlss.dist e tamén a páxina web *www.gamlss.org* onde se pode consultar un blog que vai recollendo a construción de novas familias así como as modificacións realizadas no entorno dos modelos GAMLSS.

A función de densidade e unha explicación un pouco máis detallada da maioría destas distribucións pode atoparse, por exemplo, en Stasinopoulos et al. (2008).

Bibliografía

[1] Akantziliotou C, Rigby RA, Stasinopoulos DM (2002) The R Implementation of Generalized Additive Models for Location, Scale and Shape. En M Stasinopoulos, G Touloumi (ed) Statistical Modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling, Chania, pp. 75-83.

[2] Bishop CH et al. (1995) Neural networks for pattern recognition.

[3] van Buuren S, Fredriks M (2001) Worm Plot: A Simple Diagnostic Device for Modelling Growth Reference Curves. Statistics in Medicine 20:1259-1277.

[4] Cole TJ, Green PJ (1992). Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood. Statistics in Medicine 11: 1305-1319.

[5] Cole TJ, Stanojevic S, Stocks J, Coates A, Hankinson JL, Wade AM (2009) Age and size related reference ranges: A case study of spirometry through childhood and adulthood. Statistics in Medicine 28(5):880-898.

[6] De Bastiani F, Stasinopoulos M, Rigby B (2016) gamlss.spatial: Spatial Terms in GAMLSS Models. R package version 1.3.1. https://cran.r-project.org/web/packages/gamlss.spatial/gamlss.spatial.pdf. Accedido 15 de marzo de 2016.

[7] Dunn PK, Smyth GK (1996) Randomised Quantile Residuals. Journal of Computational and Graphical Statistics 5: 236-244.

[8] Hastie TJ, Tibshirani RJ (1990). Generalized Additive Models. Chapman and Hall, London.

[9] Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. Journal of the Royal Statistical Society A 135:370-384.

[10] Rigby RA, Stasinopoulos DM (1996a). A Semi-parametric Additive Model for Variance Heterogeneity. Statistal Computing 6: 57-65.

[11] Rigby RA, Stasinopoulos DM (1996b) Mean and Dispersion Additive Models. En: W Härdle, MG Schimek (ed) Statistical Theory and Computational Aspects of Smoothing, Physica, pp. 215-230.

[12] Rigby RA, Stasinopoulos DM (2001) The GAMLSS project: a Flexible Approach to Statistical Modelling. En B Klein, L Korsholm (ed) New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling, Odense, pp. 249-256.

[13] Rigby RA, Stasinopoulos DM (2005) Generalized Additive Models for Location, Scale and Shape. Applied Statistics 54:507-554.

[14] Rigby RA, Stasinopoulos DM (2007) Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23(7):1-46.

[15] Royston P, Altman DG (1994) Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Applied Statistics 43: 429-467.

[16] Rúa Pérez C (2015) Comparación del análisis de datos funcionales con métodos tradicionales para analizar curvas de glucosa. Traballo Fin de Máster, Universidade de Vigo.

[17] Stasinopoulos M, Rigby B (2016a) gamlss.data: Data for GAMLSS models. R package version 4.3-4. https://cran.r-project.org/web/packages/gamlss.data/gamlss.data.pdf. Accedido 28 de xuño de 2016.

[18] Stasinopoulos M, Rigby B (2016b) gamlss.tr: Generating and Fitting Truncated "gamlss.family" Distributions. R package version 4.3-6. https://cran.r-project.org/web/packages/gamlss.tr/gamlss.tr.pdf. Accedido 28 de xuño de 2016.

[19] Stasinopoulos M, Rigby B (2016c) gamlss.mx: Fitting Mixture Distributions with GAMLSS. R package version 4.3-5. https://cran.r-project.org/web/packages/gamlss.mx/gamlss.mx.pdf. Accedido 28 de xuño de 2016.

[20] Stasinopoulos M, Rigby B, Akantziliotou C (2008) Instructions on how to use the gamlss package in R Second Edition. http://www.gamlss.org/wp-content/uploads/2013/01/gamlss-manual.pdf.

[21] Stasinopoulos M, Rigby B, Akantziliotou C, Heller G, Ospina R, Motpan N, McElduff F, Voudouris V, Djennad M, Enea M, Ghalanos A (2016a) gamlss.dist: Distributions to be Used for GAMLSS Modelling. R package version 4.3-6. https://cran.r-project.org/web/packages/gamlss.dist/gamlss.dist.pdf. Accedido 28 de xuño de 2016.

[22] Stasinopoulos M, Rigby B, Eilers P, Marx B, Pateras K, Kosidou L (2015a) gamlss.demo: Demos for GAMLSS. R package version 4.3-3. https://cran.r-project.org/web/packages/gamlss.demo/gamlss.demo.pdf. Accedido 28 de xuño de 2016.

[23] Stasinopoulos M, Rigby B, Lambert P (2015b) gamlss.nl: Fitting non linear parametric GAMLSS models. R package version 4.1-0. https://cran.r-project.org/web/packages/gamlss.nl/gamlss.nl.pdf. Accedido 28 de xuño de 2016.

[24] Stasinopoulos M, Rigby B, Mortan N (2016b) gamlss.cens: Fitting an Interval Response Variable Using "gamlss.family" Distributions. R package version 4.3-5. https://cran.r-project.org/web/packages/gamlss.cens/gamlss.cens.pdf. Accedido 28 de xuño de 2016.

[25] Stasinopoulos M, Rigby B, Voudouris V, Akantziliotou C, Enea M, Kiose D (2016c) gamlss: Generalised Additive Models for Location Scale and Shape. R package version 4.4-0. https://cran.r-project.org/web/packages/gamlss/gamlss.pdf. Accedido 28 de xuño de 2016.

[26] Stasinopoulos M, Rigby B, Voudouris V, Heller G, De Bastiani F (2015c) Flexible Regression and Smoothing The GAMLSS packages in R. http://www.gamlss.org/wp-content/uploads/2015/07/FlexibleRegressionAndSmoothingDraft-1.pdf.

[27] S.N. Wood. Generalized Additive Models. An introduction with R. Chapman and Hall, 2006.

[28] http://www.gamlss.org/