UNIVERSIDAD DE LA REPÚBLICA

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Informe de Pasantía

ELABORACIÓN DE CURVAS DE REFERENCIA NORMALES DE ESPIROMETRÍA EN NIÑOS URUGUAYOS MEDIANTE MODELOS GAMLSS

Pablo Palamarchuk

Tutores:

Ramón Álvarez-Vaz

Eugenia Riaño

Montevideo, Setiembre 2017.

UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba el trabajo de Pasantía:

ELABORACIÓN DE CURVAS DE REFERENCIA ESPIROMÉTRICAS NORMALES EN NIÑOS URUGUAYOS MEDIANTE MODELOS GAMLSS

Pablo Palamarchuck

Tutores:

Ramón Álvarez-Vaz

Eugenia Riaño

Licenciatura en Estadística

Puntaje .....

Tribunal

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Fecha.....

Índice general

Agradecimientos

Quisiera agradecer al Dr. Aníbal Capano por brindarme la oportunidad y depositar su confianza en quien escribe para la realización de este estudio como trabajo de pasantía.

Agradecer también a mis tutores, Ramón Álvarez-Vaz y Eugenia Riaño que me acompañaron durante el proceso y siempre dispuestos a brindar toda la ayuda que necesitaba.

A los profesores que he tenido durante la licenciatura, por brindar su tiempo y sus enseñanzas, que para bien o para mal, son parte de mi desarrollo como estudiante.

A los compañeros que he tenido durante estos años, en especial a Camila Cosentino y Leticia Colombo, con las cuales compartí muchas horas dentro y fuera de clases.

A mi familia por todo su apoyo incondicional y por soportar la cabeza en todo momento. A Fanky y Leia, quienes me acompañaron y vigilaron durante la redacción del trabajo.

A Google también!

Resumen

En este trabajo se construyen los modelos para obtener curvas de referencia espirométricas en niños uruguayos normales, utilizando los Modelos Aditivos Generalizados de Localización, Escala y Forma, para comparar los resultados con otros estudios internacionales.

La espirometría varía de acuerdo al tamaño de los pulmones, teniendo una relación directa con la estatura. Pero también varía de acuerdo a la etnia y los diferentes países. Por esta razón es necesario desarrollar valores estimados normales en una población de niños uruguayos para poder hacer una comparación dentro de las mismas condiciones ambientales, climatológicas y geográficas.

Los datos utilizados para este fin provienen de una muestra seleccionada a conveniencia de escuelas públicas y privadas del Uruguay por un grupo de investigadores del Centro Hospitalario Pereira Rossell en entre el año 1997 y 1999.

Los resultados obtenidos en este trabajo se comparan con otros, señalando similitudes y diferencias, tanto en metodología como en población muestral. Además, se presentan en modo de tablas y gráficas, los valores de referencia para las variables espirométricas de niñas y niños en relación a la estatura (talla).

Palabras claves: Ajuste de distribuciones, Espirometría, Modelos GAMLSS, Percentiles, Remuestreo, Test multivariado.

Índice de figuras

Índice de tablas

Parte I

Introducción

Capítulo 1

Estudio espirométrico

En un estudio sobre valores de espirometría es necesario identificar un modelo que permita caracterizar curvas percentilares de respuesta espirométricas por edad, sexo y demás características individuales de los participantes.

Fue Hutchinson (Spriggs, 1978) quien desarrolló en el año 1852 el primer espirómetro y las bases de los actuales conceptos sobre función respiratoria (Figura 1.1). Desde entonces se han desarrollado sobre estas bases los actuales espirómetros que constan de todos los parámetros clínicos necesarios para interpretar los estudios. Cuando se mide la función pulmonar se identifican parámetros clínicos en el volumen y en el flujo de las respiraciones, tanto en inspiración como espiración.

La maniobra más relevante es la espiratoria y en forma forzada, partiendo desde una inspiración profunda. Las 2 curvas que se presentan en este estudio son: la curva flujo/volumen y la curva volumen/tiempo. A través del esfuerzo espiratorio máximo se puede medir el Volumen Espiratorio Forzado o Capacidad Espiratoria Forzada (CVF), los flujos espiratorios forzados en el primer segundo (FEV_1) y los flujos forzados denominados periféricos (FEF_{25}, FEF_{50}, FEF_{75} y FEF_{25-75}), que corresponden a porciones de la curva Flujo/Volúmen y representan los flujos de la

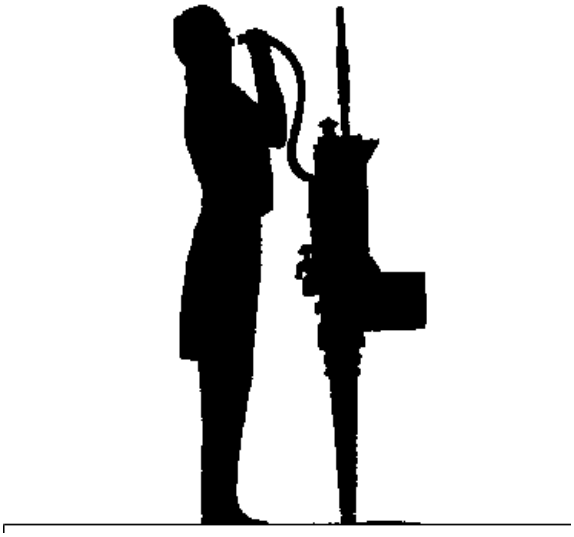

Figura 1.1: *Hutchinson y su espirómetro señalando la posición correcta para la realización de la espirometría.*

vía aérea más pequeña.

Además se mide el Indice de Gaënsler - el cual se determina con la espiración forzada expresada como un ratio FEV_1/CVF (que se interpreta como porcentaje de CVF). Ambos deberían ser iguales al realizar el mismo esfuerzo en forma forzada aunque en algunos casos el Índice de Gaënsler es menor debido al colapso de la vía aérea durante el esfuerzo y esto sucede siempre en los niños menores (Figura 1.2) .

La curva flujo/volumen comienza en el tiempo inspiratorio con una inspiración forzada y continua con una espiración forzada en el menor tiempo posible. En el tiempo espiratorio tiene un ascenso espiratorio rápido para luego descender en forma progresiva pero más lenta. En la primera parte del ascenso hasta llegar al pico de flujo espiratorio se utilizan todos los músculos espiratorios por lo que los parámetros que se miden en este tramo son esfuerzo dependientes. Luego de esta primera fase rápida comienza un descenso lento que sí corresponde a los flujos que no dependen de las fuerzas elásticas del pulmón por lo que adquieren importancia en la interpretación de las obstrucciones y restricciones.

La curva volumen/tiempo relaciona el volumen espirado con el tiempo empleado en

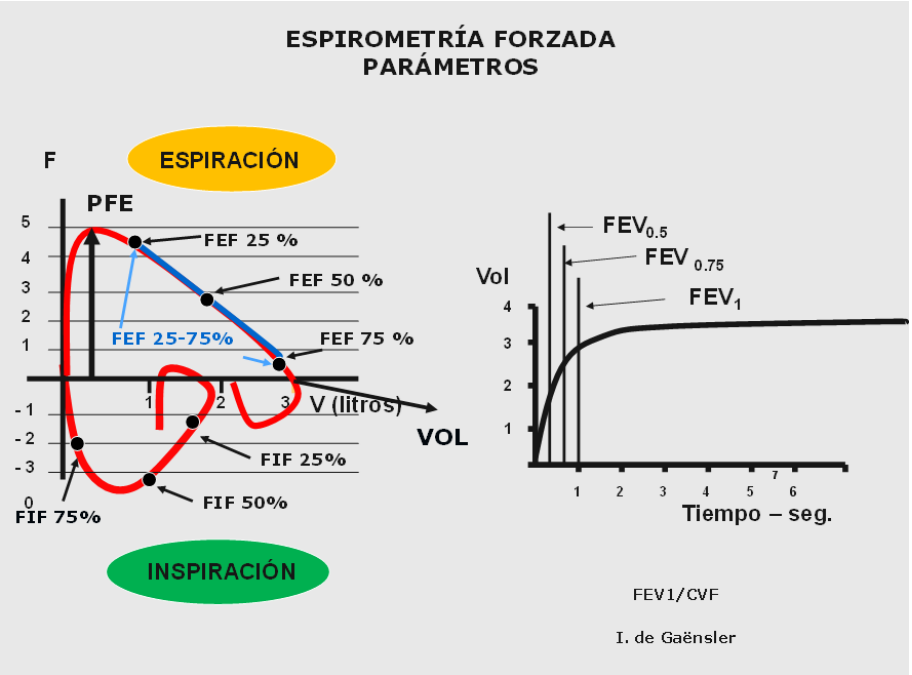

Figura 1.2: *Espirometría Forzada -Parámetros de curva flujo/volumen y volumen/tiempo*

la espiración. Tiene un ascenso rápido y luego una meseta que se prolonga hasta el final de la espiración. En ella podemos medir flujos, volúmenes y el tiempo espiratorio (Figura 1.2).

A continuación se listan los parámetros que mide un espirómetro.

PFE Pico de Flujo Espiratorio: Es el máximo volumen alcanzado en una espiración forzada. Se expresa en L/s (espirómetros) o L/min (medidores portátiles).

FEF Flujo Espiratorio Forzado: Al 25 %, 50 % y 75 % del volumen total espirado, y la porción 25-75 de la misma. Es decir, el flujo máximo cuando resta el 75 %, 50 % y 25 % del volumen a espirar. Se expresa en L/s.

FIF Flujo Inspiratorio Forzado. Al 25 %, 50 % y 75 % del volumen total inspirado. Se expresa en L/s.

Cuando existe obstrucción de la vía aérea se presenta una disminución de los flujos, tanto del FEV_1 como de los periféricos, manteniéndose la CVF. Cuando existe un

atrapamiento de aire o restricción el FEV_1 y el CVF disminuyen proporcionalmente y también la relación FEV_1/CVF, considerándose que existe una restricción.

La espirometría varía de acuerdo al tamaño de los pulmones. Por lo tanto, los valores varían de acuerdo a la edad, la talla y el peso. Pero también varían de acuerdo a la raza y a los diferentes países.

Por esta razón es necesario poseer valores estimados normales para las distintas poblaciones con el fin de que aquellos que se apartan de los rangos considerados como normales, puedan ser derivados para su estudio y controlar los tratamientos realizados en ellos. Es por eso que se desarrolla un estudio para medir la función pulmonar en una población de niños uruguayos considerados normales.

Antecedentes

A nivel nacional, no hay antecedentes de estudios con éstas características, pero hay varios estudios a nivel internacional referente a este tema, y los hay muy diversos, tanto en los rangos de edad que se manejan como en las metodologías empleadas.

Uno de los más completos fue hecho por Quanjer, Cole y Stanojevic (Quanjer *et al.*, 2012), un estudio multi étnico de la *Global Lung Iniciative* (GLI), con personas entre los 3 y 95 años, utilizando el método Lambda Mu Sigma (LMS), con un tamaño de muestra de aproximadamente 30000 personas de sexo masculino y 40000 de sexo femenino. Cuenta con una amplia etnicidad, donde se encuentran los grupos caucásicos, africanos-americanos, norasiáticos del este y sudasiáticos del este. Se incluyen datos para adultos de la ciudad de Montevideo. Utilizan un enfoque GAMLSS, pero a la hora de hacer una comparación con el presente estudio, tienen rangos de edad más amplio conteniendo más transiciones, entre niñez, juventud, adultez y vejez,

implicando un rol más importante de la edad.

Otro estudio tiene como autor a Cole (Cole y Stanojevic, 2009) realizado con personas entre 4 y 80 años con 1621 datos de distintos centros (EEUU, Bélgica, Inglaterra y Canadá). El modelo fue realizado con el enfoque GAMLSS.

Rosenthal (Rosenthal y Bain, 1993) lo desarrolla con modelos lineales partidos utilizando la talla como variable regresora, contemplando una población de niños y jóvenes entre 4 y 19 años.

Meng-Chiao (Meng-Chiao Tsai, 2010), tiene un enfoque similar a Rosenthal, sobre una población de niños entre 6 y 11 años de Taiwan, con un tamaño de muestra de 309, resultando en modelos lineales dependientes de la talla, con un rango de edades similar al presente estudio (6 a 11 años).

Objetivos

Objetivos primarios

Construir curvas percentilares de los parámetros CVF y FEV_1 a partir de modelos de regresión. Este es el objetivo primario debido a que permitirá comparar los valores de estudios espirométricos en niños uruguayos con valores de referencia obtenidos con datos de Uruguay.

Objetivos secundarios

- Explorar las diferencias entre parámetros espirométricos de los niños diagnosticados como "alérgicos" y los niños sanos ("normales").
- Estudiar la robustez del ajuste de densidades de los parámetros clínicos.

Este trabajo se estructura en cinco capítulos. En el capítulo 2 se desarrolla el marco teórico donde se hace referencia a la construcción de los modelos GAMLSS. Luego, en el capítulo 3 se aborda la aplicación, donde se hace una descripción de los datos, un análisis de las variables espirométricas entre niños alérgicos y niños normales, el estado nutricional de los mismos, el ajuste de distribuciones a las variables espirométricas, para culminar con la modelización de las mismas. En el capítulo 4 se presentan los resultados donde se hacen comparaciones con otros estudios sobre el tema, y por último, el capítulo 5, donde se presentan las principales conclusiones y consideraciones a futuro.

Parte II

Aspectos metodológicos

Capítulo 2

Marco Teórico

Metodología

El análisis de regresión es una de las técnicas estádisticas más populares y poderosas para la exploración de las relaciones entre una variable de respuesta y sus variables explicativas de interés. Los modelos de regresión se basan en ciertos supuestos que necesitan cumplirse para que éste tenga conclusiones válidas. En la práctica los supuestos normalidad, varianza constante de los errores y linealidad de la relación entre la variable de respuesta y las explicativas de los modelos de regresión lineal estándar, raramente se sostienen.

Los Modelos Lineales Generalizados (Generalized Linear Models, GLM) y los Modelos Aditivos Generalizados (Generalized Additive Models, GAM), fueron introducidos por (Nelder y Wedderburn, 1972) y por (Hastie y Tibshirani, 1986) respectivamente para superar algunas de las limitaciones de los modelos lineales estándar.

Desafortunadamente, especialmente con conjunto de datos muy grandes, estos mo-

delos han mostrado tener ajustes inadecuados o ser inapropiados en gran parte de situaciones prácticas.

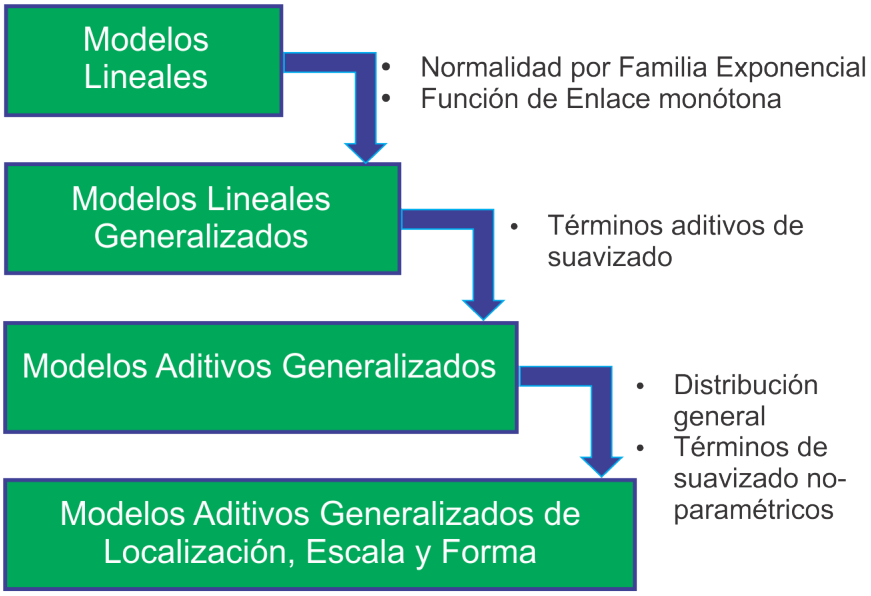

Figura 2.1: *Relación entre los diferentes tipos de Modelos lineales, Modelos lineales Generalizados, Modelos Aditivos Generalizados y Modelos Aditivos Generalizados de Localización, Escala y Forma*

Los Modelos Aditivos Generalizados de Localización, Escala y Forma (Generalized Additive Models for Location Scale and Shape, GAMLSS), son un marco de referencia que corrige algunos de los problemas de los GLM y GAM. Un GAMLSS es un modelo de regresión general, que asume que la variable de respuesta (dependiente), tiene alguna distribución paramétrica. Además, todos los parámetros de la distribución de la variable de respuesta pueden ser modelados como funciones de variables explicativas disponibles. Esto contrasta con los GLM y GAM, donde la distribución de la variable de respuesta está restringida a distribuciones de la familia exponencial y solo la media (parámetro de localización) de la distribución puede ser modelizada.

Entonces, la principal característica de los modelos GAMLSS es la habilidad de permitir que, la localización, la escala y la forma de la distribución de la variable de respuesta, varíen de acuerdo a los valores de las variables explicativas.

Los GAMLSS fueron introducidos por (Rigby y Stasinopoulos, 2001)(Rigby y Stasinopoulos, 2005), (Stasinopoulos y Rigby, 2007) y (Akanztiliotou *et al.*, 2002) como una forma de superar algunas de las limitaciones asociadas con los modelos lineales generalizados (GLM) y modelos aditivos generalizados (GAM) (Nelder y Wedderburn, 1972) y (Hastie y Tibshirani, 1986), respectivamente).

Modelos Lineales Generalizados

Los modelos lineales generalizados (GLM) (Nelder y Wedderburn, 1972) suponen una distribución de la familia exponencial (EF) para la variable de respuesta $\mathbf{Y_i}$, una función monótona de *enlace* (*link*), g(.) relacionando la media de la variable $\mathbf{Y_i}$, μ_i con el predictor lineal introducido η_i:

$$\mathbf{Y_i} \sim EF(\mu_i, \phi)$$

$$g(\mu_i) = \eta_i = \mathbf{x_i}^{\mathsf{T}}\beta \tag{2.1.1}$$

independientemente para $i = 1, 2, ..., n$. En notación vectorial es representado de la forma:

$$\mathbf{Y} \sim EF(\mu, \phi)$$

$$g(\mu) = \eta = \mathbf{x}^{\mathsf{T}}\beta \tag{2.1.2}$$

La distribución de la familia exponencial $EF(\mu, \phi)$ se define por la función de probabilidad (densidad) $f_{\mathbf{Y}}(\mathbf{y}; \mu, \phi)$ de $\mathbf{Y}$ de la forma:

$$f_{\mathbf{Y}} = exp\left\{\frac{\mathbf{y}\theta - b(\theta)}{\phi} + c(\mathbf{y}, \phi)\right\} \tag{2.1.3}$$

donde $E(\mathbf{Y}) = \mu = b'(\theta)$ y $Var(\mathbf{Y}) = \phi Var(\mu)$ donde la *función de varianza*

$V(\mu) = b''[\theta(\mu)]$. La forma de (2.1.3) incluye a varias distribuciones de las más utilizadas, incluyendo la normal, Poisson, Gamma, Gaussiana inversa, y también distribuciones binomial y binomial negativa.

Modelos Aditivos Generalizados

Un Modelo Aditivo Generalizado (GAM) (Hastie y Tibshirani, 1986) es un modelo lineal generalizado con un predictor lineal involucrando una suma de funciones de suavizado sobre las variables explicativas. En general, el modelo tiene una estructura de este estilo:

$$g(\mu_i) = \mathbf{X}_i^*\theta + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{2i}) + f_3(\mathbf{x}_{3i}, \mathbf{x}_{4i}) + ... \tag{2.1.4}$$

donde $\mu_i \equiv E(\mathbf{Y_i})$ y $\mathbf{Y_i}$ tiene alguna distribución de la familia exponencial. $\mathbf{Y_i}$ es una variable de respuesta, $\mathbf{X_i^*}$ es una fila de la matriz de modelo para cualquier componente estrictamente paramétrico del modelo, θ es el vector parámetro correspondiente, y las f_j son funciones de suavizado de las covariables, $\mathbf{x_k}$. El modelo permite una especificación más flexible de la dependencia de la variable de respuesta sobre las variables regresoras, pero especificando el modelo sólo en términos de funciones de suavizado en lugar de relaciones paramétricas detalladas, haciendo posible evitar modelos incómodos y complejos. Esta flexibilidad y conveniencia se produce a costa de dos nuevos problemas teóricos: la representación de las funciones de suavizado y la determinación del grado de suavidad (*smoothing*) que deberían tener.

Funciones de suavizado univariantes

Puede ser mejor introducir la representación de las funciones de suavizado considerando un modelo que contiene una función de suavizado de una sola variable explicativa,

$$\mathbf{y_i} = f(\mathbf{x_i}) + \epsilon_i, \tag{2.1.5}$$

donde $\mathbf{y_i}$ es una variable de respuesta, $\mathbf{x_i}$ es una variable explicativa, f es una función de suavizado, y los ϵ_i son variables aleatorias independientes e idénticamente distribuidas. $N(0, \sigma^2)$. Para simplificar aún más, supongamos que los $\mathbf{x_i}$ se encuentran en el intervalo [0,1].

Representando una función de suavizado: Splines de regresión

Para estimar f, usando los métodos anteriores, se requiere que f sea representada de alguna manera tal que (2.1.5) sea un modelo lineal. Esto puede hacerse eligiendo una *base*, definiendo el espacio de funciones donde f (o una aproximación cercana a ésta) sea un elemento del mismo. La elección de una base, equivale a la elección de algunas *funciones de base*, las cuales serán tratadas como completamente conocidas: si $b_j(\mathbf{x})$ es la *j-ésima* función de base, entonces f se asume que tiene una representación

$$f(\mathbf{x}) = \sum_{j=1}^{q} b_j(\mathbf{x})\beta_j \tag{2.1.6}$$

para algunos valores de los parámetros β_j desconocidos. Sustituyendo (2.1.6) en (2.1.5), el resultado es un modelo lineal.

Controlando el grado de suavizado con splines de regresión penalizados

Una alternativa para controlar la suavidad es mantener la dimensión de la base fija. Con un tamaño un poco más grande de lo que se cree razonable, podría ser necesario controlar la suavidad del modelo mediante la adición de una penalización de "ondulación", al ajuste de mínimos cuadrados objetivo. Por ejemplo, en lugar de ajustar el modelo minimizando

$$\|\mathbf{y} - \mathbf{X}\beta\|^2,$$

podría ser ajustado haciendo mínimo

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \int_0^1 \left[f''(\mathbf{x})\right]^2 d\mathbf{x}$$

donde la integral de la segunda derivada de f al cuadrado penaliza los modelos que son muy sinuosos. El intercambio entre el ajuste del modelo y la suavidad del modelo es controlado por el *parámetro de suavizado*, λ. Un $\lambda \to \infty$ conduce a una estimación de línea recta para f, mientras que $\lambda = 0$ da lugar a una estimación de regresión spline no-penalizada.

Como f es lineal en los parámetros, β_i, la penalización siempre puede ser escrita como una forma cuadrática en β.

$$\int_0^1 \left[f''(\mathbf{x})\right]^2 d\mathbf{x} = \beta^\intercal \mathbf{S} \beta$$

donde $\mathbf{S}$ es una matriz de coeficientes conocidos. Es ahora que la forma algo complicada de la base spline, que se utiliza aquí, demuestra su valor, porque resulta que $\mathbf{S}_{i+2,j+2} = R(x_i^*, x_j^*)$ para $i, j = 1, ..., q-2$ donde las primeras dos filas y columnas de $\mathbf{S}$ son nulas.

Por lo tanto, el problema de ajuste de regresión spline penalizado es minimizar

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\beta^\top \mathbf{S}\beta \tag{2.1.7}$$

con respecto a β. El problema de estimar el grado de suavidad para el modelo es ahora el problema de estimar el parámetro de suavizado λ.

Modelos aditivos

Ahora supongamos que dos variables explicativas, x y z, pueden ser utilizadas para estudiar su relación con una variable de respuesta y. Un modelo aditivo simple con estructura

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i \tag{2.1.8}$$

es apropiado. Las f_j son funciones de suavizado, y ϵ_i son variables aleatorias i.i.d. $N(0, \sigma^2)$. Por simplicidad, asumamos que tanto x_i como z_i tienen dominio en el intervalo $[0, 1]$.

Hay que notar dos cosas acerca de este modelo. En primer lugar, la suposición de poner efectos aditivos es fuerte: $f_1(x) + f_2(z)$ es un caso especial muy restrictivo de la función de suavizado general de dos variables $f(x, z)$. Y en segundo lugar, el hecho de que el modelo ahora contenga más de una función, presenta un problema de identificabilidad: f_1 y f_2 son estimables cada una dentro de una constante aditiva. Cualquier constante puede ser simultáneamente agregada a f_1 y ser sustraída de f_2 sin que se vea alterada la predicción del modelo. Por lo tanto, las limitaciones de identificabilidad tienen que ser impuestas en el modelo antes de su ajuste.

Proporcionada la identificabilidad, el modelo puede expresarse usando regresiones spline penalizadas, estimadas por mínimos cuadrados penalizados, y el grado de sua-

vizado estimado por validación cruzada, de la misma manera que el modelo simple univariado.

Ajustando modelos aditivos vía mínimos cuadrados penalizados

Los parámetros β del modelo (2.1.8) son obtenidos minimizando la función objetivo

$$\|y - \mathbf{X}\beta\|^2 + \lambda_1 \beta^\mathsf{T} S_1 \beta + \lambda_2 \beta^\mathsf{T} S_2 \beta$$

donde los parámetros de suavizado λ_1 y λ_2 controlan la ponderación del objetivo de hacer a f_1 y f_2 suaves, relativo al objetivo de ajustar lo más cercano posible los datos de respuesta. Asumamos por el momento que estos parámetros de suavizado vienen dados. Definiendo $S \equiv \lambda_1 \mathbf{S_1} + \lambda_2 \mathbf{S_2}$, la función objetivo puede ser reescrita como

$$\|y - \mathbf{X}\beta\|^2 + \beta^\mathsf{T} \mathbf{S} \beta = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix} \beta \right\|^2$$

donde $\mathbf{B}$ es cualquier matriz tal que $\mathbf{B}^\mathsf{T}\mathbf{B} = \mathbf{S}$. Al igual que en el caso de suavizado simple, la expresión del lado derecho es la función de mínimos cuadrados objetivo sin penalización para una versión aumentada del modelo y sus correspondientes datos de respuesta: entonces el modelo puede ser ajustado por una regresión lineal estándar.

Modelos Aditivos Generalizados

Los modelos aditivos *generalizados* (GAM) provienen de los modelos aditivos, como los modelos lineales generalizados de los modelos lineales. Esto es, el predictor lineal ahora predice alguna función de suavizado monótona del valor esperado de la variable de respuesta, y la respuesta puede seguir cualquier distribución de la familia exponencial, o simplemente tener una relación media-varianza conocida, permitiendo el enfoque de una cuasi-verosimilitud.

Como una ilustración, supongamos que se tiene un conjunto de datos de tres variables:

Girth Diámetro del árbol en pulgadas

Height Altura en pies

Volume Volumen de leña en pies cúbicos

Supongamos que queremos modelizar el conjunto de datos usando un GAM de la forma

$$log\left\{E\left(Volume_i\right)\right\} = f_1\left(Girth_i\right) + f_2\left(Height_i\right), Volume_i \sim gamma$$

Este modelo es quizás más natural que el modelo aditivo, ya que uno esperaría que el volumen sea producto de alguna función de la circunferencia y alguna función de la altura, y es probable que sea razonable esperar que la variación en el aumento del volumen se relacione con el volumen medio.

Mientras que el modelo aditivo se estima mediante mínimos cuadrados penalizados, el modelo aditivo generalizado se estima mediante la maximización de la verosimilitud penalizada: en la práctica esto se logrará mediante mínimos cuadrados penalizados iterativos, pero no hay un truco sencillo para producir un GLM no penalizado cuya verosimilitud sea equivalente a la verosimilitud penalizada del GAM que deseamos ajustar.

Para ajustar el modelo, simplemente iteramos el siguiente esquema de mínimos cuadrados penalizados reponderados iterativamente (P-IRLS) hasta la convergencia.

1. Dada la estimación actual del parámetro $\beta^{[k]}$ y su vector de respuesta media estimado $\mu^{[k]}$ correspondiente, calculamos:

$$\omega_i \propto \frac{1}{V\left(\mu_i^{[k]}\right) g'\left(\mu_i^{[k]}\right)} \text{ y } z_i = g\left(\mu_i^{[k]}\right)\left(y_i - \mu_i^{[k]}\right) + X_i\beta^{[k]}$$

donde $\text{Var}(Y_i)=V(\mu^{[k]})\phi$, y X_i es la i-ésima fila de la matriz X.

2. Minimizar

$$\left\|\sqrt{W}(z - X\beta)\right\|^2 + \lambda_1\beta^{\mathsf{T}}S_1\beta + \lambda_2\beta^{\mathsf{T}}S_2\beta$$

respecto de β para obtener $\beta^{[k+1]}$. W es una matriz diagonal tal que $W_{ii} = w_i$.

En este caso, la función de enlace, g, es el logaritmo, por lo tanto $g'(\mu_i) = \mu_i^{-1}$, mientras que para la distribución *gamma*, $V(\mu_i) = \mu_i^2$. Entonces, para el enlace logarítmico, el modelo *gamma* para los errores, tenemos:

$$w_i = 1 \text{ y } z_i = (y_i - \mu_i^{[k]})/\mu_i^{[k]} + \mathbf{X_i}\beta^{[k]}$$

Modelos Mixtos y Modelos Aditivos Mixtos Generalizados

Un abordaje diferente para estimar y hacer inferencia con los GAM se basa en representar los GAM como modelos mixtos con los términos de suavizado como *efectos aleatorios* (random effects).

(Pinheiro y Bates, 2000) ofrecen una amplia variedad de modelos que logran abarcar casi todas las formas de modelización con modelos mixtos lineales en R, mientras que (Ruppert *et al.*, 2003) incluye una explicación clara de las funciones de suavizado como componentes de los modelos mixtos.

Modelos lineales mixtos en general

El modelo lineal mixto puede ser escrito convenientemente como

$$y = \mathbf{X}\beta + \mathbf{Z}b + \epsilon, b \sim N(\mathbf{0}, \psi_\theta), \epsilon \sim N(\mathbf{0}, \mathbf{\Lambda}\sigma^{\mathbf{2}}) \tag{2.1.9}$$

donde ψ_θ es una matriz de varianzas y covarianzas definida positiva para los efectos aleatorios **b**, y **Z** es una matriz de coeficientes fijos que describen como la variable de respuesta, y, depende de los efectos aleatorios (es una matriz del modelo para los efectos aleatorios). ψ_θ depende de algunos parámetros, θ, que será el objetivo principal de inferencia estadística sobre los efectos aleatorios. Finalmente, $\mathbf{\Lambda}$, es una matriz definida positiva, que usualmente tiene una estructura simple dependiendo de pocos o ningún parámetros desconocidos: en ocaciones es utilizada par modelizar las correlación de los residuos, pero generalmente es una simple matriz identidad.

Se podría combinar el vector residual y los efectos aleatorios en un único, no independiente, vector residual variable-varianza, $\mathbf{e}= \mathbf{Zb} + \epsilon$. **e** es un vector normal multivariante de media cero, y su matriz de covarianza es $\mathbf{Z}\psi_\theta\mathbf{Z}^\mathsf{T} + \mathbf{I}\sigma^{\mathbf{2}}$. Entonces, (2.1.9) puede ser reescrita como:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}, \mathbf{e} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_\theta\sigma^{\mathbf{2}}) \tag{2.1.10}$$

donde $\mathbf{\Sigma}_\theta = \mathbf{Z}\psi_\theta\mathbf{Z}^\mathsf{T}/\sigma^{\mathbf{2}} + \mathbf{I}$, y θ, enfatiza la dependencia de Σ_θ en el vector paramétrico de covarianza. Entonces, si θ fuera conocido entonces se podría estimar β.

Estimación de modelos mixtos lineales

En general θ debe ser estimado, y la estimación por máxima verosimilitud provee el sistema básico para hacer esto. La verosimilitud de β, θ y σ^2 será

$$L(\beta, \theta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n \|\Sigma_\theta|}} exp\left[-(\mathbf{y} - \mathbf{X}\beta)^\intercal \Sigma_\theta^{-1}(\mathbf{y} - \mathbf{X}\beta)/(2\sigma^2)\right] \qquad (2.1.11)$$

y maximizando L respecto a β, θ y σ^2 nos dará sus estimaciones. Usualmente esta maximización pude ser simplificada por una verosimilitud *perfil*. La idea es, desde que ya se conoce exactamente como hallar las estimaciones máximo verosímiles de β y σ^2, para un θ dado, estos estimadores pueden ser insertados en la verosimilitud como funciones implícitas de θ, que derivan en la verosimilitud *perfil*

$$L_p(\theta) = \frac{1}{\sqrt{(2\pi\hat{\sigma}_\theta^2)^n \|\Sigma_\theta|}} exp\left[-(\mathbf{y} - \mathbf{X}\hat{\beta}_\theta)^\intercal \Sigma_\theta^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_\theta)/(2\hat{\sigma}_\theta^2)\right]$$

$\hat{\beta}_\theta$ y $\hat{\sigma}_\theta^2$ son las estimaciones por máxima verosimilitud/mínimos cuadrados de β y σ^2 dado θ. Para propósitos de maximización numérica, L_p puede ser tratada como función solamente de θ: cualquier valor que maximice L_p con respecto a θ automáticamente maximiza L.

Modelos mixtos lineales generalizados

Los modelos mixtos lineales generalizados (Generalized Linear Mixed Model - GLMM) provienen de los modelos mixtos lineales, del mismo modo que los modelos lineales generalizados provienen de los modelos lineales. Sea $\mu^b \equiv \mathbb{E}(y|b)$. Entonces un modelo mixto lineal generalizado tiene la forma

$$g(\mu_i^b) = \mathbf{X}\beta + \mathbf{Z_i}\mathbf{b}, \mathbf{b} \sim N(0, \psi) \text{ y } y_i|\mathbf{b} \sim \textit{distribución familia exponencial}$$

donde g es una función de enlace monótona, y la matriz de varianza-covarianzas, ψ, de los efectos aleatorios, usualmente es parametrizada en términos del vector pa-

ramétrico θ. Los $y_i|\mathbf{b}$ son independientes.

La verosimilitud de un GLMM es obtenida más fácilmente considerando la distribución conjunta de la respuesta y los efectos aleatorios.

$$f_{\beta,\theta,\phi}(\mathbf{y},\mathbf{b}) \propto |\psi|^{-1/2}\exp\big(logf(\mathbf{y}|\mathbf{b}) - \tfrac{1}{2}\mathbf{b}^{\mathsf{T}}\psi^{-\mathbf{1}}\mathbf{b}\big)$$

donde $f(\mathbf{y}|\mathbf{b})$ es la distribución conjunta de la variable respuesta condicionada al efecto aleatorio. Ahora la distribución marginal de $\mathbf{y}$, y por tanto la verosimilitud, se obtiene integrando respecto a los efectos aleatorios

$$L(\beta,\theta,\phi) \propto |\psi|^{-1/2}\int\exp\big(l(\beta,\mathbf{b}) - \tfrac{1}{2}\mathbf{b}^{\mathsf{T}}\psi^{-\mathbf{1}}\mathbf{b}\big)\,d\mathbf{b}$$

donde $l(\beta,\mathbf{b})$ es $f(\mathbf{y}|\mathbf{b})$ con la observación $\mathbf{y}$ insertada, considerada como función de β y $\mathbf{b}$, el logaritmo de la verosimilitud de un modelo lineal generalizado que resultaría del tratamiento de β y $\mathbf{b}$ como efectos fijos.

Desafortunadamente, esta integral generalmente tiene que ser aproximada, o evaluada numericamente, siendo esta última cada vez más difícil de evaluar a medida que aumenta la dimensión de $\mathbf{b}$.

Una simple aproximación es obtenida reemplazando $l_p = l(\beta,\mathbf{b}) - \frac{1}{2}\mathbf{b}^{\mathsf{T}}\psi^{-\mathbf{1}}\mathbf{b}$ por una aproximación cuadrática sobre los valores estimados $\hat{\beta}$ y $\hat{\mathbf{b}}$, que maximizan l_p: es la aproximación de Laplace a la integral.

La verosimilitud aproximada del modelo está dada por

$$L^*(\beta,\theta,\phi) \propto |\psi|^{-1/2}\int exp\left(-\tfrac{1}{2\phi}\left\|\mathbf{W}^{1/2}(\mathbf{z}-\mathbf{X}\beta-\mathbf{Z}\mathbf{b})\right\|^2 - \tfrac{1}{2}\mathbf{b}^{\mathsf{T}}\psi^{-1}b\right)d\mathbf{b}$$

que es simplemente la verosimilitud de un modelo lineal mixto ponderado, donde $z_i = g'\left(\hat{\mu}_i^b\right)\left(y_i - \hat{\mu}_i^b\right) + \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{b}}$ y $W_{ii} = \frac{1}{V\left(\hat{\mu}_i^b\right)g'\left(\hat{\mu}_i^b\right)^2}$.

Por lo tanto se puede obtener un máximo aproximado de L maximizando iterativamente L^*, usando los métodos ya discutidos para modelos lineales mixtos. Obsérvese que los resultados son sólo estimaciones aproximadas de máxima verosimilitud ahora, aún cuando el tamaño de la muestra tiende al infinito. La aproximación depende de cuán buena sea la aproximación de Laplace a la integral, y también de los W_{ii}.

Entonces el algoritmo para ajustar un modelo lineal mixto generalizado es:

1. Obtener elementos iniciales, $\hat{\beta}^{[1]}$ y $\hat{\mathbf{b}}^{[1]}$, por ejemplo, tomando $\hat{\mathbf{b}}^{[1]} = 0$ y ajustar el modelo lineal generalizado resultante para obtener $\hat{\mathbf{b}}^{[1]}$.

2. Tomar $k = 1$ e iterar los siguientes pasos hasta la convergencia.

3. Dado $\hat{\mathbf{b}}^{[k]}$ y $\hat{\mathbf{b}}^{[k]}$, buscar $\mathbf{z}$ y $\mathbf{W}$ como se definieron en el punto anterior.

4. Estimar los efectos del modelo lineal mixto

$$\mathbf{z} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon, \mathbf{b} \sim N(\mathbf{0}, \psi), \epsilon \sim N(\mathbf{0}, \mathbf{W}^{-1}\phi)$$

   para obtener estimaciones de $\hat{\beta}^{[k+1]}$, $\hat{\theta}^{[k+1]}$ y $\hat{\phi}^{[k+1]}$, y predicciones de $\hat{\mathbf{b}}^{[k+1]}$. Incrementando k de a uno.

Este método aproximado se conoce generalmente como Quasi Verosimilitud Penalizada (QVP), como resultado de la manera en que (Breslow y Clayton, 1993) trataron de justificarlo teóricamente.

Modelos aditivos mixtos generalizados

Un modelo aditivo mixto generalizado (Generalized Additive Mixed Model - GAMM) es sólo un modelo lineal mixto generalizado (GLMM) en el que la parte del predictor lineal es especificado en términos de funciones de suavizado de covariables (Lin y

Zhang, 1999). Por ejemplo, un Modelo Aditivo Mixto tiene una estructura similar a

$$y_i = \mathbf{X}_i\beta + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + ... + \mathbf{Z}_i\mathbf{b} + \epsilon_i \tag{2.1.12}$$

donde y_i es una respuesta univariada; θ es un vector de parámetros fijos; $\mathbf{X}_i$ es una fila de la matriz modelo de los efectos fijos; los f_j son funciones de suavizado de covariables x_k; $\mathbf{Z}_i$ es una fila de la matriz modelo de los efectos aleatorios; $\mathbf{b} \sim N(0, \psi)$ es un vector de coeficientes de los efectos aleatorios, con matriz de varianza ψ definida positiva desconocida; $\epsilon \sim N(0, \mathbf{\Lambda})$ es un vector residual con matriz de covarianza $\mathbf{\Lambda}$, la cual por lo general tiene algún diseño simple. La generalización de GLM a GAM requirió el desarrollo de la teoría para la regresión penalizada, para evitar el sobreajuste, pero los métodos de GLMM no requieren ningún ajuste para hacer frente a los GAMM: es posible escribir cualquiera de los *smoothers* de la regresión penalizada considerados, como componentes de un modelo mixto, mientras que el tratamiento de sus parámetros de suavizado como parámetros de componentes de la varianza, que se estima por máxima verosimilitud, métodos de máxima verosimilitud restringida (REML) o QVP.

Suavizados como componentes de modelos mixtos

En esta sección se explica, con más detalle, cómo se puede utilizar cualquier suavizado cuadráticamente penalizado, como un componente convencional de un modelo lineal mixto.

Primero consideremos un suavizado con un solo parámetro de suavizado. Por ejemplo,

$$f(\mathbf{x}) = \sum_{j=1}^{J} b_j(\mathbf{x})\beta_j$$

con la medida de ondulación asociada, $J(f) = \beta^{\mathsf{T}}\mathbf{S}\beta$, donde $\mathbf{S}$ es una matriz de coeficientes semi-definida positiva (semi-definida porque la mayoría de las penalizaciones tratan algún espacio de funciones como tener nula ondulación). Dado $(y_i, \mathbf{x}_i)$ es fácil producir una matriz modelo $\mathbf{X}^f$, donde $X^f_{ij} = b_j(\mathbf{b})$, así que $\mathbf{X}^f\beta$ es un vector de los valores de $f(\mathbf{x}_i)$.

Modelos Aditivos Generalizados de Localización, Escala y Forma

Los Modelos Aditivos Generalizados de Localización, Escala y Forma (Generalized Additive Model for Localization, Scale and Shape - GAMLSS) son un tipo de modelos de regresión semi-paramétricos. Son paramétricos, en el sentido que éstos requieren de una suposición de que la variable de respuesta tenga una distribución paramétrica, y "semi" en el sentido de que el modelado de los parámetros de la distribución, como función de las variables explicativas, pueden involucrar el uso de funciones de suavizado - *smoothing*- no paramétricas.

En los GAMLSS el supuesto de la familia exponencial para la distribución de la variable de respuesta (Y) es suspendido y remplazado por una distribución general, incluyendo distribuciones continuas y discretas con alto grado de asimetría y/o kurtósis. La parte sistemática del modelo es expandida para permitir el modelado, no solo de la media (o localización), sino que también de otros parámetros de la distribución de Y como función lineal y/o no lineal, paramétrica y/o suavizados

no-paramétricos de las variables explicativas y/o efectos aleatorios. Por lo tanto los GAMLSS están especialmente indicados para modelar una variable de respuesta que no sigue una distribución de la familia exponencial, o que presenta heterogeneidad (por ejemplo, cuando la escala y la forma de la distribución de la variable de respuesta cambian según las variables explicativas).

El modelo GAMLSS

Un modelo GAMLSS asume que, para $i = 1, 2, ..., n$, observaciones independientes de la variable de respuesta Y_i, ésta tiene función de densidad $f_Y\left(y_i|\theta^i\right)$ condicional en $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$, un vector de cuatro parámetros de distribución, donde cada uno puede ser una función de las variables explicativas.

Esto es denotado por $Y_i|\theta^i \sim D(\theta^i)$, o como $Y_i|(\mu_i, \sigma_i, \nu_i, \tau_i) \sim D(\mu_i, \sigma_i, \eta_i, \tau_i)$, independientemente para $i = 1, 2, ..., n$, donde D representa la distribución de Y. Nos vamos a referir a $(\mu_i, \sigma_i, \nu_i, \tau_i)$ como los *parámetros de distribución*. Los primeros dos parámetros de distribución de la población, μ_i y σ_i, se caracterizan normalmente por ser el parámetro de localización, y el parámetro de escala, mientras que los restantes, si los hay, son caracterizados como parámetros de forma (asimetría y kurtosis).

Sea $\mathbf{Y}^\mathsf{T} = (Y_1, Y_2, ..., Y_n)$ el vector de largo n de la variable de respuesta. (Stasinopoulos y Rigby, 2007) definen la formulación original de un modelo GAMLSS de la siguiente manera. Para $k = 1, 2, 3, 4$, sea $g_k(.)$ una función de enlace monótona conocida que relaciona el parámetro de distribución θ_k al predictor η_k.

$$g_k(\boldsymbol{\theta}_k) = \eta_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk} \tag{2.1.13}$$

llevado al caso

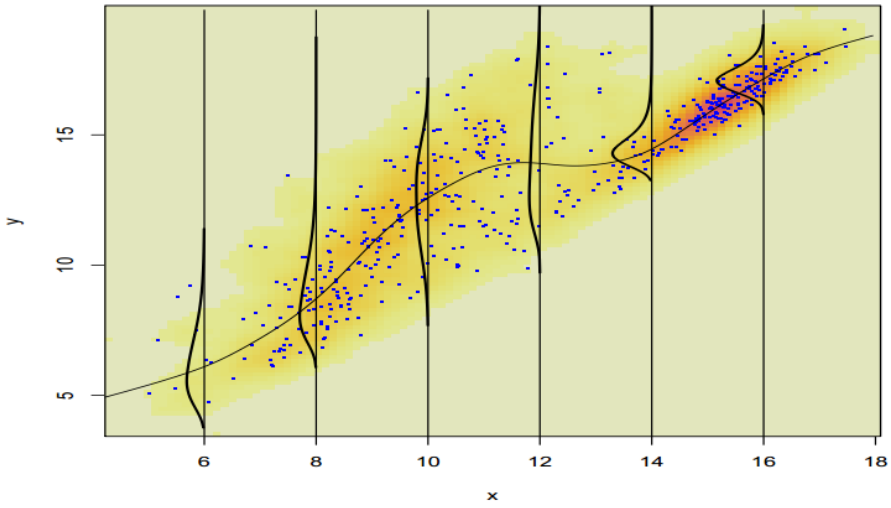

Figura 2.2: *Supuestos del modelo de regresión GAMLSS (Fuente: A flexible regression approach using GAMLSS in R, Rigby y Stasinopoulos, 2010)*

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}$$

donde $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$, y, para $k = 1, 2, 3, 4$, $\boldsymbol{\theta}_k$ y $\boldsymbol{\eta}_k$ son vectores de largo n, $\boldsymbol{\beta}_k^{\mathsf{T}} = (\beta_{1k}, \beta_{2k}, ..., \beta_{J'_k k})$ es un vector de parámetros de largo J'_k, $\mathbf{X}_k$ es una matriz de diseño fija conocida de dimensión $n \times J'_k$, $\mathbf{Z}_{jk}$ es una matriz de diseño fija conocida de $n \times q_{jk}$ y γ_{jk} es una variable aleatoria q_{jk}-dimensional que se asume que se distribuye $N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{\mathbf{jk}}^{-\mathbf{1}})$, donde $\mathbf{G}_{\mathbf{jk}}^{-\mathbf{1}}$ es la matriz inversa (generalizada) de una matriz simétrica de $q_{jk} \times q_{jk}$ $\mathbf{G}_{jk} = \mathbf{G}_{\mathbf{jk}}(\boldsymbol{\lambda}_{jk})$, la cual puede depender de un vector de hiperparámetros $\boldsymbol{\lambda}_{jk}$, y donde si $\mathbf{G}_{jk}$ es singular. Se entiende entonces que γ_{jk} tiene una función de densidad impropia a priori, proporcional a $\exp(-\frac{1}{2}\gamma_{jk}^{\mathsf{T}}\mathbf{G}_{jk}\gamma_{jk})$, mientras que si no es singular, entonces γ_{jk} tiene una distribución normal q_{jk}-variada con media $\mathbf{0}$ y matriz de varianza-covarianza $\mathbf{G}_{\mathbf{jk}}^{-\mathbf{1}}$.

El modelo (2.1.13) permite al usuario modelar cada uno de los parámetros de distribución como una función lineal de variables explicativas y/o como funciones lineales de variables estocásticas (efectos aleatorios). Se debe tener en cuenta que rara vez todos los parámetros de la distribución deberán ser modelizados utilizando variables explicativas.

Hay muchos casos particulares importantes de los GAMLSS. Por ejemplo, para aquellos que estén familiarizados con el suavizado, la siguiente formulación puede ser más familiar. Sea $\mathbf{Z}_{jk} = \mathbf{I}_n$, donde $\mathbf{I}_n$ es la matriz identidad de $n \times n$, y $\gamma_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ para todas las combinaciones de j y k en el modelo (2.1.13), entonces tenemos la formulación *aditiva semi-paramétrica* de GAMLSS dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{2.1.14}$$

donde h_{jk} es una función desconocida de la variable explicativa X_{jk} y $\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ es el vector el cual evalúa la función h_{jk} en $\mathbf{x}_{jk}$. Si no hubiera término aditivo ninguno de los parámetros de distribución, tenemos el modelo GAMLSS *lineal paramétrico* simple,

$$g_1(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k \tag{2.1.15}$$

El modelo (2.1.14) puede ser extendido para permitir términos paramétricos no-lineales para ser incluidos en el modelo para μ, σ, ν y τ, de la siguiente manera:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{2.1.16}$$

Nos vamos a referir al modelo (2.1.16) como *aditivo semi-paramétrico no-lineal*. Si, para $k = 1, 2, 3, 4$, $J_k = 0$, esto es, si para todos los parámetros de distribución no

tenemos términos aditivos, (2.1.16) se reduce a un modelo GAMLSS *paramétrico no-lineal*:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k). \tag{2.1.17}$$

Si, adicionalmente, $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) = \mathbf{X}_k^{\mathsf{T}\boldsymbol{\beta}_k}$ para $i = 1, 2, ..., n$ y $k = 1, 2, 3, 4$, entonces el modelo (2.1.17) se reduce a un modelo paramétrico lineal (2.1.15). Se debe destacar que algunos de los términos en cada $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k)$ pueden ser lineales, en cuyo caso el modelo GAMLSS es una combinación de términos paramétricos lineales y no-lineales. Vamos a referirnos a cualquier combinación de (2.1.15) o (2.1.17) como modelos GAMLSS paramétricos.

Los vectores paramétricos β_k y los parámetros de los efectos aleatorios γ_{jk}, para $j = 1, 2, ..., J_k$ y $k = 1, 2, 3, 4$, son estimados dentro del marco referencial GAMLSS (para valores fijos de los hiperparámetros de suavizado λ_{jk}) mediante la maximización de la función de verosimilitud penalizada $\ell_p(\beta, \gamma)$ dada por

$$\ell_p(\beta, \gamma) = \ell(\beta, \gamma) - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_{jk}^{\mathsf{T}}\mathbf{G}_{jk}\gamma_{jk} \tag{2.1.18}$$

donde $\ell(\beta, \gamma) = \sum_{i=1}^{n} log f_Y(y_i|\boldsymbol{\theta}^i) = \sum_{i=1}^{n} log f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ es la función log-verosimilitud de los parámetros de distribución dados los datos. Notar que se usa (β, γ) como argumento en la log-verosimilitud penalizada para enfatizar que es maximizado; (β, γ) representa todos los $\beta_k's$ y los $\gamma_{jk}'s$, para $j = 1, 2, ..., J_k$ y $k = 1, 2, 3, 4$. Para modelos GAMLSS paramétricos (2.1.15) o (2.1.17), $\ell_p(\beta, \gamma)$ se reduce a $\ell(\beta)$, y los β_k para $k = 1, 2, 3, 4$, son estimados maximizando la función de verosimilitud $\ell(\beta)$.

Implementación de GAMLSS en R: librería gamlss

Una de las formas para aplicar los GAMLSS es utilizando la librería **gamlss** implementada en el software R. La misma contiene una gran variedad de familias de distribución, tanto contínuas como discretas o mixtas y términos aditivos.

Distribuciones disponibles

La Tabla 2.1 lista todas las familias de distribución contínuas, mientras que las Tablas 2.2 y 2.3 contienen a las familias de distribución discretas y mixtas contenidas en la librería **gamlss**.

La forma de la distribución asumida por la variable de respuesta Y, $f_Y(y|\mu, \sigma, \nu, \tau)$, puede ser muy general. La única restricción que la implementación en R de los GAMLSS tiene es que la función $logf_Y(y|\mu, \sigma, \nu, \tau)$ y su primeras derivadas respecto a cada uno de los parámetros de $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$ deben ser computables. Las derivadas explícitas son preferibles, pero las derivadas numéricas pueden ser usadas.

Términos aditivos disponibles

Los GAMLSS permiten modelizar todos los parámetros de distribución μ, σ, ν y τ como funciones paramétricas lineales o no-lineales y/o funciones de suavizado paramétricas o no-paramétricas de las variables explicativas y/o términos de efectos aleatorios. En la implementación en R, la función `gamlss()` en el paquete **gamlss** permite fórmulas para todos los parámetros de distribución. Para modelar funciones lineales, se usa la forma usada por R en la función `lm()`, y `glm()`. Para ajustar funciones no-lineales o no-paramétricas (suavizado) y/o términos de efectos aleatorios, se deben incluir términos aditivos apropiados. En la Tabla 2.4 se presentan las

Tabla 2.1: *Tabla de distribuciones contínuas disponibles en el paquete* ***gamlss*** *(con funciones de enlace predeterminadas)*

Distribución	**Nombre R**	μ	σ	ν	τ
beta	BE()	logit	logit	-	-
Box-Cox Cole and Green	BCCG()	id	log	id	-
Box-Cox power exponential	BCPE()	id	log	id	log
Box-Cox t	BCT()	id	log	id	log
exponencial	EXP()	log	-	-	-
exponencial Gaussiana	exGAUS()	id	log	log	-
exponencial gen. Beta tipo 2	EGB2()	id	id	log	log
gamma	GA()	log	log	-	-
beta generalizada tipo 1	GB1()	logit	logit	log	log
beta generalizada tipo 2	GB2()	log	id	log	log
gamma generalizada	GG()	log	log	id	-
Gaussiana inversa gen.	GIG()	log	log	id	-
t generalizada	GT()	id	log	log	log
Gumbel	GU()	id	log	-	-
Gaussiana inversa	IG()	log	log	-	-
Johnson's SU (media μ)	JSU()	id	log	id	log
Johnson's SU original	JSUo()	id	log	id	log
logistica	LO()	id	log	-	-
log Normal	LOGNO()	log	log	-	-
log Normal (Box-Cox)	LNO()	log	log	fijo	-
NET	NET()	id	log	fijo	fijo
normal	NO()	id	log	-	-
familia normal	NOF()	id	log	id	-
power exponencial	PE()	id	log	log	-
Gumbel reversa	RG()	id	log	-	-
skew power exponencial tipo 1	SEP1()	id	log	id	log
skew power exponencial tipo 2	SEP2()	id	log	id	log
skew power exponencial tipo 3	SEP3()	id	log	log	log
skew power exponencial tipo 4	SEP4()	id	log	log	log
sinh-arcsinh	SHASH()	id	log	log	log
skew t tipo 1	ST1()	id	log	id	log
skew t tipo 2	ST2()	id	log	id	log
skew t tipo 3	ST3()	id	log	log	log
skew t tipo 4	ST4()	id	log	log	log
skew t tipo 5	ST5()	id	log	id	log
familia t	TF()	id	log	log	-
Weibull	WEI()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-
Weibull (media μ)	WEI3()	log	log	-	-

Distribución	**Nombre R**	μ	σ	ν
beta binomial	BB()	logit	log	-
binomial	BI()	logit	-	-
logarítmica	LG()	logit	-	-
Delaporte	DEL()	log	log	logit
binomial negativa tipo 1	NBI()	log	log	-
binomial negativa tipo 2	NBII()	log	log	-
Poisson	PO()	log	-	-
Poisson Gaussiana inv.	PIG()	log	log	-
Sichel	SI()	log	log	id
Sichel (media μ)	SICHEL()	log	log	id
beta binomial zero alterada	ZABB()	logit	log	logit
beta binomial zero alterada	ZABB()	logit	log	logit
binomial zero alterada	ZABI()	logit	logit	-
logarítmica zero alterada	ZALG()	logit	logit	-
binomial negativa zero alterada	ZANBI()	log	log	logit
Poisson zero alterada	ZAP()	log	logit	-
beta binomial zero aumentada	ZIBB()	logit	log	logit
binomial zero aumentada	ZIBI()	logit	logit	-
binomial negativa zero aumentada	ZINBI()	log	log	logit
Poisson zero aumentada	ZIP()	log	logit	-
Poisson zero aumentada (media μ)	ZIP2()	log	logit	-
Poisson Gaussiana inv. zero aumentada	ZIPIG()	log	log	logit

Tabla 2.2: *Tabla de distribuciones discretas disponibles en el paquete* ***gamlss*** *(con funciones de enlace predeterminadas)*

Distribución	**Nombre R**	μ	σ	ν	τ
beta aumentada (en 0)	BEOI()	logit	log	logit	-
beta aumentada (en 0)	BEINF0()	logit	logit	log	-
beta aumentada (en 1)	BEZI()	logit	log	logit	-
beta aumentada (en 1)	BEINF1()	logit	logit	log	-
beta aumentada (en 0 y 1)	BEINF()	logit	logit	log	log
gamma ajustada zero	ZAGA()	log	log	logit	-
gamma inversa zero ajustada	ZAIG()	log	log	logit	-

Tabla 2.3: *Tabla de distribuciones mixtas disponibles en el paquete* ***gamlss*** *(con funciones de enlace predeterminadas)*

distintas opciones de términos aditivos implementados dentro de la librería **gamlss**.

Términos aditivos	**Nombre de funciones en R**
boosting	`boost()`
splines de base cúbica	`cs()`, `scs()`, `vc()`
árboles de decisión	`tr()`
polinomios de potencia y fraccionales	`pp()`, `fp()`
suavizado de nodo libre	`fk()`
`loess`	`lo()`
redes neuronales	`nn()`
ajuste no-lineal	`nl()`
bases spline Beta penalizadas	`pb()`, `ps()`, `cy()`, `tp()`, `pvc()`
efectos aleatorios	`random()`, `ra()`, `rc()`, `re()`
regresión de cresta	`ri()`, `ridge()`
GAM de Simon Wood	`ga()`

Tabla 2.4: *Términos aditivos implementados en el paquete* ***gamlss***

Algoritmos para ajustar modelos de regresión paramétricos

Un típico modelo de regresión paramétrico dentro del marco de GAMLSS asume que la variable de respuesta $Y_i \sim D(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$, independientemente para $i = 1, ..., n$, donde los vectores de largo n de los parámetros de distribución pueden ser modelados como funciones de variables explicativas como

$$g1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\beta_1$$

$$g2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\beta_2$$

$$g3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\beta_3$$

$$g4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\beta_4$$

donde las matrices $\mathbf{X}$ contienen los valores de las variables explicativas, los $\boldsymbol{\eta}$ son los predictores lineales, los $g()$ son funciones de enlace conocidas (usualmente para garantizar que los parámetros distribucionales tengan un rango adecuado) y los $\boldsymbol{\beta}$

los coeficientes a estimar.

La verosimilitud a ser maximizada respecto a los parámetros $\boldsymbol{\beta}$ será

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4) \tag{2.1.19}$$

Resulta que la verosimilitud en (2.1.19) se puede maximizar usando un algoritmo iterativo (descrito abajo) que utiliza repetidamente regresiones lineales ponderadas simples. Las cantidades necesarias para el algoritmo RS son:

- Función de score: $\mathbf{u}_k = \frac{\partial \ell}{\partial \boldsymbol{\eta}_k}$, para $k = 1, 2, 3, 4$.
- variables dependientes ajustadas: $\mathbf{z}_k = \boldsymbol{\eta}_k + [\mathbf{W}_{kk}]^{-1}\mathbf{u}_K$, para $k = 1, 2, 3, 4$.
- Matrices diagonales de pesos iterativos: $\mathbf{W}_{kk}$ que pueden tener una de las siguientes formas $-\frac{\partial^2 \ell}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_k^T}$, $-E[\frac{\partial^2 \ell}{\partial \eta_k \partial \eta_k^T}]$ o $diag\left\{\left[\frac{\partial \ell}{\partial \eta_k}\right]^2\right\}$, es decir, la información observada, la información esperada o la función de score de producto, dependiendo respectivamente de si se utiliza un algoritmo Newton-Raphson, score de Fisher o cuasi Newton-Raphson (Lange, 1999).

A continuación de describe una versión simplificada del algoritmo RS (el método por defecto utilizado en la función `gamlss()`) utilizado para ajustar los modelos. Sea r el índice de iteración del ciclo externo (outer cycle), k el índice del parámetro e i el índice de iteración del ciclo interno (inner cycle). Esencialmente, el algoritmo RS tiene un ciclo externo que comprueba la maximización de la verosimilitud general con respecto a los $\boldsymbol{\beta}$ y un ciclo interno para ajustar un modelo para cada parámetro distribucional, para $k = 1, 2, 3, 4$, donde los otros parámetros de distribución se fijan a sus valores actuales. Obsérvese en cada cálculo del algoritmo el uso de los valores actualizados más recientes de todas las cantidades. Notar que $\boldsymbol{\theta}^\mathsf{T} = (\theta_1, \theta_2, \theta_3, \theta_4) = (\mu, \sigma, \nu, \tau)$. El algoritmo RS se puede describir como sigue:

- **Inicio**: inicializar valores ajustados $\boldsymbol{\theta}_k^{(1,1)}$ para $k = 1, 2, 3, 4$ de los vectores de parámetros distribucionales de largo n, $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}$ y $\boldsymbol{\tau}$ respectivamente. Se evalúa el predictor lineal inicial $\boldsymbol{\eta}_k^{(1,1)} = g_k\left[\boldsymbol{\theta}_k^{(1,1)}\right]$, para $k = 1, 2, 3, 4$.

- **INICIAR OUTER CYCLE** $r = 1, 2, ...$ **HASTA CONVERGENCIA PARA** $k = 1, 2, 3, 4$

  - **INICIAR INNER CYCLE** $i = 1, 2...$ **HASTA CONVERGENCIA**

    - Evaluar los $\mathbf{u}_k^{(r,i)}, \mathbf{W}_{kk}^{(r,i)}, \mathbf{z}_k^{(r,i)}$ actuales.

    - Se regresan los $\mathbf{z}_k^{(r,i)}$ actuales contra la matriz de diseño $\mathbf{X}_k$ usando los ponderadores iterativos $\mathbf{W}_{kk}^{(r,i)}$ para obtener los parámetros estimados $\boldsymbol{\beta}_k^{(r,i)}$ actualizados.

  - **FIN INNER CYCLE** en la convergencia de $\boldsymbol{\beta}_k^{(r,.)}$ y tomar $\boldsymbol{\beta}_k^{(r+1,1)} = \boldsymbol{\beta}_k^{(r,i)}$, $\boldsymbol{\eta}_k^{(r+1,1)} = \boldsymbol{\eta}_k^{(r,.)}$ y $\boldsymbol{\theta}_k^{(r+1,1)} = \boldsymbol{\theta}_k^{(r,.)}$, de lo contrario, actualizar i y continuar con el ciclo interno.

  ACTUALIZAR valor de k

- **FIN OUTER CYCLE**: Si el cambio en la verosimilitud penalizada es suficientemente pequeño, de lo contrario actualizar r y continuar el ciclo externo

Criterios de selección del modelo

Selección del modelo en GAMLSS

Sea $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \Lambda\}$ un modelo GAMLSS como se definió anteriormente. Los componentes de $\mathcal{M}$ son definidos de la siguiente manera: (i) $\mathcal{D}$ especifica la distribución de la variable de respuesta; (ii) $\mathcal{G}$ especifica el conjunto de funciones de enlace; (iii) $\mathcal{T}$ especifica las términos que aparecen en todos los predictores para μ, σ, ν y τ; (iv) Λ especifica los hiperparámetros de suavizado que determina el grado del mismo en las funciones $h_{jk}()$.

En la búsqueda de un modelo GAMLSS apropiado para cualquier conjunto de datos nuevo, se deben especificar cada uno de los cuatro componentes lo más objetivamente posible.

Componente $\mathcal{D}$: Selección de la distribución

La selección de la distribución apropiada puede lograrse en dos etapas, la etapa de *ajuste* y la etapa de *diagnóstico*. La etapa de ajuste involucra la comparación de diferentes modelos ajustados utilizando en criterio de información de Akaike generalizado (GAIC). El modelo con el menor valor de GAIC(k), para un valor de k elegido, es seleccionado.

El GAIC consiste en el logarítmo de la verosimilitud y un factor de penalización k fijo que multiplica los grados de libertad efectivos totales (df), definido de la forma

$$GAIC(k) = -2 \sum_{i=1}^{n} log \left[f(y_i | \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}) \right] + k.df$$

.

Las etapas de diagnóstico involucran el uso de *worm plots*. Los worm plots han sido introducidos por (Buuren y Fredriks, 2001) y son QQ-plots normales sin tendencia de los cuantiles de los residuos (z-scores). Éstos permiten la detección de inadecuaciones en el modelo, tanto globalmente como dentro de un rango específico de una (o dos) variables explicativas.

Componente $\mathcal{G}$: Selección de las funciones de enlace

La elección de las funciones de enlace para cada uno de los parámetros de distribución es usualmente determinada por el rango del parámetro en cuestión. Por ejemplo, en una distribución Pareto II (PARETO II), tanto μ como σ toman valores positivos, por lo que una función de enlace log() es una manera natural de asegurarse que ambos parámetros permanezcan positivos (cualquiera sea el valor de sus predictores). Para una distribución normal, $-\infty < \mu < \infty$ y $0 < \sigma < \infty$, entonces una función de enlace identidad para μ y una función enlace log() para σ asegura que ambos parámetros estén en su rango.

Hay ocasiones en cual la elección de las funciones de enlace es importante desde el punto de vista interpretativo. Por ejemplo, si creemos que una variable explicativa afecta el parámetro de distribución de manera multiplicativa en vez de aditivamente, entonces un enlace logarítmico es más apropiado.

La elección del enlace puede mejorar el ajuste del modelo considerablemente. Diferentes funciones de enlace se pueden comparar directamente usando la *global deviance* (GD). La mejor función de enlace resulta en la GD más baja.

Componente $\mathcal{T}$: Selección de los términos aditivos en el modelo

Sea $\mathcal{X}_i$ un *pool* de términos disponibles a considerar para el parámetro $\boldsymbol{\theta}_i$ para $i = 1, 2, 3, 4$, donde $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3}, \boldsymbol{\theta_4}) = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$. Típicamente $\mathcal{X}_i$ contendrá términos aditivos lineales y de suavizado. Por ejemplo, sean dos factores, f_1 y f_2, y x_1, x_2, x_3 y x_4 variables explicativas contínuas. Entonces, por ejemplo,

$$\mathcal{X}_i = \{f_1 * f_2 + s(x_1) + s(x_2, x_3) + x_4\}$$

permite interacciones de segundo orden para los dos factores, función de suavizado para x_1, un suavizado con interacción para x_2 y x_3, y un término lineal para x_4. Puntos a destacar:

- Dada la distribución para una variable de respuesta, la selección de los términos tiene que hacerse para **todos** los parámetros de la distribución asumida, no solo el parámetro de localización. Los procedimientos habituales de *forward*, *backward* y *stepwise* pueden ser aplicadas para cada parámetro pero también se debe pensar en como aplicar estos procedimientos cuando se eligen los términos para cada parámetro.

- Los términos aditivos pueden influenciar a los parámetros de la distribución de diferentes formas. Por ejemplo, en el ejemplo anterior la interacción de los factores f_1 y f_2 afectan el parámetro de interés. La variable x_4 lo afecta linealmente, la variable x_1 no-linealmente, mientras que una interacción no lineal de suavizado entre x_2 y x_3 afectan el parámetro de interés.

- El tamaño de los términos disponibles $\mathcal{X}_i$ relativo al número de observaciones en la muestra importa tanto como la selección de los términos. Por ejemplo, si el número de variables explicativas contínuas es pequeño, digamos 5, todas las $2^5 = 25$ diferentes combinaciones de cómo esas variables pueden influenciar un parámetro pueden probarse. Por otra parte, cuando se trata con un mayor

número de variables contínuas, como 50, hay $2^{50} = 1{,}13 \times 10^5$ combinaciones diferentes que no todas pueden ser ajustadas, por lo que se debe implementar otra estrategia.

Hay muchas funciones dentro de **gamlss** para asistir con la selección de términos para las variables explicativas cuando todos los datos son utilizados para la selección de variables. Las funciones básicas son `addterm()` y `dropterm()` que permite la adición o sustracción de un término en el predictor de un parámetro respectivamente. Estas funciones son bloques de construcción de la función `stepGAIC()` disponible para la selección *stepwise* de términos para un parámetro de distribución de un modelo GAMLSS usando el criterio de información de Akaike generalizado (GAIC).

Componente Λ: Selección de los parámetros de suavizado

Cada término de suavizado seleccionado para cualquiera de los parámetros de la distribución tiene al menos un parámetro (o hiper) de suavizado λ asociado con él. Se denota con Λ al conjunto de todos los hiperparámetros de la distribución, $\Lambda = \{\lambda_{\mu,1}, \lambda_{\mu,2}, \lambda_{\sigma,1}, \lambda_{\nu,1}\}$.
Los parámetros de suavizado pueden se fijados o estimarse con los datos. La manera estándar de fijar un parámetro de suavizado es fijando los grados de libertad efectivos para el suavizado. Muchos de los procedimientos de suavizado dentro de los paquetes de **gamlss** permiten al usuario hacer eso. Generalmente es deseable estimar el parámetro de suavizado automáticamente.

Los siguientes son tres métodos comunes para estimar los parámetros de suavizado:

- Validación cruzada generalizada (GCV)
- GAIC

- Método de máxima verosimilitud

Cada método puede implementarse de dos maneras:

localmente: cuando el método es aplicado cada vez dentro del algoritmo iterativo GAMLSS

globalmente: cuando el método es aplicado fuera del algoritmo iterativo GAMLSS

Según (Stasinopoulos y Rigby, 2007), los métodos locales suelen ser mucho más rápidos y a menudo producen resultados similares a los métodos globales. Los métodos globales suelen ser más confiables.

Selección de todos los componentes usando un conjunto de datos de validación

Para conjuntos de datos grandes, dentro de GAMLSS, el modelizador estadístico puede permitirse separar los datos en diferentes partes, por ejemplo:

1. Puede usarse *datos de entrenamiento* para ajustar el modelo (minimizando su DG)

2. Puede usarse *datos de validacíon* para la selección del modelo, en particular, la distribución, funciones de enlace, términos de los predictores y los parámetros de suavizado (minimizando su DG, denotado por VGD)

3. Puede usarse *datos de prueba* para evaluar el poder predictivo del modelo elegido por (2) y ajustado por (1) y aplicado a los datos de prueba (usando la DG, denotado por TGD)

Hay varias funciones que asisten en la elección del modelo en estos casos, las cuales

se presentan en la Tabla 2.6.

Componente	**Datos completos**	**Validación cruzada K-fold**	**Datos de validación y prueba**
$\mathcal{D}$	`GAIC(), wp()`	`gamlssCV(), CV()`	`gamlsVGD(), VGD(), getTGD(), TGD()`
$\mathcal{G}$	`deviance()`	`gamlssCV(), GV()`	como arriba
$\mathcal{T}$	`drop1(), add1(), add1ALL(), drop1ALL(), stepGAIC(), stepGAICALL(), stepGAICALL.A(), stepGAICALL.B()`	`gamlssCV(), CV()`	`drop1TGD(), add1TGD(), stepTGD()`
Λ global	`findhyper(), optim()`	`optim()`	`optim()`

Tabla 2.6: *Diferentes funciones de selección de modelos de acuerdo a que componente de la distribución es usado y de acuerdo a diferentes configuraciones de los datos.*

Comparación entre modelos anidados.

La elección entre modelos es importante porque los modelos GAMLSS son flexibles y por tanto permite diferentes escenarios posibles para un conjunto de datos dado. Se debe ser capaz de elegir entre esos escenarios de una manera consistente.

Test de razón de verosimilitudes (LRT)

Sean M_0 y M_1 dos modelos diferentes. M_0 está anidado en M_1, es decir, M_0 es un caso especial de M_1. Supongamos que M_0 es el modelo más simple y M_1 el más complicado.

Dos modelos GAMLSS paramétricos anidados, donde M_1 es un sub modelo de M_0, con *deviance* globales ajustadas GD_0 y GD_1 y los grados de libertad de los errores df_{e0} y df_{e1} respectivamente, pueden ser comparados usando el estadístico de prueba $LRT = GD_0 - GD_1$ que tiene una distribución asintótica Chi-cuadrado cuando

M_0 es referencia, con $d = df_{e0} + df_{e1}$ grados de libertad. Se rechaza H_0 si $LRT \geq \chi^2_{d,\alpha}$.

Comparación entre modelos no anidados.

Para comparar modelos GAMLSS no anidados (por ejemplo, con familias de distribución distintas), para penalizar el sobreajuste (*overfitting*) se puede usar el criterio de información de Akaike generalizado (GAIC). El modelo con menor GAIC(k) es seleccionado. El criterio de información de Akaike (AIC) y criterio bayesiano de Schwartz (SBC) son casos especiales del criterio GAIC(k), donde $k = 2$ y $k = log(n)$ respectivamente.

Ambos criterios, AIC y SBC, se justifican asintóticamente como predicción del grado de ajuste en un nuevo conjunto de datos, es decir, aproximaciones al error predictivo medio. La justificación para el uso de SBC viene también como una aproximación a los factores de Bayes. En la práctica se encuentra generalmente que, mientras que el AIC original es muy permisivo en la selección de modelos, el SBC es demasiado restrictivo.

Diagnóstico del modelo

En el modelo de regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + e_i$ se definen los residuos como la diferencia entre los valores observados y los ajustados $\hat{\epsilon}_i = y_i - \hat{y}_i$ donde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ para $i = 1, 2, ..., n$. Algunas veces los $\hat{\epsilon}_i$ son llamados residuos *crudos* para distinguirlos de los *residuos estandarizados* los cuales son definidos como $(y_i - \hat{y}_i)/\hat{\sigma}\sqrt{(1 - h_{ii})}$, donde h_{ii} son los valores de la diagonal de la matriz $\mathbf{H}$ (ecuación **??**). El problema con los *raw residuals* es que son difíciles de generalizar a otras distribuciones diferentes a la distribución normal. Por ejemplo, dentro de la literatura

de los modelos lineales generalizados los *residuos de desvianza* $r_i^d = sign(y_i - \hat{\mu}_i)/\sqrt{d_i}$ donde $d_i = -2log(L_i^c/L_i^s)$ o los *residuos de Pearson* $r_i^P = (y_i - \hat{\mu}_i)/se(\hat{\mu}_i)$ a menudo son utilizados. Desafortunadamente, los residuos de desvianza no están bien definidos con múltiples parámetros para la distribución de y, mientras que los residuos de Pearson pueden estar lejos de una distribución normal y tampoco son apropiados para datos altamente sesgados o kurtóticos. Por lo tanto para GAMLSS se usan los *cuantiles normalizados (aleatorizados) residuales* (Dunn y Smyth, 1996).

Cuantiles residuales normalizados (aleatorizados)

La principal ventaja es que, sea cual sea la distribución de la variable de respuesta, sus verdaderos valores r_i,$i = 1, 2, ..., n$ siempre tienen una distribución normal estándar bajo el supuesto de que el modelo es correcto. Dado que dentro de la literatura estadística se comprueba que el supuesto de normalidad esté bien establecido, los cuantiles residuales normalizados brindan un camino sencillo para comprobar la adecuación de un modelo GAMLSS ajustado.

Dada la distribución $f(y, \boldsymbol{\theta})$ ajustada a las observaciones y_i para $i = 1, 2, ..., n$, los cuantiles residuales normalizados ajustados son dados por $\hat{r}_i = \Phi^{-1}(\hat{\mu}_i)$, donde Φ^{-1} es función de distribución acumulada inversa de una variable normal estándar. Los $\hat{\mu}_i$ son cuantiles residuales definidos diferentemente para variables de respuesta contínuas y discretas.

Si y es una observación de una variable de respuesta contínua, entonces, sean $u = F(y|\boldsymbol{\theta})$ y $\hat{u} = F(y|\hat{\boldsymbol{\theta}})$ el modelo y las funciones de distribución acumulada respectivamente. Si el modelo está correctamente especificado, u tiene una distribución uniforme entre cero y uno. u es transformada en *z-score*, r, usando $r = \Phi^{-1}(u)$, por lo que r tendrá una distribución normal estándar. Notar que $r = \Phi^{-1}[F(y|\boldsymbol{\theta})]$.

Similarmente $\hat{u}$ es transformada en $\hat{r}$ por $\hat{r} = \Phi^{-1}(\hat{u}) = \Phi^{-1}[F(y|\hat{\boldsymbol{\theta}})]$ y $\hat{r}$ tiene aproximadamente una distribución normal estándar. Si y es una observación de una variable de respuesta discreta entera, entonces u es un valor aleatorio de la distribución uniforme en el intervalo $[u_1, u_2] = [F(y-1|\boldsymbol{\theta}), F(y|\boldsymbol{\theta})]$ y $\hat{u}$ es un valor aleatorio de una distribución uniforme en $[\hat{u}_1, \hat{u}_2] = [F(y-1|\hat{\boldsymbol{\theta}}), F(y|\hat{\boldsymbol{\theta}})]$. El procedimiento es similar al de una variable de respuesta contínua.

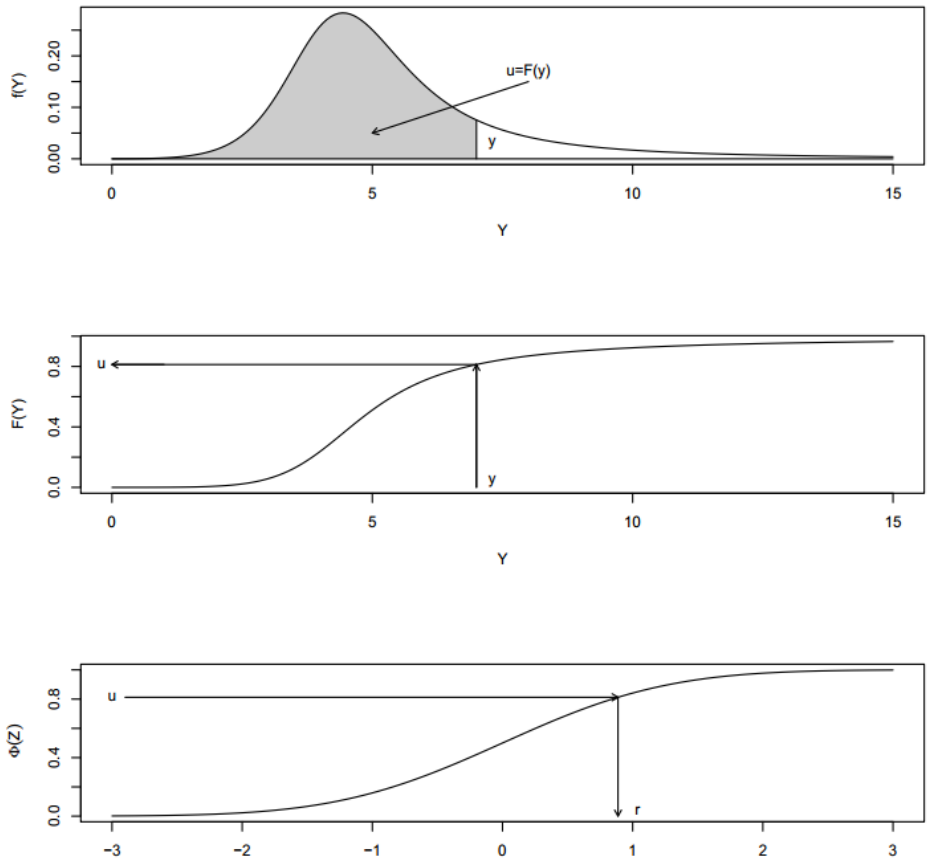

Figura 2.3: *Una descripción de como se obtienen los residuos r para una distribución contínua. Las funciones graficadas son la función de densidad del modelo $f(y)$, la función de distribución acumulada $F(y)$ y la función de distribución acumulada de la variable normal estandarizada $\Phi(z)$, en la cual y es transformada en u y luego de u a r. Los residuos r son el z-score para una observación específica y tiene una distribución normal estándar si el modelo es correcto. Fuente: Flexible Regression and Smoothing The GAMLSS packages in **R**, Stasinopoulos - 2015*

Hay varias funciones implementadas que usan los cuantiles normalizados (aleatori-

zados) residuales.

- La función `plot.gamlss()` es para un chequeo general de los residuos
- La función de worm plots, `wp()`, la cual puede ser usada para identificar si la distribución ajustada es adecuada de forma global o dentro de rangos que no se superponen de una o dos variables explicativas.
- La función de los estadísticos Q (Royston y Wright, 2000), `Q.stats()`, para detectar si los residuos son "significativamente" diferentes a una distribución normal en su media, varianza, oblicuidad (skewness) y curtosis (y más potencialmente qué parámetro de distribución del modelo no se ajustó adecuadamente) en los rangos de la variable explicativa.
- La función `rqres.plot()` designada para una aleatorización repetida de los residuos (para cuando la variable de respuesta no es contínua)

Parte III

Resultados

Capítulo 3

Aplicación

A continuación se van a presentar las características del estudio. Se dará cuenta de las características de la muestra, los aspectos técnicos referente a los estudios espirométricos llevados adelante y los aspectos éticos del mismo. Luego se hace una descripción de los datos, donde se verán las relaciones entre las variables antropométricas, y luego su relación con las variables espirométricas CVF y FEV_1.

Posteriormente se presenta un análisis de las variables espirométricas según niños alérgicos y niños normales, y el estado nutricional de los niños para poder compararlos con los valores de la OMS.

Luego se aborda el problema de encontrar una distribución para las variables CVF y FEV_1, introduciendo una prueba de robustez y sus resultados en el contexto global y separado por sexo.

Se da paso luego a la parte de modelización de las variables CVF y FEV_1 en los distintos escenarios.

Características del estudio

El estudio es llevado adelante por un grupo de investigadores del Centro Hospitalario Pereira Rossell entre los años 1997 y 1999.

Ante la imposibilidad de realizar un estudio aleatorizado de todas las escuelas públicas y privadas del país, se seleccionó una muestra de escuelas a las únicas que podía acceder el equipo de investigación por temas logísticos y de permisos (distancias, transaldos, equipo disponible, permiso de Consejo de Educación Primaria), incluyendo zonas en donde existe ascendencia indígena (Tacuarembó) y de distintos niveles de contaminación ambiental. En la muestra no se encontró población afrodescendiente.

Los escolares participantes del estudio provienen de 7 escuelas públicas y privadas del interior del país (Colonia, Dolores, Paysandú, Tacuarembó) y de Montevideo, se desconoce la tasa de no respuesta y no puede determinarse a priori si existió algún sesgo de selección.

Los criterios de selección de los niños fueron los siguientes:

- Niños con examen físico normal al momento del estudio.
- No haber presentado antecedentes luego del 1er año de vida de: sibilancias, asma, broncoespasmo inducido por el ejercicio, y/o bronquitis reiteradas.
- Haber realizado la maniobra de espiración forzada en forma satisfactoria.

De un total de 1021 niños participantes, 878 cumplieron con los criterios de inclusión

(412 varones y 466 niñas). El proceso de como se llega a los datos definitivos para su análisis se detalla en la Figura 3.1 (es necesario aclarar que el proceso de eliminación de datos, tanto por ser poco confiables como por duplicación, se hizo previo a la implementación de los modelos).

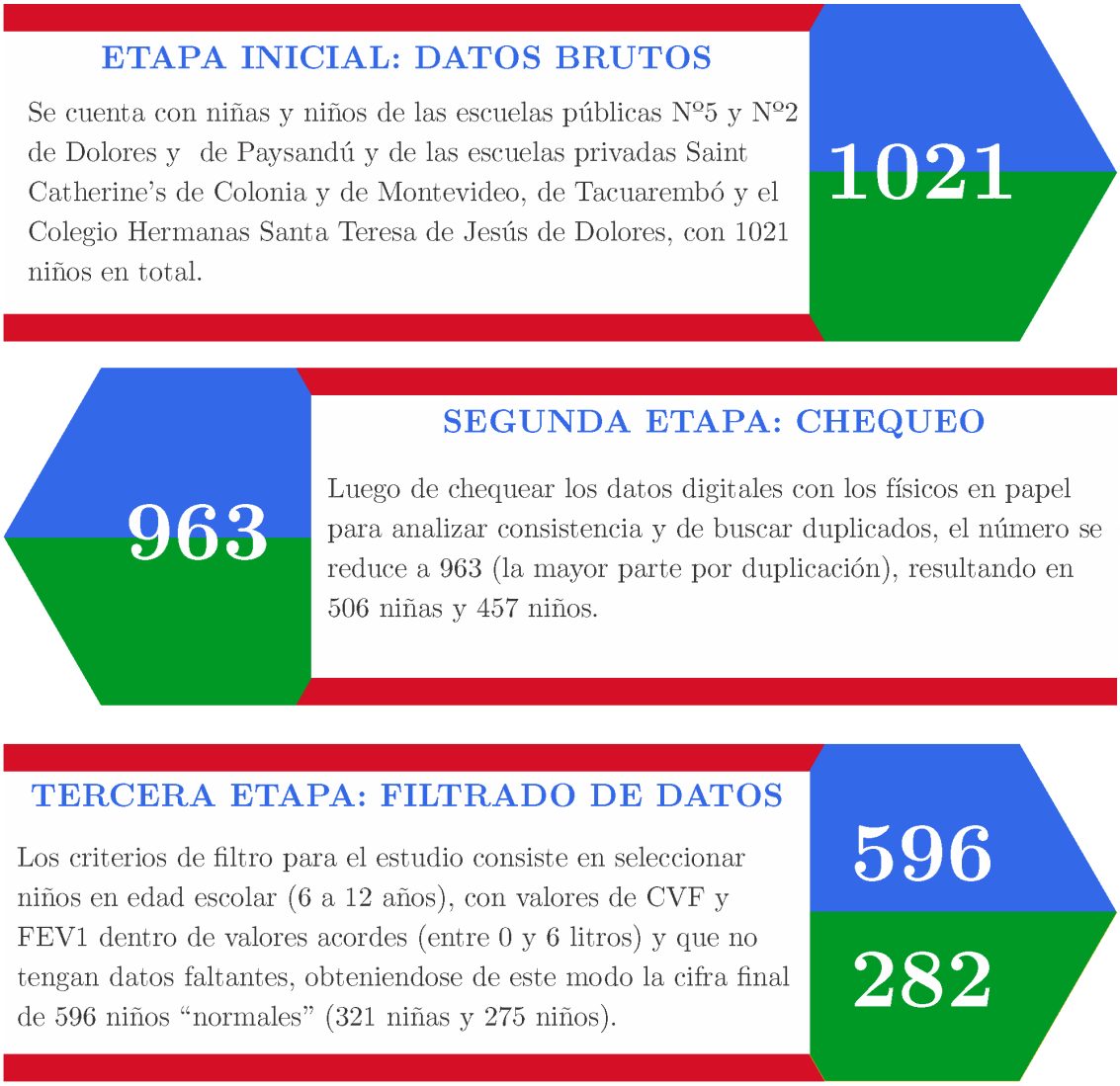

Figura 3.1: *Etapas de depuración del conjunto de datos, donde se muestra la cantidad de observaciones implicadas y la descripción de las mismas*

Los equipos médicos estaban constituídos por neumólogos pediatras que realizaban los estudios mediante dos espirómetros (Brentwood-Spiroscan 2000 y Fukuda) los cuales cumplían con las normas de ATS para estos registros y la presencia de enfermeras universitarias. La maestra de la clase del niño estaba presente durante la realización del estudio. Se utilizaron piezas bucales descartables para cada niño.

Los niños fueron pesados con ropas livianas en una balanza electrónica marca Sohenle Personal Scale 7306.00 (error +- 0.1kg) y se midió su talla (estatura) descalzos mediante un pediómetro digital Sohenle 5001 (error +- 0.5 cm) en un ambiente térmicamente adecuado.

Previamente se habían recabado datos sobre los antecedentes de los niños mediante un formulario escrito enviado a los padres.

El análisis de los datos se realizó con el programa R (R Core Team, 2017) a través de la UI (interfáz de usuario) RStudio (RStudio Team, 2016) utilizando las librerías *readr*(Wickham *et al.*, 2017b), *tibble*(Müller y Wickham, 2017), *tidyr*(Wickham, 2017), *dplyr*(Wickham *et al.*, 2017a), *ggplot2*(Wickham, 2009), *gamlss*(Rigby y Stasinopoulos, 2005), *ICSNP*(Nordhausen *et al.*, 2015) y *MASS*(Venables y Ripley, 2002).

Aspectos Éticos

El Consejo de Educación Primaria aprobó la realización del estudio en las escuelas públicas.

Un comité de notables de cada escuela privada aprobó el desarrollo del trabajo, explicándosele previamente el protocolo a seguir.

Se requirió la firma de cada padre aprobando la realización del estudio.

Descripción de los datos

Variables dentro del estudio

Las variables comprendidas en este estudio se listan a continuación en la Tabla 3.1.

Variable	Tipo de variable	Descripción
Edad	Contínua	Edad del niño expresada en años al momento del estudio.
Talla	Contínua	Talla del niño al momento de estudio. Expresada en centímetros.
Peso	Contínua	El peso expresado en kilogramos.
Sexo	Categórica Nominal	El sexo del niño, con valores F para femenino y M para masculino.
Alergicos	Categórica Nominal	Variable que refiere a antecedentes patológicos.
ContFab	Categórica Nominal	Presencia de contaminación ambiental por actividades industriales .
Fuman	Categórica Nominal	Si en la casa hay alguien que fuma, ya sea madre, padre, abuelos u otros.
Escuela	Categórica Nominal	Escuela a la cual pertenece el niño. Puede considerarse también como una variable geográfica.
CVF	Contínua	Capacidad Vital Forzada, expresada en litros (L).
FEV1	Contínua	Volumen Espiratorio Forzado en el primer segundo (FEV_1), expresado en litros (L).
FEF2575	Contínua	Flujo Espiratorio Forzado medido en la mitad de la espiración (FEF_{25-75}) ó mesoflujo.
PFE	Contínua	Pico de Flujo Espirométrico. Se mide en litros por minuto (L/min).
IGaensler	Contínua	La relación FEV_1/CVF, también conocido como Índice de Gaënsler.

Tabla 3.1: *Descripción de las variables espirométricas del estudio.*

Variable	Edad	Talla	Peso
Mínimo	6.10	107.00	17.00
1^{er} cuartil	8.02	127.00	27.50
Mediana	9.44	135.00	33.00
3^{er} cuartil	10.70	143.00	40.10
Máximo	12.00	173.0	82.10
Media	9.35	135.30	34.72
Varianza	2.64	128.85	101.83
Desvío estándar	1.63	11.35	10.10

Tabla 3.2: *Medidas de localización y dispersión para las variables antropométricas contínuas*

Sexo	Alergicos	ContFab	Fuman
466 Femenino (F)	596 No	818 No	527 No
412 Masculino (M)	282 Si	60 Si	351 Si

Tabla 3.3: *Frecuencias absolutas de las variables categóricas.*

Medidas de resumen

A continuación se muestran las medidas de resumen para las variables del estudio, tanto para las variables de respuesta como para las variables explicativas. En la Tabla 3.2 se muestra el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, el máximo, la media y la varianza de las variables antropométricas contínuas. En la Tabla 3.3 se muestra la distribución de frecuencias de las variables categóricas Sexo, Alergicos, ContFab y Fuman. En la Tabla 3.4 se muestran los mismos valores que en la Tabla 3.2 para las variables de respuesta de los parámetros de espirométricos.

Variable	CVF	FEV1	FEF2575	PFE
Mínimo	0.33	0.33	0.22	39.60
1^{er} cuartil	1.54	1.48	2.15	195.20
Mediana	1.89	1.78	2.61	235.10
3^{er} cuartil	2.27	2.14	3.08	280.60
Máximo	4.58	4.37	6.49	550.00
Media	1.93	1.82	2.66	244.90
Varianza	0.27	0.21	0.57	4763.04
Desvío estándar	0.52	0.46	0.76	69.01

Tabla 3.4: *Medidas de resumen para las variables espirométricas*

Relaciones entre variables antropométricas

Edad, Talla y Peso

Como primera observación (ver Figura 3.2), se puede decir que la relación entre la Talla y el Peso no es necesariamente lineal. También se puede observar que no hay simetría respecto de la línea azul, que representa la media local, es decir, no hay homocedasticidad.

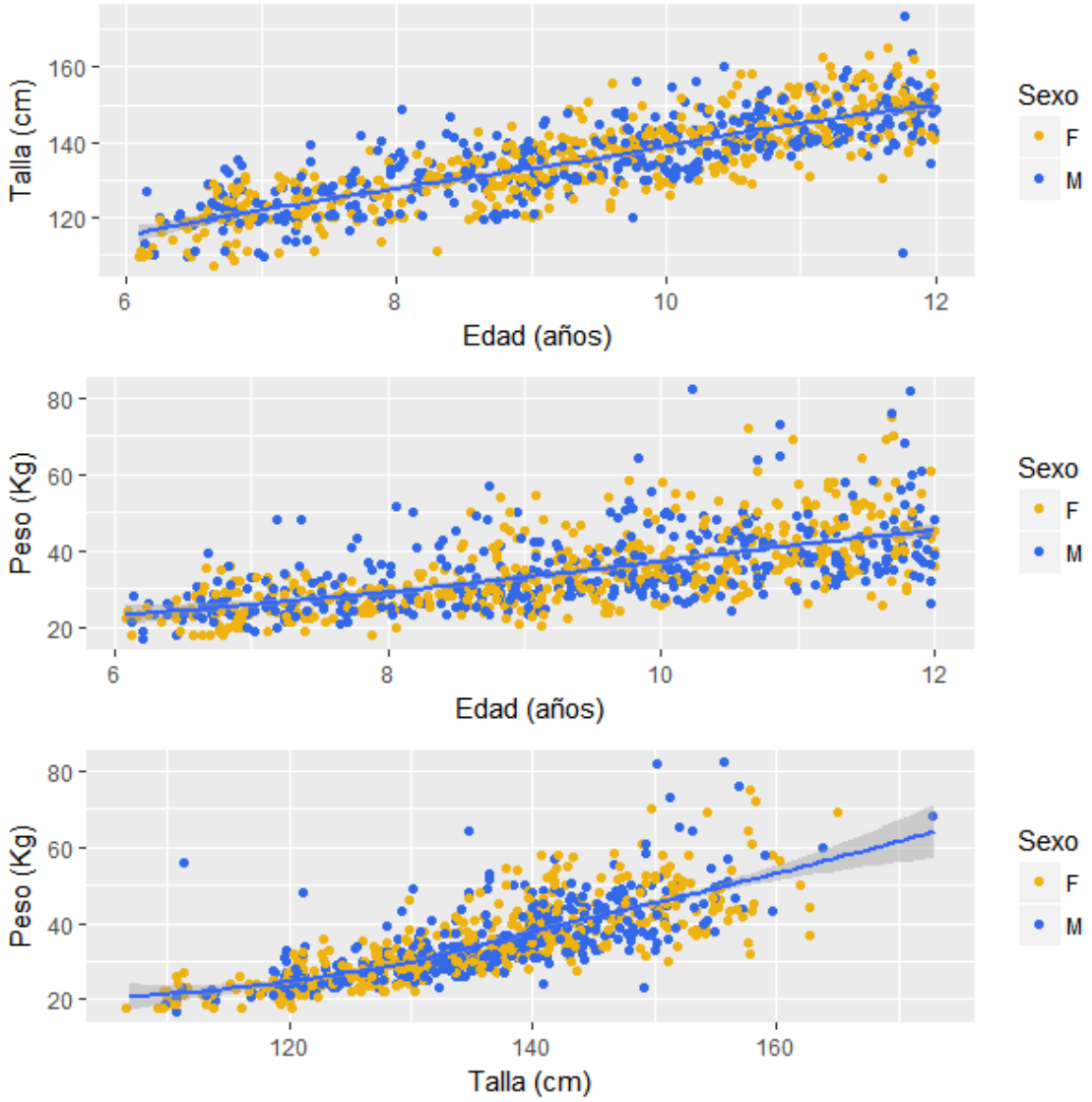

Figura 3.2: *Gráfico de dispersión entre las variables Edad, Talla y Peso, coloreado por Sexo, donde F refiere a femenino y M a masculino: (arriba) Edad y Talla; (centro) Edad y Peso; (abajo) Talla y Peso.*

En el caso del Peso y Edad la relación aparenta ser más lineal. Se observa que respecto a la media local hay una mayor dispersión cuando se incrementa la edad. La heterocedasticidad es mayor.

La relación entre Talla y Edad, podría ser lineal, con una menor dispersión que en los gráficos anteriores, aunque igual con cierta heterocedasticidad respecto a la media local.

Relación entre Capacidad Vital Forzada (CVF) y variables antropométricas.

A continuación se presentan los gráficos que relacionan la variable de la capacidad vital forzada (CVF) con las variables explicativas Talla, Edad y Peso (Figura 3.3).

En la Figura 3.3, se observa que las relaciones entre el CVF con la Edad y Talla es lineal. No se genera una curva leve en ninguno de los dos casos. Para la variable Peso se observa una concavidad negativa, con una mayor concentración en la primera mitad del gráfico que la relaciona con la Talla.

Relación entre Volumen Espiratorio Forzado en el primer segundo ($\mathbf{FEV_1}$) y las variables antropométricas.

Análogamente, se presentan los gráficos que relacionan la variable del volumen espiratorio forzado en el primer segundo (FEV_1) con las variables explicativas Talla, Edad y Peso.

La Figura 3.4 presenta los gráficos para la variable FEV_1, que muestran una similitud con los de la variable CVF. El gráfico central de la Figura 3.4 da muestras de una

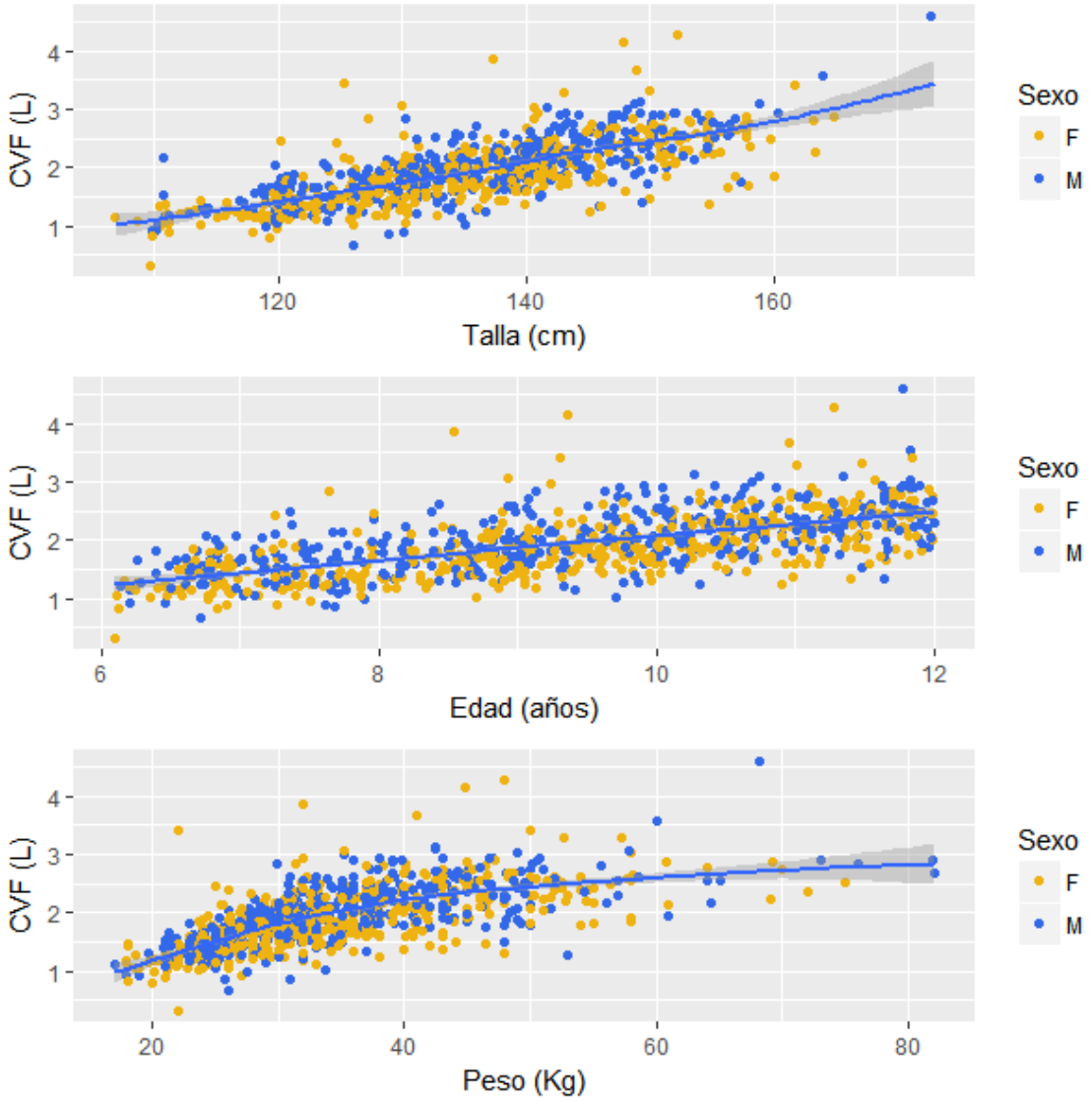

Figura 3.3: *CVF en relación a: (arriba) Talla; (centro) Edad y (abajo) Peso. La línea azul es un ajuste de la media local y en color gris aparece el intervalo de confianza del mismo.*

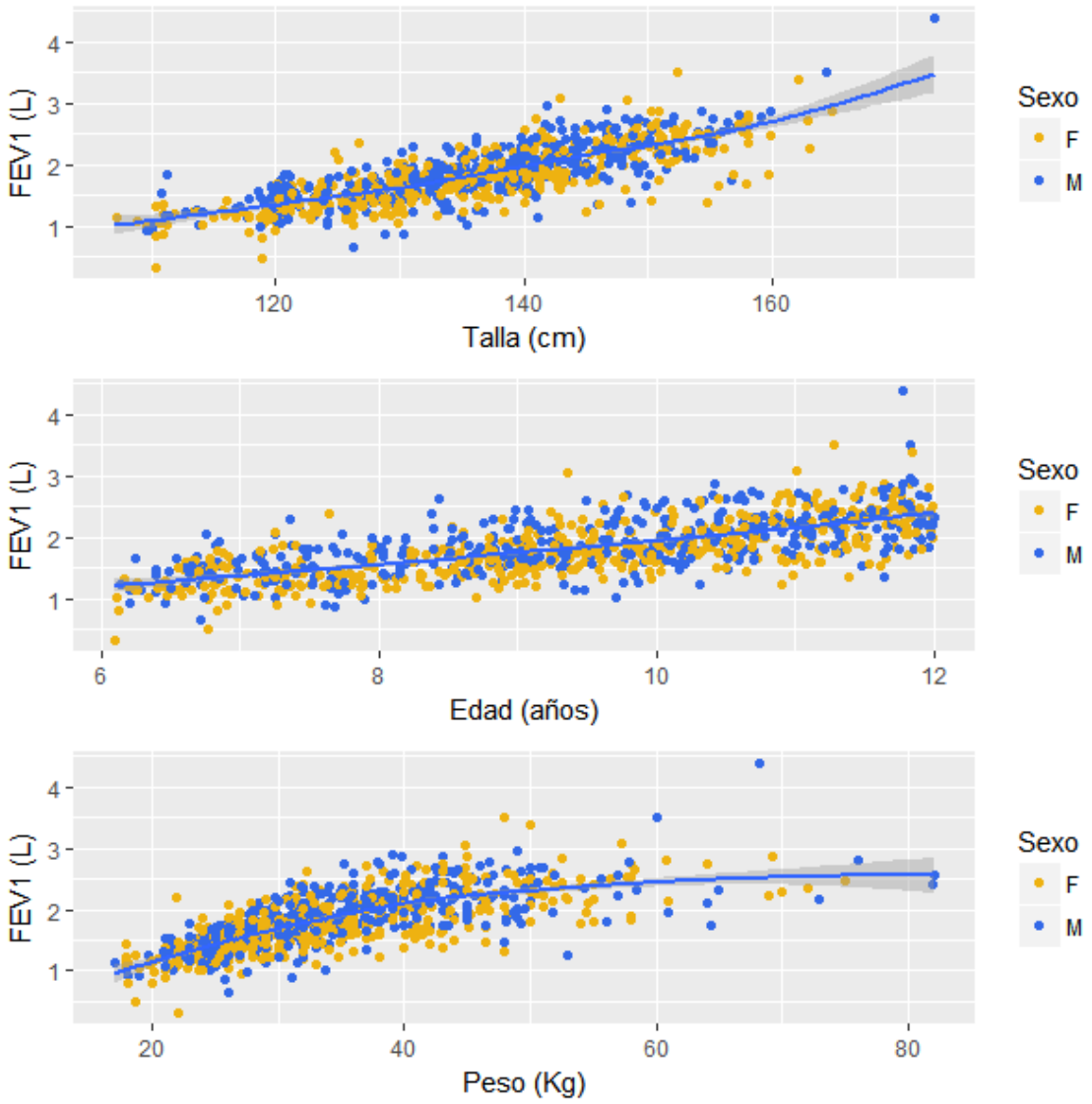

Figura 3.4: *FEV_1 en relación a: (arriba) Talla; (centro) Edad y (abajo) Peso.*

relación lineal entre el FEV_1 y la Edad, y la presencia de heterocedasticidad leve. En el gráfico superior se observa una leve concavidad positiva en la relación de FEV_1 con Talla. Mientras tanto, el gráfico inferior muestra una concavidad negativa para la variable Peso.

Las apreciaciones que se pueden obtener de ambas variables son similares.

Análisis de las variables espirométricas según niños alérgicos y niños normales

Se busca estudiar si los niños con antecedentes de patología respiratoria se diferencian de los niños que no la presentan. Para ello se comparan los valores de los parámetros respiratorios comprendidos dentro del estudio. La Figura 3.5 muestra la dispersión de las variables CVF, FEV_1, FEF_{25-75} y PFE en relación a la edad, expresada en años, donde los puntos verdes hacen referencia a los niños normales y los puntos rojos a los niños con antecedentes patológicos (alérgicos).

En las gráficas se observa que los puntos verdes y rojos aparecen mezclados más que tener alguna diferencia de nivel en sus valores. Se esperaría que los niños con antecedentes tuvieran algún tipo de disminución en los valores paramétricos, con valores por debajo de los normales. Esto no parece ocurrir en los casos de las variables CVF y FEV_1. Si se mira con detenimiento en la Figura 3.5, a partir de los 9 años de edad hasta los 11, los niños alérgicos parecen seguir lo esperable. Pero a simple vista no es fácil de apreciar.

La Figura 3.6 muestra las densidades de las variable CVF, FEV_1, FEF_{25-75} y PFE para los niños normales y alérgicos. Se puede observar que las diferencias entre ambas son puntuales. Se observa que hay similitudes entre niños alérgicos y normales en sus valores de CVF, como de FEV_1. En tanto, en los valores del FEF_{25-75} se observa, no

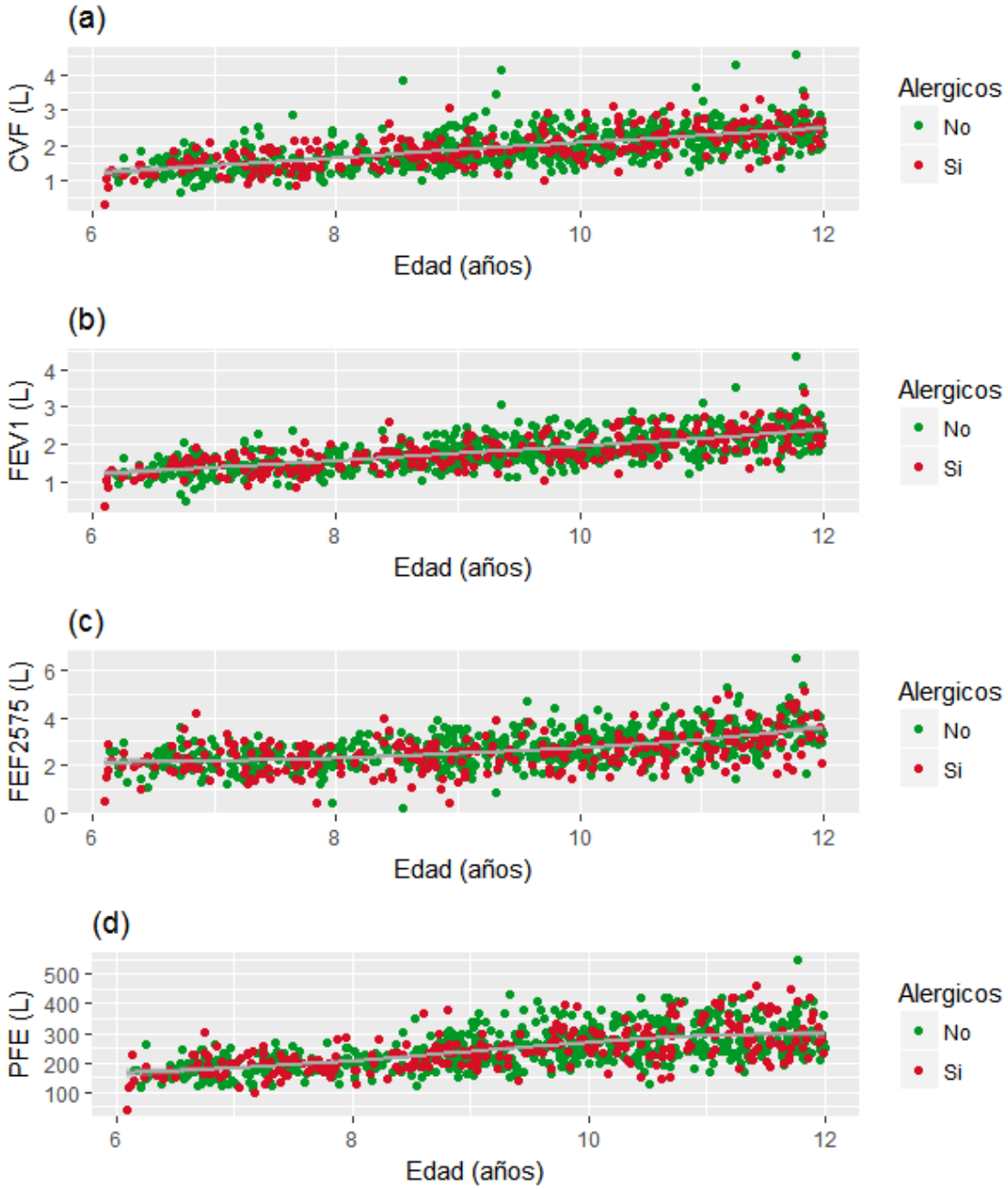

Figura 3.5: *Gráfico de (a) CVF, (b) FEV_1, (c) FEF_{25-75} y (d) PFE en función de la edad, donde se distinguen entre niños alérgicos (rojos) y normales (verdes).*

Tabla 3.5: *Comparación de variables espirométricas entre niños normales y alérgicos.*

Variable	Alérgicos		Normales	
	Media	Varianza	Media	Varianza
CVF	1.93	0.26	1.92	0.28
FEV_1	1.81	0.21	1.82	0.21
FEF_{25-75}	2.56	0.62	2.70	0.54
PFE	244.20	5144.63	245.20	4590.50

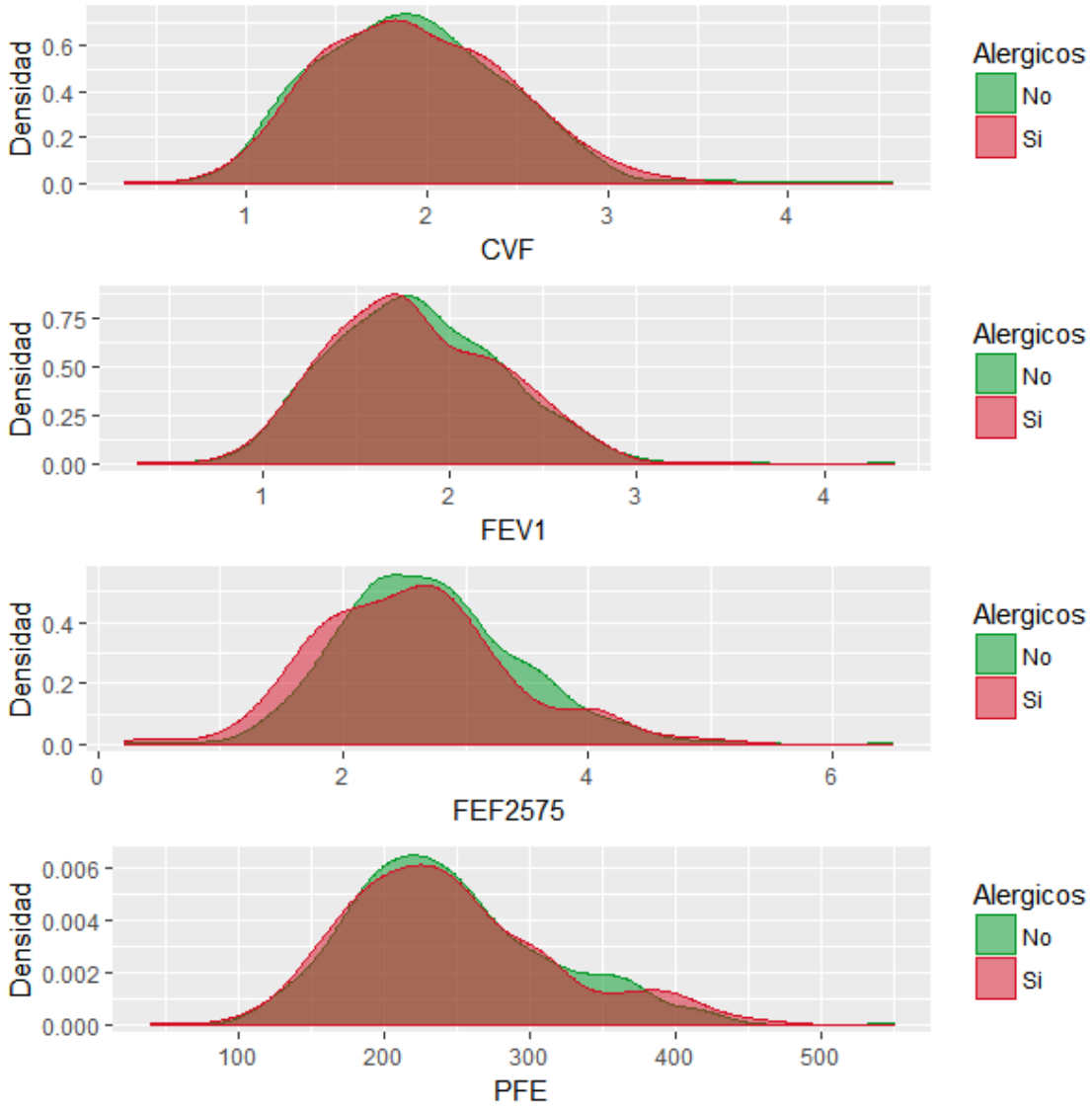

Figura 3.6: *Gráfico comparativo de densidades para los parámetros CVF, FEV_1, FEF_{25-75} y PFE de los niños alérgicos (rojo) y normales (verde).*

solo diferencias en la forma de la densidad, sino la existencia de un corrimiento en media, donde los niños alérgicos presentan valores menores que los niños normales.

Las variables CVF, FEV_1, FEF_{25-75} y PFE se obtienen de la misma maniobra de espirometría, por lo que las variables no son independientes, por ello se debe abordar el estudio como un problema multivariado. Con éste fin, se realiza un test de Hotelling T^2 modificado para dos muestras sin asumir que las matrices de covarianzas sean iguales. Supongamos que se tienen dos muestras p-variadas. El estadístico de prueba se define como

Grupo	Variable	CVF	FEV_1	FEF_{2575}	PFE
normales	CVF	0.28	0.23	0.18	18.39
	FEV_1	0.23	0.22	0.22	17.29
	FEF_{2575}	0.18	0.22	0.55	23.33
	PFE	18.39	17.29	23.33	4590.50
alérgicos	CVF	0.26	0.22	0.19	22.06
	FEV_1	0.22	0.21	0.25	20.87
	FEF_{2575}	0.19	0.25	0.63	27.72
	PFE	22.06	20.87	27.72	5144.64

Tabla 3.6: *Matriz de varianzas y covarianzas de los parámetros espirométricos CVF, FEV_1, FEF_{2575} y PFE para el grupo de los niños normales (arriba) y los niños alérgicos (abajo).*

$$T_u^2 = (\bar{\mathbf{X}}_\mathbf{1} - \bar{\mathbf{X}}_\mathbf{2})^\intercal \tilde{\mathbf{S}}^{-1} (\bar{\mathbf{X}}_\mathbf{1} - \bar{\mathbf{X}}_\mathbf{2})$$

con

$$\tilde{\mathbf{S}} = \frac{S_1}{n_1} + \frac{S_2}{n_2}$$

donde S_1 es la matriz de covarianzas de la población 1 y S_2 es la matriz de covarianzas de la población 2.

(James, 1954) sugiere que el estadístico de prueba sea comparado con $2h(\alpha)$, una distribución χ^2 corregida cuya forma es

$$2h(\alpha) = \chi^2(A + B\chi^2)$$

donde

$$A = 1 + \frac{1}{2p}\sum_{i=1}^{2}\frac{\left(tr\tilde{\mathbf{S}}^{-1} - \tilde{\mathbf{S}}_i\right)^2}{n_i - 1}$$

$$y$$

$$B = \frac{1}{p(p+2)}\left[\frac{1}{2}\sum_{i=1}^{2} tr\left(\tilde{\mathbf{S}}^{-1} - \tilde{\mathbf{S}}_i\right)^2 + \frac{1}{2}\frac{\left(tr\tilde{\mathbf{S}}^{-1} - \tilde{\mathbf{S}}_i\right)^2}{n_i - 1}\right]$$

Para llevar a cabo la prueba T^2 de Hotelling se utilizó la función `HotellingsT2` incluída en la librería `ICSNP`, que contiene herramientas de análisis multivariado no paramétrico.

Prueba de Hotelling T^2

Variables: CVF, FEV1, FEF2575, PFE

Estadístico de prueba	2.4842
Gr. Libertad Numerador	4
Gr. Libertad Denimonador	873
P-valor	0.04229

Tabla 3.7: *Resultado de la prueba T^2 de Hotelling para las variables CVF, FEV_1, FEF_{25-75} y PFE entre niños con antecedentes patología respiratoria (alérgicos) y sin antecedentes de patología (normales).*

El p-valor de la prueba (p-valor<0.05) en la Tabla 3.7 indica que no hay evidencia suficiente para decir que provienen de distribuciones iguales. En otras palabras, no podemos decir que los niños con antecedentes de patología respiratoria y los niños normales provengan de una misma población.

La evidencia indicaría que, estadísticamente, hay diferencias entre las poblaciones de niños normales y alérgicos, por lo que no se pueden considerar como un mismo grupo para el objetivo de encontrar curvas de referencia de los parámetros espirométricos.

A efectos de lograr las curvas de referencia de los distintos parámetros espirométricos,

se utilizarán solo a los niños normales, al igual que otros autores, para así también poder comparar con los distintos resultados de estudios internacionales.

Estado nutricional

A continuación se presentan los valores de Z-score de Talla e Índice de Masa Corporal (IMC) por Edad. Como instrumento de comparación de los niños en la muestra de este estudio con los valores de referencia a nivel mundial procedentes de la Organización Mundial de la Salud (OMS), se utiliza un macro para R del programa Anthro Plus (de Onis *et al.*, 2007), desarrollado para facilitar la aplicación de los valores referenciales OMS 2007 de 5 a 19 años para controlar el crecimiento de los niños y adolescentes en edad escolar.

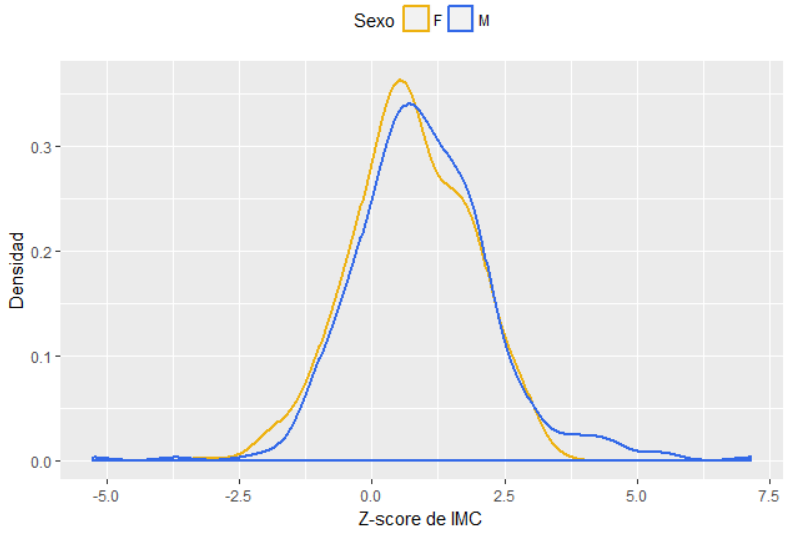

Figura 3.7: *Densidades de los Z-score del IMC por edad de niñas (naranja) y niños(azul).*

Se puede observar en la Figura 3.7 que no hay grandes diferencias en el IMC entre niñas y niños, sin embargo ambos presentan valores de Z-score en promedio mayores a 0, indicando que un 60.5 % de la muestra tiene el peso alterado, donde un 42.3 % presenta riesgo de sobrepeso y un 17.2 % presenta sobrepeso y obesidad, como lo

muestra la Tabla 3.8.

Tabla 3.8: *IMC por Edad*

Edad	N	Peso Alterado	Peso Normal	Riesgo de Sobrepeso	Sobrepeso y Obesidad
6 a 12	873	60,5	39,5	42,3	17,2
6	97	56,6	43,4	41,2	14,4
7	117	56,4	43,6	41	15,4
8	141	66,7	33,3	43,3	23,4
9	176	63,5	36,5	46	16,4
10	169	62,8	37,2	44,4	16,6
11	172	54,7	45,3	36,6	16,3
12	1	100	0	100	0

En cuanto a la Talla, en la Tabla B.3 se puede ver que un 25.7 % de los niños presenta la Talla alterada, donde el 24.1 % es más alto de lo normal para su edad.

Distribución de CVF y FEV_1

La meta es encontrar familias de distribuciones para las variables de respuesta CVF y FEV_1 para utilizarlas en su modelización. Es necesario entonces encontrar una distribución paramétrica que se ajuste a los datos para cada una.

Una de las formas para llevar a cabo este cometido es a través de la función `fitdistr()` de la librería **MASS**. El paquete **gamlss**, dentro de sus funciones, incluye la función `fitDist()` que sirve como alternativa para el ajuste de una familia de distribución a los datos.

Diferencias entre las funciones `fitDist()` y `fitdistr()`

La gran diferencia entre ambas funciones, es que la función `fitdistr()` (**MASS**) necesita de antemano establecer la distribución (por ejemplo, decirle que es normal)

y ésta estima los parámetros de la distribución (media y desvío estándar, para el ejemplo). La función de la librería **gamlss** es más flexible, ya que a partir de los datos, estima la familia y todos los parámetros correspondientes.

Prueba de robustez

Para estudiar la robustez de las distribuciones ajustadas con la función `fitDist()` del paquete **gamlss**, y la dependencia de los resultados en función de los datos, se realiza un proceso iterativo donde se seleccionan muestras de los datos con un tamaño de muestra del 80 % del número de observaciones y se analiza cual es la familia de distribuciones que más se repite en las N iteraciones, y se calcula la variabilidad de los parámetros estimados para cada caso. Se presentan medidas de ajuste basadas en el AIC.

Explicación de la rutina para la prueba de robustez.

1. Establece un número de iteraciones en primera instancia (`iter`).

2. Luego fija un valor de tamaño de muestra del total de datos (`nS`).

3. Crea un *data frame* para guardar los resultados de la rutina, donde se va a guardar la familia resultante de la función `fitDist()` (con el menor GAIC, con penalización k=2), luego el valor del AIC, y el valor estimado de los parámetros de la familia de distribución, que van de 2 a 4 (`Rob_res`).

4. Para cada iteración, se selecciona una muestra de tamaño `nS` (`muestra`).

5. Aplica la función `fitDist()` a la muestra para la variable analizada CVF y guarda al objeto gamlss resultante como `AjDist`.

6. El valor de `df.fit`, es el equivalente a los grados de libertad del ajuste, que no es otra cosa que la cantidad de parámetros que tiene la familia.

7. se almacenan los valores de los cuatro posibles parámetros que pueden caracterizar a la distribución, la familia y el AIC dentro de `Rob_res`.

Resultados para CVF

A continuación se muestran los resultados obtenidos para la variable CVF de la prueba de robustez de la función `fitDist()`, para una cantidad de mil iteraciones, considerando a todos los niños normales.

En la Tabla 3.9 se muestran las distintas familias de distribución que fueron ajustadas, la cantidad de parámetros, su frecuencia relativa, el AIC promedio, junto a la media de cada parámetro de distribución según corresponda.

Cabe recordar, que en cada iteración, la familia ajustada es aquella que tuvo el menor GAIC con penalización k, en este caso con un valor de $k = 2$.

Familia	k	Frec.Rel	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
SEP4	4	0.420	717.45	1.88	0.76	3.52	1.66
exGAUS	3	0.247	720.68	1.59	0.40	0.34	-
ST5	4	0.212	724.27	-1.72	0.06	0.25	0.01
GT	4	0.079	724.24	1.92	0.76	0.57	13.09
JSU	4	0.039	723.85	1.93	0.53	43.51	4.92

Tabla 3.9: *Resultado de la prueba de robustez para la variable CVF.*

La Figura 3.8 es una representación gráfica de la frecuencia relativa para cada familia de distribución.

Como se puede apreciar en la Figura 3.9, los ajustes son similares, presentando, a simple vista, leves diferencias entre ellos. De aquí la precaución que se debe tener al utilizar este método de ajuste de distribuciones. La que tiene un mejor ajuste es la familia de distribución SEP4, que es la que tiene más frecuencia relativa y el menor AIC medio.

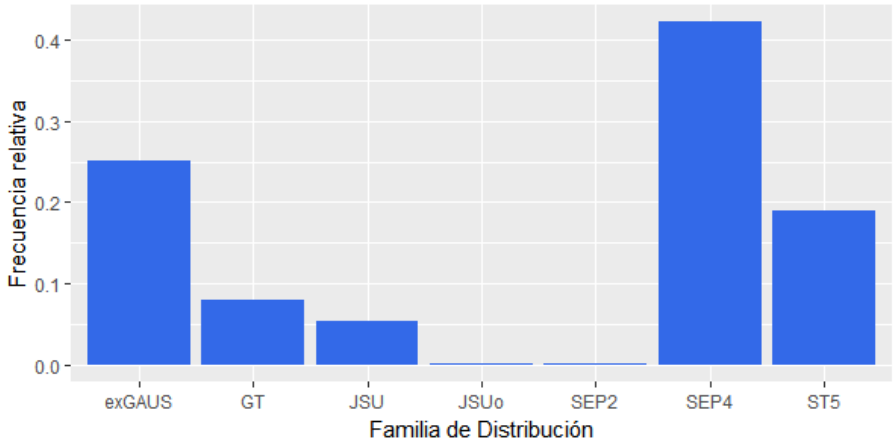

Figura 3.8: *Distribución en el muestreo de las familias de distribución para la variable CVF*

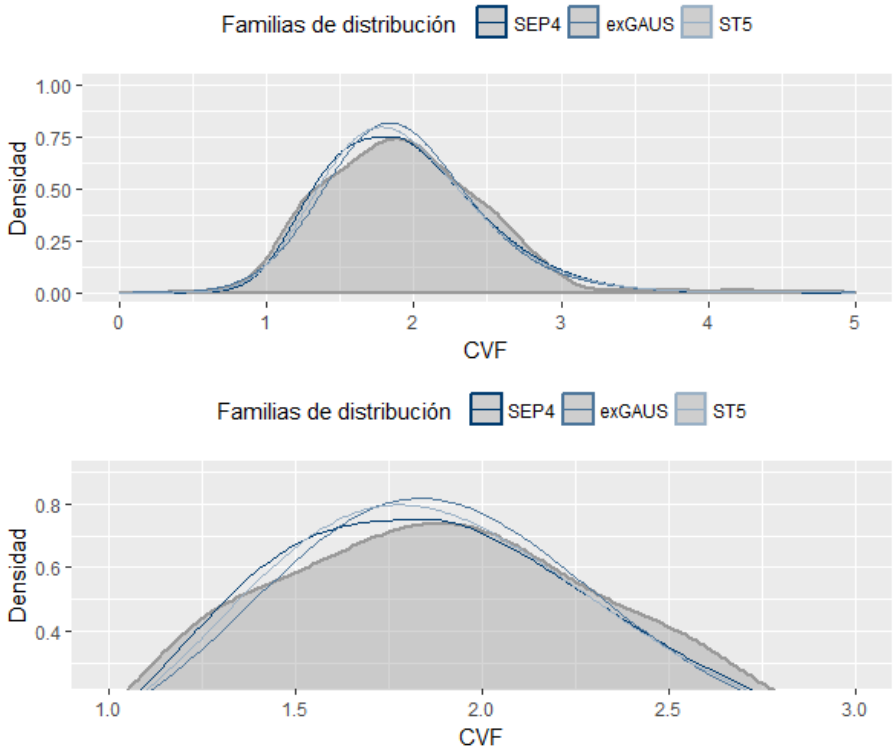

Figura 3.9: *Densidades ajustadas para CVF: skewed power exponential type 4 (SEP4); exponential Gaussian (exGAUS); skewed t type 5 (ST5). El gráfico inferior es un zoom de la zona central.*

Resultados para niñas

A continuación se muestran los resultados obtenidos del proceso iterativo para el conjunto de datos de niños normales de sexo femenino.

En la Tabla 3.10 se muestran las distintas familias de distribución que fueron ajustadas para este caso, la cantidad de parámetros, su frecuencia relativa, el AIC promedio, junto a la media de cada parámetro de distribución según corresponda.

Familia	k	Frec.Rel	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
SEP4	4	0.676	391.10	1.82	0.77	5.34	1.57
RG	2	0.285	395.59	1.62	0.45	-	-
SN2	3	0.039	393.04	1.37	0.41	2.02	-

Tabla 3.10: *Resultado de la prueba de robustez para la variable CVF de niños normales de sexo femenino .*

La Figura 3.10 es una representación gráfica de la frecuencia relativa para cada familia de distribución.

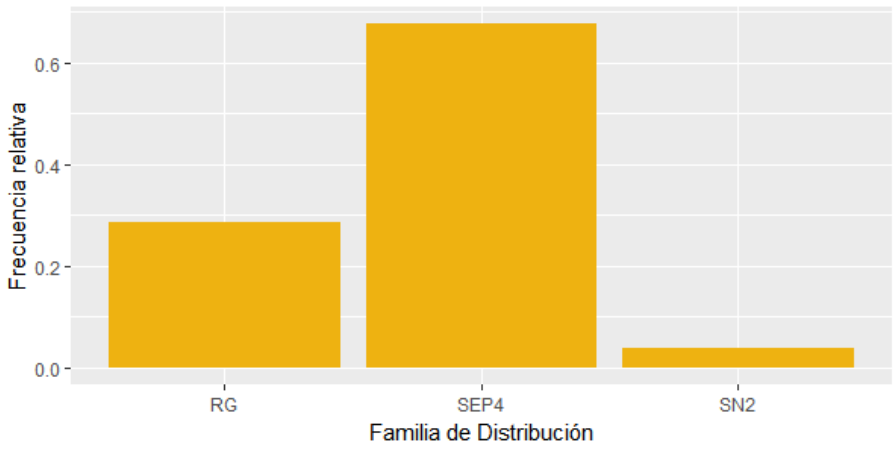

Figura 3.10: *Gráfico de frecuencia de familias de distribución para la variable CVF de los niños normales de sexo femenino*

Se puede observar en la Figura 3.11 que en ningún caso, no se logra captar en su totalidad la densidad de los datos. Sin embargo, la que se ajusta mejor parecería ser la familia de distribución SEP4, aunque es similar a la familia RG. La familia

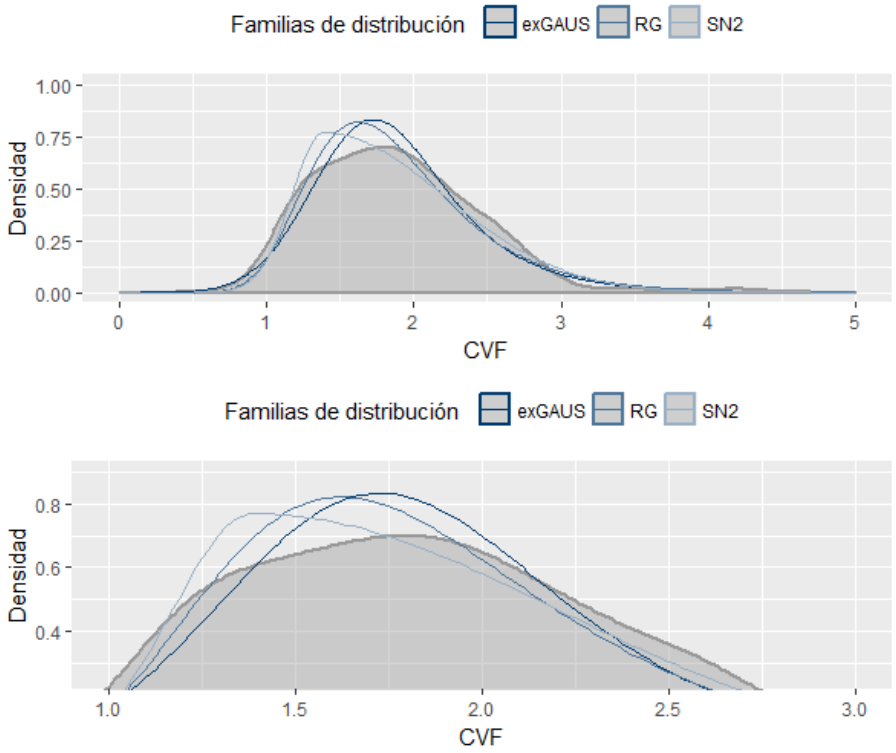

Figura 3.11: *Densidades ajustadas para CVF de niños de sexo femenino: exponential Gaussian (exGAUS); Revers Gumbel (RG); skewed Normal type 2 (SN2). El gráfico inferior es un zoom de la zona central.*

SEP4 tiene menor AIC y cuenta con cuatro parámetros, lo que permitiría una mayor flexibilidad.

Resultados para niños

A continuación se muestran los resultados obtenidos del proceso iterativo para el conjunto de datos de niños normales de sexo masculino de la variable CVF.

En la Tabla 3.11 se muestran las distintas familias de distribución que fueron ajustadas para este caso, la cantidad de parámetros, su frecuencia relativa, el AIC promedio, junto a la media de cada parámetro de distribución según corresponda.

Familia	k	Frec.Rel	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
exGAUS	3	0.626	320.30	1.70	0.41	0.29	-
GT	4	0.199	319.53	1.98	0.72	1.06	4.96
NO	2	0.142	303.25	1.98	0.48	-	-
PE	3	0.008	288.94	1.98	0.47	2.92	-
PE2	3	0.008	300.89	1.98	0.76	2.78	-
SEP2	4	0.008	292.61	2.48	0.82	-4.30	3.50
SN2	3	0.005	298.06	1.85	0.46	1.19	-
SEP1	4	0.003	299.09	1.36	0.92	2.98	3.41
SEP4	4	0.001	303.52	1.97	0.77	3.49	2.39

Tabla 3.11: *Resultado de la prueba de robustez para la variable CVF de niños normales de sexo masculino.*

La Figura 3.12 es una representación gráfica de la frecuencia relativa para las seis familias de distribución con mayor frecuencia relativa.

Se observa en la Figura 3.13 que los datos para los niños de sexo masculino se acerca mucho a una distribución normal, sin embargo, no capta de manera correcta la media. Por otra parte, la familia exGAUS si parece hacerlo, y pareciera tener un mejor ajuste general.

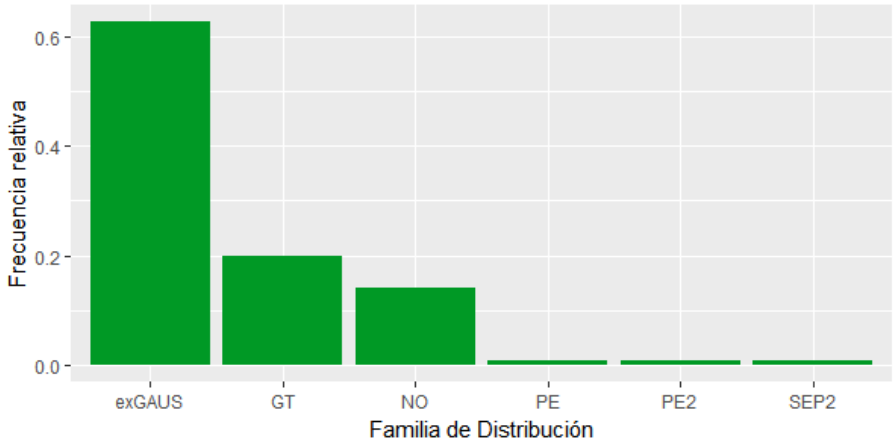

Figura 3.12: *Gráfico de frecuencia de familias de distribución para la variable CVF de los niños normales de sexo masculino*

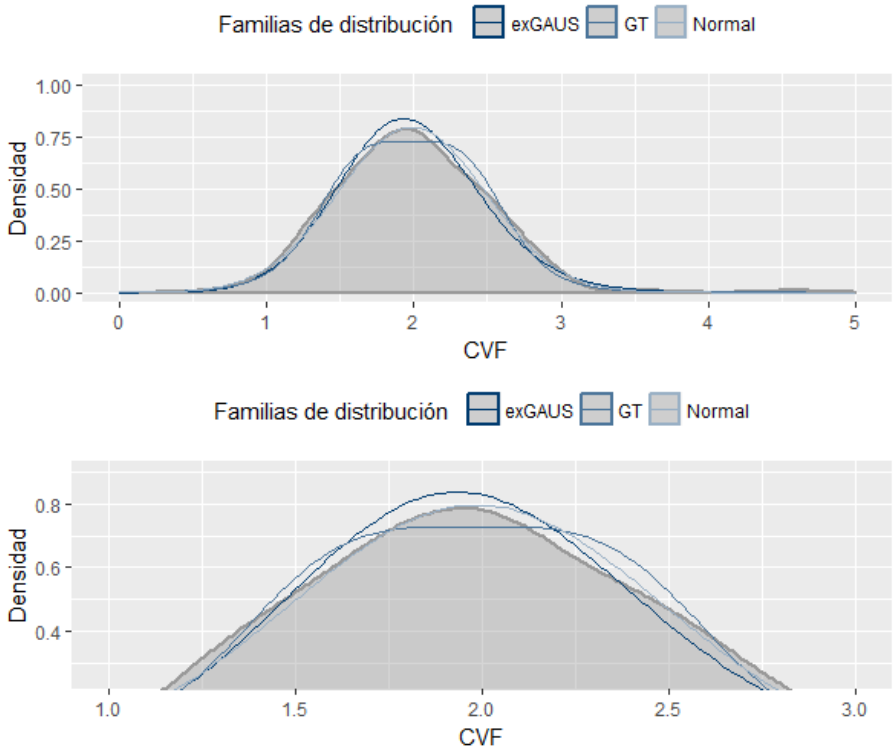

Figura 3.13: *Densidades ajustadas para CVF de niños de sexo masculino: exponential Gaussian (exGAUS); Generalized t (GT); Normal (Normal). El gráfico inferior es un zoom de la zona central.*

Resultados para FEV_1

Primero se muestra los resultados correspondientes al conjunto de todos los niños normales. La Tabla 3.12 muestra los resultados de la prueba de robustez. En la tabla aparece el nombre abreviado, el número de parámetros que caracterizan a la distribución (k), la frecuencia relativa, el AIC medio, y las medias de cada uno de los parámetros correspondiente a cada familia.

Familia	k	Frec.Rel	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
exGAUS	3	0.640	613.13	1.54	0.37	0.28	-
SN2	3	0.102	597.82	1.67	0.44	1.25	-
ST5	4	0.094	609.69	-1.20	0.14	0.21	0.01
SEP4	4	0.065	599.28	1.79	0.67	3.29	1.75
SN1	3	0.028	596.87	1.39	0.62	1.69	-
JSU	4	0.022	606.17	1.83	0.47	44.51	5.33
SEP2	4	0.015	582.87	1.30	0.76	3.93	2.58
GT	4	0.014	604.09	1.82	0.68	1.33	4.50
SEP1	4	0.004	571.86	2.19	0.69	-1.24	3.65
SHASHo	4	0.004	579.73	1.69	0.58	0.21	1.22
SST	4	0.004	608.88	1.82	0.47	1.37	21.78
ST3	4	0.003	609.75	1.59	0.41	1.38	22.60
SHASH	4	0.002	596.05	1.80	0.57	1.26	1.12
SHASHo2	4	0.002	587.05	1.71	0.48	0.18	1.20
SEP3	4	0.001	600.68	1.67	0.52	1.26	2.53

Tabla 3.12: *Resultado de la prueba de robustez para la variable FEV_1 de niños normales.*

La Figura 3.14 muestra la frecuencia relativa de las 6 familias con mayor frecuencia relativa que se ajustaron en el proceso iterativo.

En la Figura 3.15 se observa que las distribuciones exGAUS y ST5 son muy similares, pero en detalle, se puede ver que la familia que ajusta mejor la zona media es la exGAUS, ya que la familia ST5 presenta una leve inclinación hacia la izquierda.

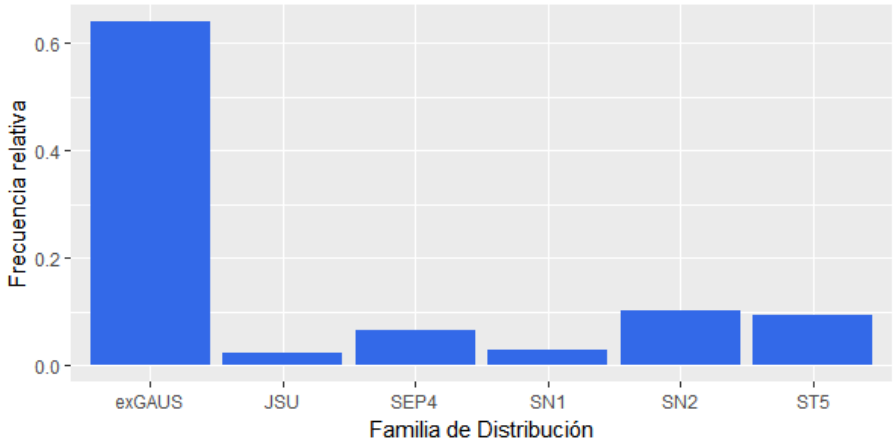

Figura 3.14: *Gráfico de frecuencia de familias de distribución para la variable* FEV_1

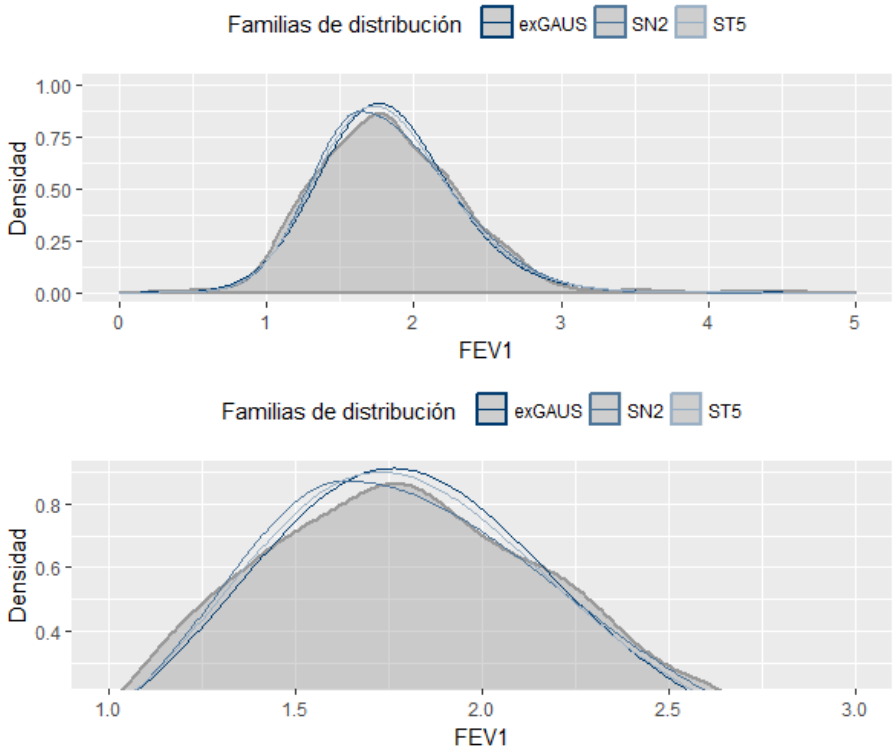

Figura 3.15: *Densidades ajustadas para* FEV_1*: exponential Gaussian (exGAUS); skewed Normal type 2 (SN2); skewed t type 5 (ST5). El gráfico inferior es un zoom de la zona central.*

Resultados para niñas

A continuación, en la Tabla 3.13 se muestran los resultados obtenidos del proceso iterativo para el conjunto de datos de niños normales de sexo femenino de la variable FEV_1.

Familia	k	Frec.Rel	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
SN2	3	0.719	328.25	1.55	0.43	1.38	-
SEP4	4	0.117	316.50	1.75	0.69	4.59	1.82
SN1	3	0.076	325.03	1.30	0.66	2.10	-
SEP3	4	0.038	320.66	1.50	0.51	1.51	2.69
SEP2	4	0.017	312.83	1.40	0.79	3.94	2.92
SHASHo2	4	0.013	320.51	1.55	0.48	0.40	1.38
SEP1	4	0.009	314.87	1.80	0.72	0.51	3.40
SHASH	4	0.005	305.53	1.72	0.90	2.07	1.48
SHASHo	4	0.005	314.15	1.57	0.63	0.37	1.33
GT	4	0.001	333.14	1.81	0.72	0.89	6.86

Tabla 3.13: *Resultado de la prueba de robustez para la variable FEV_1 de niños normales de sexo femenino.*

La Figura 3.16 muestra la frecuencia relativa de las 6 familias con mayor frecuencia relativa que se ajustaron en el proceso iterativo.

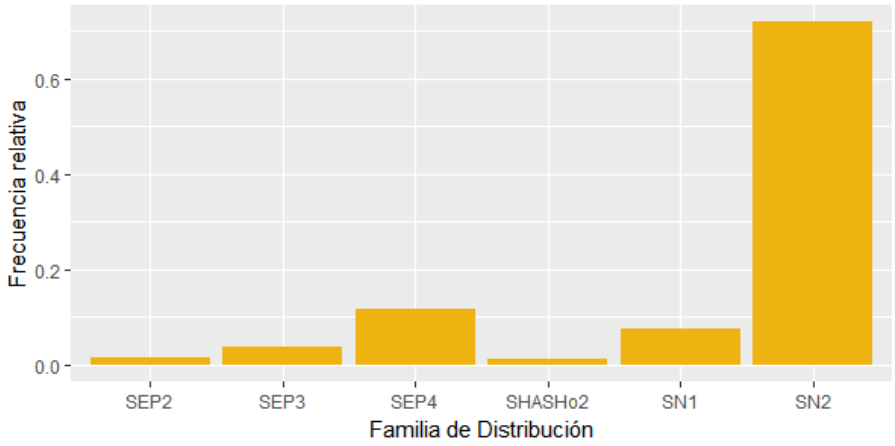

Figura 3.16: *Gráfico de frecuencia de familias de distribución para la variable FEV_1 de los niños normales de sexo femenino*

En la Figura 3.17 se observa que la familia de distribución SN2, aquella con mayor frecuencia relativa, no parece ajustar bien en la zona media, lo que si logra hacer la familia de distribución SEP4 que tiene un ajuste general más cercano.

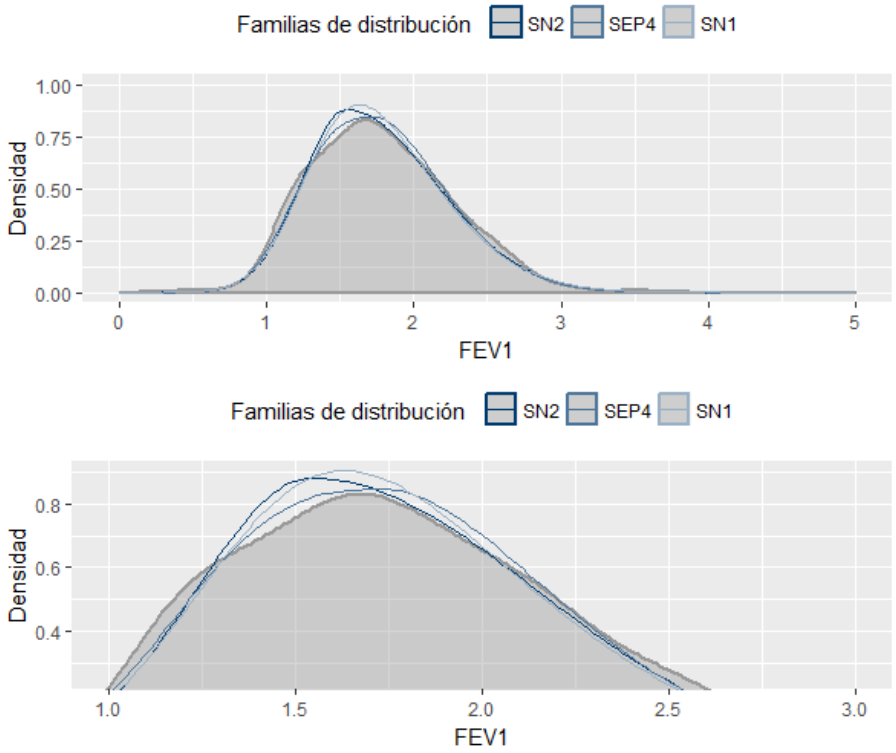

Figura 3.17: *Densidades ajustadas para FEV_1 de niños de sexo femenino: skewed Normal type 2 (SN2); skewed power Exponential type 4 (SEP4); skewed Normal type 1 (SN1). El gráfico inferior es un zoom de la zona central.*

Resultados para niños

A continuación, en la Tabla 3.14 se muestran los resultados obtenidos del proceso iterativo para el conjunto de datos de niños normales de sexo masculino de la variable FEV_1.

Familia	k	Frec.Rel	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
exGAUS	3	0.798	279.40	1.60	0.36	0.28	-
NO	2	0.089	263.25	1.88	0.44	-	-
GT	4	0.033	278.98	1.88	0.62	0.89	4.94
SN2	3	0.032	266.40	1.76	0.43	1.19	-
SN1	3	0.021	264.72	1.48	0.59	1.53	-
ST4	4	0.008	270.83	1.85	0.40	138.95	5.41
PE	4	0.007	250.61	1.88	0.43	2.69	-
SEP2	4	0.004	252.79	1.98	0.64	-0.18	2.91
PE2	4	0.003	243.06	1.88	0.67	2.73	-
SEP1	4	0.003	240.07	1.69	0.81	1.07	3.89
LO	2	0.001	258.23	1.87	0.24	-	-
SEP4	4	0.001	256.12	1.86	0.66	3.42	1.92

Tabla 3.14: *Resultado de la prueba de robustez para la variable FEV_1 de niños normales de sexo masculino.*

La Figura 3.18 muestra la frecuencia relativa de las 6 familias con mayor frecuencia relativa que se ajustaron en el proceso iterativo.

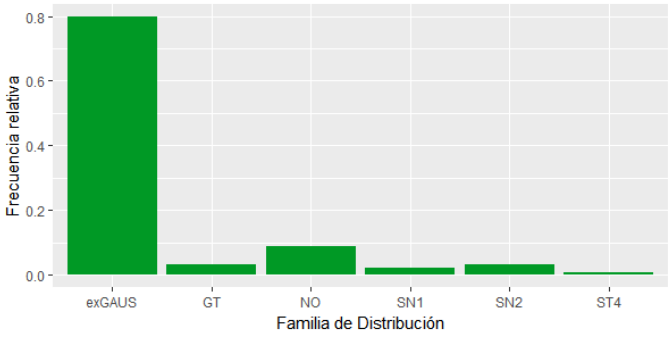

Figura 3.18: *Gráfico de frecuencia de familias de distribución para la variable FEV_1 de los niños normales de sexo masculino*

Se observa en la Figura 3.19 que la familia de distribución exGAUS es la que mejor ajusta la zona media, ya que la media de la distribución normal presenta un leve

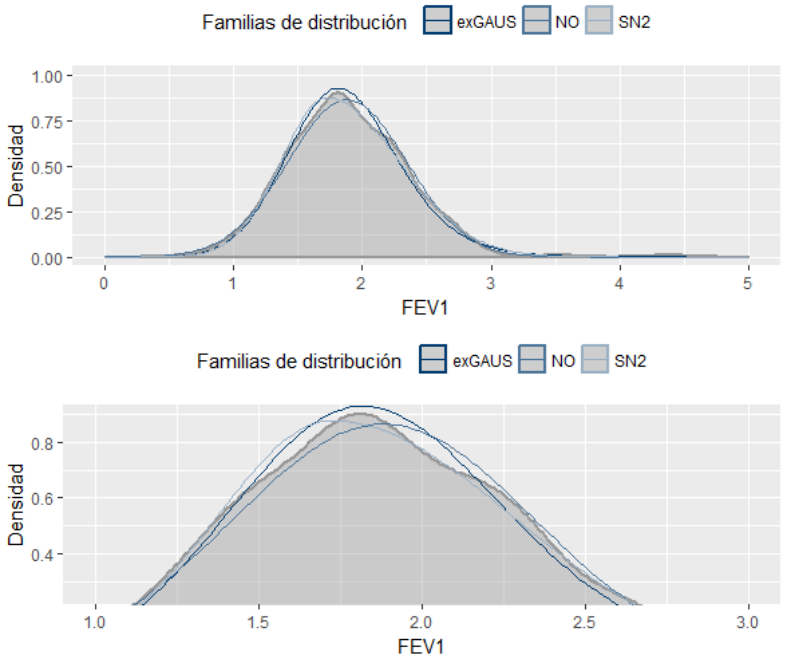

Figura 3.19: *Densidades ajustadas para FEV_1 de niños de sexo masculino: exponential Gaussian (exGAUS); Normal (NO); skewed Normal type 2 (SN2). El gráfico inferior es un zoom de la zona central.*

corrimiento hacia la derecha y la familia de distribución SN2 presenta un corrimiento de la media a la izquierda.

Modelización de CVF y FEV_1

La estrategia que se adoptó para la modelización de las variables espirométricas CVF y FEV_1 es la siguiente:

- Separar el conjunto de datos de los niños normales en dos: uno de entrenamiento y otro de validación del modelo, con una relación de 0.8 y 0.2 respectivamente.

- Utilizar los resultados de las pruebas de iteración para seleccionar la o las familias de distribución y ajustar modelos con la función `gamlss()`, utilizando el conjunto de entrenamiento.

- Luego utilizar el subconjunto de validación para hacer predicción y comparar los distintos modelos.

En cada escenario, habiendo elegido la familia de distribución, se debe modelizar cada uno de los parámetros presentes en ésta. Se plantean tres alternativas en cuanto a las distribuciones; utilizar la distribución normal, aquella que a través del proceso de iteración tuvo mayor frecuencia relativa y la familia BCPE (Box-Cox Power Exponential).

Los modelos, tanto para la variable CVF como para la variable FEV_1, se ajustaron con la siguiente estrategia:

1. Se ajustan modelos con la función `gamlss()` con las distintas familias de distribución mencionadas con el término `pb(Talla)` en el parámetro de localización μ (con el resto de los parámetros constantes).

2. Partiendo del modelo con mejor ajuste, se le incluye un término de suavizado `pb()` con la variable `Edad`, y utilizando la función `drop1()` con penalización SBC, determinar si dicha inclusión es estadísticamente importante.

Mediana μ

Para predecir el parámetro μ, se opta por seguir la forma general:

$$\mu = a_\mu + pb(Talla, edf_{\mu T}) + pb(Edad, edf_{\mu E}) \tag{3.6.1}$$

donde `pb()` es la forma de expresar en R un spline Beta penalizado, donde $edf_{\mu T}$ y $edf_{\mu E}$ son los grados de libertad efectivos del término de suavizado para Talla y Edad respectivamente. El término a_μ es una constante.

Se incluirá la variable Sexo como factor regresor para el caso del modelo global.

Variabilidad σ

La modelización del parámetro σ, se hace en términos de Talla y Edad utilizando el enlace *log*, de la forma:

$$log(\sigma) = a_\sigma + pb(Talla, edf_{\sigma T}) + pb(Edad, edf_{\sigma E}) \tag{3.6.2}$$

El término a_σ es una constante

Asimetría ν y curtosis τ

El parámetro de asimetría, ν, se opta por modelizarlo de la forma:

$$\nu = a_\nu + pb(Talla, edf_{\nu T}) + pb(Edad, edf_{\nu E}) \tag{3.6.3}$$

mientras tanto, el parámetro τ se estructura de la siguiente manera:

$$log(\tau) = a_\tau + pb(Talla, edf_{\tau T}) + pb(Edad, edf_{\tau E}) \tag{3.6.4}$$

donde los términos a_ν y a_τ son constantes.

Para comparar los modelos globales con los de cada sexo, se plantea de la siguiente forma:

- **Modelo GAMLSS con variable Sexo**: La idea es hacer un modelo GAMLSS para las variable de respuesta CVF y FEV_1, con la inclusión de la variable Sexo dentro del mismo, en los parámetros de distribución que sean necesarios.
- **Modelo GAMLSS por sexo**: Separar a los niños normales por sexo y luego ajustar un modelo para las variables CVF y FEV_1 para los niños y otro para las niñas.

Modelos GAMLSS

Antes de dar paso a la presentación de la modelización de cada parámetro espirométrico en sus distintos escenarios, se muestra la nomenclatura de los modelos dentro de los mismos.

Los distintos modelos siguen una estructura de la forma **m_parámetro espirométrico_abc**.

m refiere a la clase de objeto que representa, en este caso es un modelo.

parámetro espirométrico hace referencia a cual de las dos variables de respuesta es modelizada, donde cvf refiere a CVF y fev al FEV_1.

a el primer número se asocia con la característica de si es un modelo global (0), de niñas (1) o de niños (2).

b es un número que hace referencia al nivel de desarrollo de un modelo; es decir, el modelo base es representado por 0, y los siguientes modelos con agregados

prosigue la serie con 1, 2, 3,..., hasta el modelo final. Por ejemplo, un modelo base con un término con una variable tendrá un 0, y si a ese modelo base se le agrega un término más con otra variable, tendrá un 1, y así.

c el último número hace referencia a la familia de distribución utilizada para el modelo, donde el valor 1 se relaciona con la distribución normal, 2 con la que mayor frecuencia obtuvo en las pruebas iterativas y 3 a la distribución BCPE.

Variable de respuesta: CVF

		Predictor lineal para μ							
Modelo	Distribución	Talla	$edf_{\mu T}$	Edad	$edf_{\mu E}$	Sexo	df	*deviance*	SBC
m_cvf_001	NO	pb()	6.8429	-	-	-	7.84	391.90	440.26
m_cvf_002	SEP4	pb()	2.0034	-	-	-	5	330.80	361.65
m_cvf_003	BCPE	pb()	2.0036	-	-	-	5	304.41	335.26
m_cvf_013	BCPE	pb()	2.0044	-	-	Si	6	279.22	316.24
m_cvf_023	BCPE	pb()	2.0042	pb()	2.0116	Si	7	275.61	318.87

Tabla 3.15: *Diferentes modelos GAMLSS para CVF, con predictor lineal para la mediana μ, los grados de libertad del ajuste, y los valores de deviance y SBC, donde cada fila es un modelo separado.*

En la Tabla 3.15 se observa que un modelo con la familia de distribución BCPE ajusta mejor que uno con la familia SEP4 o NO (normal). También que la diferencia de un modelo que incluye la variable Edad y uno sin ella, no es significativa. Con un criterio conservador, con el SBC (GAIC(k=log(n))), la inclusión de la variable Edad aumenta el indicador. Además, se utiliza un grado más de libertad con la inclusión. Teniendo en cuenta lo anterior, se opta por el modelo con las variables Talla y Sexo.

	df	SBC	LRT	Pr(Chi)
Eliminaciones de un solo término para μ				
ninguno		318.87		
pb(Talla)	1.0096	463.59	150.940	< 2.2e-16
pb(Edad)	1.0114	316.24	3.605	0.0586
Sexo	1.0025	336.56	23.869	1.037e-06

Tabla 3.16: *Test de eliminación de variables en el parámetro μ del modelo m_cvf_023.*

En la Tabla 3.16 se muestra las pruebas de eliminación de los componentes en el modelo m_cvf_023, donde muestra los grados de libertad utilizados, *df* (calculados

como la suma de los grados efectivos de libertad del término de suavizado *edf* y los grados de libertad paramétricos), el valor de SBC, el test de razón de verosimilitud (LRT) y los p-valores de la prueba Chi-cuadrado de eliminación de cada término, $Pr(Chi)$. Se puede apreciar que la inclusión de la variable Sexo es significativa, ya que los p-valores de las pruebas son menores al nivel de prueba $\alpha = 0{,}05$, mientras que la inclusión de la variable Edad tiene un p-valor mayor al nivel α, por lo que no es necesaria su inclusión.

A continuación se presentan las fórmulas de los predictores lineales en detalle de los distintos parámetros de distribución para la variable CVF. Para esto, se utiliza la función `stepGAICAll.A()`, donde se parte del modelo m_cvf_013 y con el argumento `scope` de la función de modo que el modelo más "simple.[en] cada parámetro sea una constante ($lower =\sim 1$) y lo más complejo sea de la forma `pb(Edad) + pb(Talla) + Edad + Talla + Sexo`, y utilizando para la elección de los términos un criterio SBC ($k = log(n)$).

El modelo final m_cvf_033, resultante de la función `stepGAICAll.A()`, con familia de distribución BCPE, es de la forma:

$$\mu = a_\mu + pb(\mathsf{Talla}, 2{,}0041) + \mathsf{Sexo}$$

$$log\sigma = a_\sigma$$

$$\nu = a_\nu + \mathsf{Sexo}$$

$$log\tau = a_\tau$$

con un SBC=313.75 y 7 grados de libertad.

En la Tabla 3.17 se puede ver que el hecho de ser niño respecto a ser niña, aumenta en 0.127 unidades (litros) el CVF en promedio, siendo éstos de la misma talla. Además, considerando que $edf_{\mu T}$ es casi 2 y por lo tanto la relación entre Talla y CVF es casi lineal, se puede decir que al incrementar en una unidad (en este caso expresado en

	Estimación	**Error Std.**	**t**	**p**
Parámetro de localización				
función enlace μ: identidad				
coeficientes μ				
Intercepto	-2.848	0.019	-143.94	$<$ 2e-16
pb(Talla)	0.034	1.8e-04	184.50	$<$ 2.2e-16
SexoM	0.127	0.024	5.27	2.07e-07
Parámetro de escala				
función enlace σ: log				
coeficientes σ				
Intercepto	-1.726	0.046	-37.34	$<$ 2e-16
Parámetro de asimetría				
función enlace ν: identidad				
coeficientes ν				
Intercepto	-0.288	0.29	-0.99	0.322
SexoM	1.6191	0.714	2.27	0.024
Parámetro de curtósis				
función enlace τ: log				
coeficientes τ				
Intercepto	0.156	0.09	1.70	0.09

Tabla 3.17: *Coeficientes de la regresión lineal del modelo m_cvf_033*

centímetros) la Talla, el CVF aumenta 0.034 unidades en promedio dentro del mismo sexo. Respecto al parámetro de asimetría, se observa que para el caso de las niñas dicho valor es de $-0{,}288$, lo que lleva a tener una media por debajo de la mediana, mientras que en el caso de los niños el valor es de $1{,}332(1{,}612 - 0{,}288)$, por lo que su media será mayor que la mediana.

Resumen de los Cuantiles Residuales	
media	-0.03404459
varianza	0.9977899
coef. de asimetría	0.001233773
coef. de curtosis	3.034787
coef. de correlación de Filliben	0.9978815

Tabla 3.18: *Cuantiles residuales de los errores para el modelo m_cvf_033.*

Observando la Figura 3.20 se observa que no hay evidencia de mal ajuste del modelo m_cvf_033. Los residuos se aproximan a una distribución normal, dado que mirando el QQ-plot se puede observar que los puntos no se alejan de la recta, y además, los valores en la Tabla 3.18 de media, varianza, coeficiente de asimetría y de curtosis,

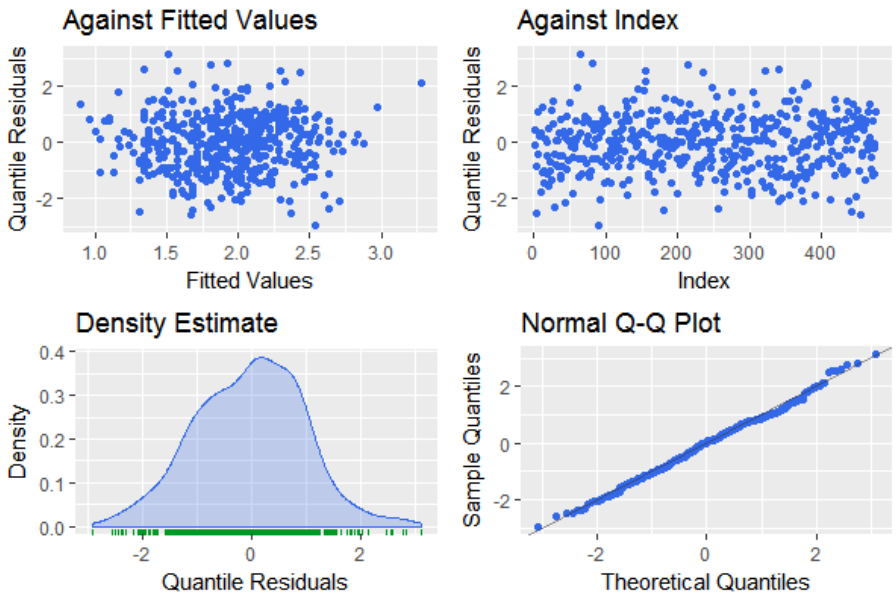

Figura 3.20: *Gráfico de los residuos del modelo m_cvf_033.*

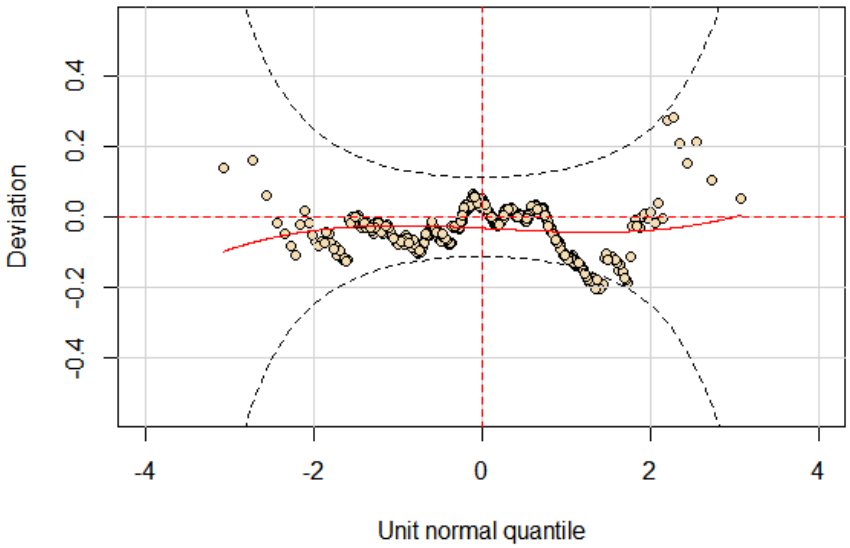

Figura 3.21: *Worm plot del modelo m_cvf_033.*

son similares a los de una normal (media = 0, varianza = 1, coef. de asimetría = 0, coef. de curtosis = 3). Lo mismo se puede decir a través del coeficiente de correlación de Filliben.

El *worm plot* de la Figura 3.21 muestra que el ajuste es bueno. Hay muy pocos

puntos fuera de la región de confianza del 95 % (Ver figura A.3 en página 136)

Variable de respuesta: $\mathbf{FEV}_1$

		Predictor lineal para μ							
Modelo	Distribución	Talla	$edf_{\mu T}$	Edad	$edf_{\mu E}$	Sexo	df	*deviance*	SBC
m_fev_001	NO	`pb()`	7.47	-	-	-	8.47	152.48	204.74
m_fev_002	exGAUS	`pb()`	7.50	-	-	-	9.5	152.03	210.63
m_fev_003	BCPE	`pb()`	3.32	-	-	-	6.3	116.71	155.70
m_fev_013	BCPE	`pb()`	3.72	-	-	Si	7.72	95.01	142.62
m_fev_023	BCPE	`pb()`	3.83	`pb()`	2.01	Si	8.85	86.73	141.38

Tabla 3.19: *Desarrollo de los modelos GAMLSS para FEV_1, con predictor lineal para la mediana μ, los grados de libertad del ajuste, y los valores de deviance y SBC, donde cada fila es un modelo separado.*

Comparando los diferentes modelos ajustados, la Tabla 3.19 muestra que el modelo con mejor ajuste es aquel que tiene una familia de distribución BCPE. Se puede ver que la inclusión de la variable Edad en el modelo reduce en poco más de una unidad el valor del SBC e insume algo más que un grado de libertad adicional.

En la Tabla 3.20 se observa que la eliminación de la variable Edad en el modelo no es estadísticamente significativa, ya que el valor de la prueba Chi-cuadrado es 5.013e-03, que es menor al nivel de prueba $\alpha = 0{,}05$. Por otro lado, la disminución del SBC es de 1.3 unidades. En este caso no se elimina ninguna de las variables.

El modelo final m_fev_023, resultante de la función `stepGAICAll.A()`, con familia

	df	SBC	LRT	Pr(Chi)
Eliminaciones de un solo término para μ				
ninguno		141.32		
pb(Talla)	1.9982	315.52	186.520	< 2.2e-16
Sexo	1.4265	155.52	22.994	3.859e-06
pb(Edad)	1.133	142.62	8.29	5.013e-03

Tabla 3.20: *Test de eliminación de variables en el parámetro de localización del modelo m_fev_023.*

de distribución BCPE, es de la forma:

$$\mu = a_\mu + pb(\mathsf{Talla}, 3{,}83) + pb(\mathsf{Edad}, 2{,}01) + \mathsf{Sexo}$$

$$log(\sigma) = a_\sigma$$

$$\nu = a_\nu$$

$$log(\tau) = a_\tau$$

con un SBC=141.38 y 8.85 grados de libertad.

	Estimación	**Error Std.**	**t**	**p**
Parámetro de localización				
función enlace μ: identidad				
coeficientes μ				
Intercepto	-2.327	0.036	-64.742	$<$ 2e-16
pb(Talla)	0.028	6.035e-04	46.378	$<$ 2e-16
SexoM	0.107	0.022	4.791	2.23e-06
pb(Edad)	0.034	0.010	3.753	1.97e-04
Parámetro de escala				
función enlace σ: log				
coeficientes σ				
Intercepto	-1.904	0.045	-42.56	$<$ 2e-16
Parámetro de asimetría				
función enlace ν: identidad				
coeficientes ν				
Intercepto	1.234	0.312	3.957	8.77e-05
Parámetro de curtósis				
función enlace τ: log				
coeficientes τ				
Intercepto	0.360	0.093	3.867	1.26e-04

Tabla 3.21: *Coeficientes de la regresión lineal del modelo m_fev_023*

En este caso sólo un parámetro de la distribución fue modelizado, y a diferencia de lo ocurrido con la variable CVF, el sexo no influye en la asimetría, ya que no aparece como término en el parámetro ν. Los niños tienen un valor diferencial de 0.107 unidades, es decir, de 107 mL superior en FEV_1 respecto a las niñas de su misma edad y talla. El término de suavizado de la variable Talla tiene un coeficiente con valor 0.028 y el de la variable Edad un coeficiente de 0.034.

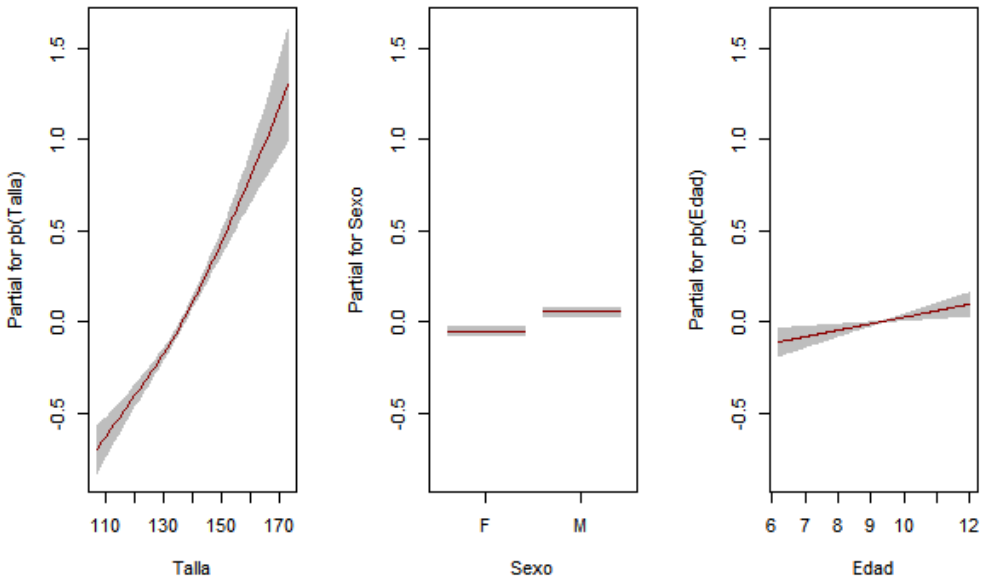

Figura 3.22: *Efecto de las variables Talla, Edad y Sexo sobre el predictor lineal de μ.*

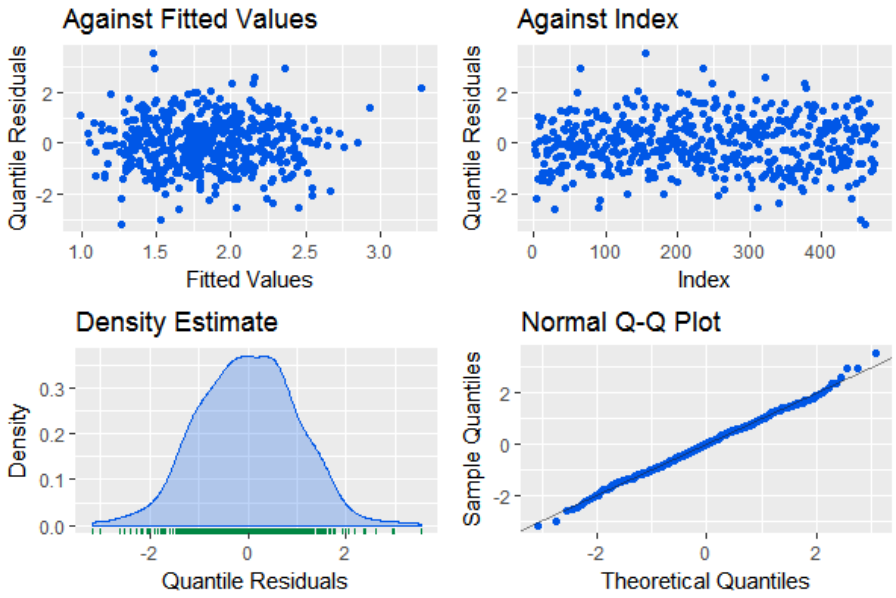

Figura 3.23: *Gráfico de los residuos del modelo m_fev_023.*

En la Figura 3.22 se puede ver como la variable Talla tiene un aporte mayor en el valor del FEV_1 que la variable Edad, que es casi nulo.

En la Figura 3.23 se puede observar que la densidad estimada de los residuos del modelo m_fev_023 se asemejan a una distribución normal. Además, los cuantiles

Resumen de los Cuantiles Residuales	
media	-0.001619348
varianza	1.000402
coef. de asimetría	0.01330202
coef. de curtosis	3.191898
coef. de correlación de Filliben	0.9986347

Tabla 3.22: *Cuantiles residuales de los errores para el modelo m_fev_023.*

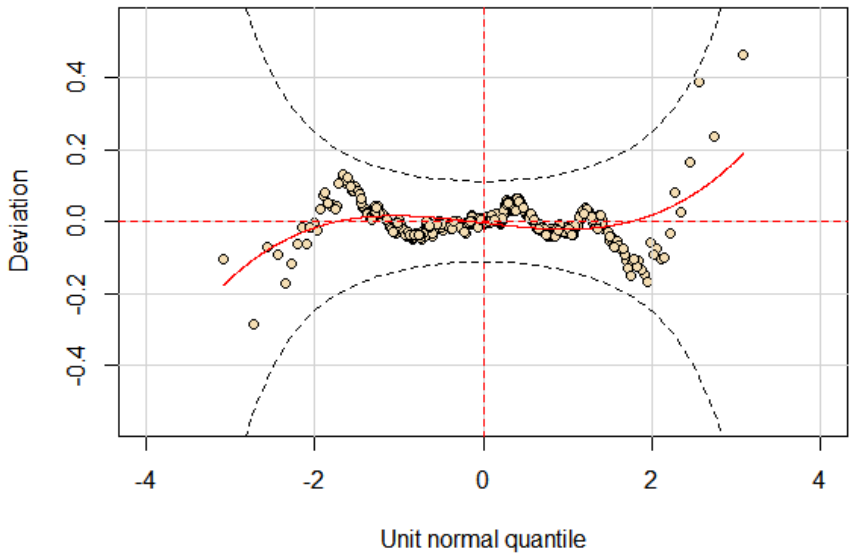

Figura 3.24: *Worm plot del modelo m_fev_023.*

residuales tienen un coeficiente de correlación de Filliben de 0.998 (Tabla 3.22), lo que es indicador de alta de normalidad (la normalidad tiene un valor de 1). La Figura 3.24 muestra el *Worm plot* del modelo m_fev_023, donde se puede notar que algunos puntos tienen una desviación respecto a la media superior a 0.4, lo que podría indicar la presencia de datos atípicos, sin embargo, todos los puntos se encuentran en la región de confianza.

Modelos GAMLSS por sexo

Variable de respuesta: CVF

Niñas

A continuación se plantea la modelización del parámetro espirométrico CVF separando por sexo. El procedimiento es similar al que comprendía a todos los niños (sin separar por sexo), solo que en este caso la variable Sexo no aparece.

		Predictor lineal para μ						
Modelo	Distribución	Talla	$edf_{\mu T}$	Edad	$edf_{\mu E}$	df	*deviance*	SBC
m_cvf_101	NO	`pb()`	2.0017	-	-	3	242.75	259.37
m_cvf_102	SEP4	`pb()`	2.0031	-	-	5	168.39	196.10
m_cvf_103	BCPE	`pb()`	2.0037	-	-	5	154.98	182.69
m_cvf_113	BCPE	`pb()`	2.0035	`pb()`	2.0051	6	150.68	183.95

Tabla 3.23: *Desarrollo de los modelos GAMLSS para CVF de las niñas, con predictor lineal para la mediana μ, los grados de libertad del ajuste, y los valores de deviance,y SBC, donde cada fila es un modelo separado.*

La Tabla 3.23 hace referencia a la comparación entre los modelos de CVF para las niñas. Se observa que el ajuste con la distribución BCPE es más apropiado. Además, los grados efectivos de libertad de las funciones `pb()` son levemente superiores a 2, lo que es indicativo de una relación lineal. Por un tema de simplificación en la interpretación, se decide por incluir las variables Talla y Edad como términos lineales.

	df	SBC	LRT	Pr(Chi)
Eliminaciones de un solo término para μ				
ninguno		183.95		
Talla	1	245.90	67.496	< 2e-16
Edad	1	182.69	4.293	0.03828

Tabla 3.24: *Test de eliminación de variables para el modelo m_cvf_113.*

En la Tabla 3.24 se puede ver que el test de eliminación de la variable Edad tiene un p-valor en la prueba menor al nivel α de 0.05, aumentando en 1.25 el SBC, además de utilizar un grado más de libertad. Al correr la función `stepGAICAll.A()` (con penalización $k = log(n)$), se incluye a la variable Edad de forma lineal.

El modelo final m_cvf_123, con familia de distribución BCPE es de la forma:

$$\mu = a_\mu + Talla + Edad$$

$$log(\sigma) = a_\sigma$$

$$\nu = a_\nu$$

$$log(\tau) = a_\tau$$

con un SBC=182.69 y 5 grados de libertad, donde a_σ, a_ν y a_τ son valores constantes.

	Estimación	**Error Std.**	**t**	**p**
Parámetro de localización				
función enlace μ: identidad				
coeficientes μ				
Intercepto	-2.521	0.045	-55.635	< 2e-16
Talla	0.029	3.968e-04	73.874	< 2e-16
Edad	0.044	8.301e-03	5.308	2.45e-07
Parámetro de escala				
función enlace σ: log				
coeficientes σ				
Intercepto	-1.69942	0.05694	-29.84	< 2e-16
Parámetro de asimetría				
función enlace ν: identidad				
coeficientes ν				
Intercepto	-0.5888	0.3066	-1.92	0.0559
Parámetro de curtósis				
función enlace τ: log				
coeficientes τ				
Intercepto	0.2891	0.1246	2.32	0.0212

Tabla 3.25: *Coeficientes de la regresión lineal del modelo m_cvf_123 para niñas*

A través de la Tabla 3.25 podemos ver que el incremento en el volumen del CVF respecto a la Talla es de 29mL por unidad de medida (cm), con la Edad fija. Por otro lado, si mantenemos la Talla constante, el incremento en el CVF respecto a la Edad es de 44mL por unidad de medida (años).

Mirando la Figura 3.26 junto al Resumen de los Cuantiles Residuales, se observa que si bien la densidad estimada de éstos no es simétrica, en el QQ-plot y el *Worm*

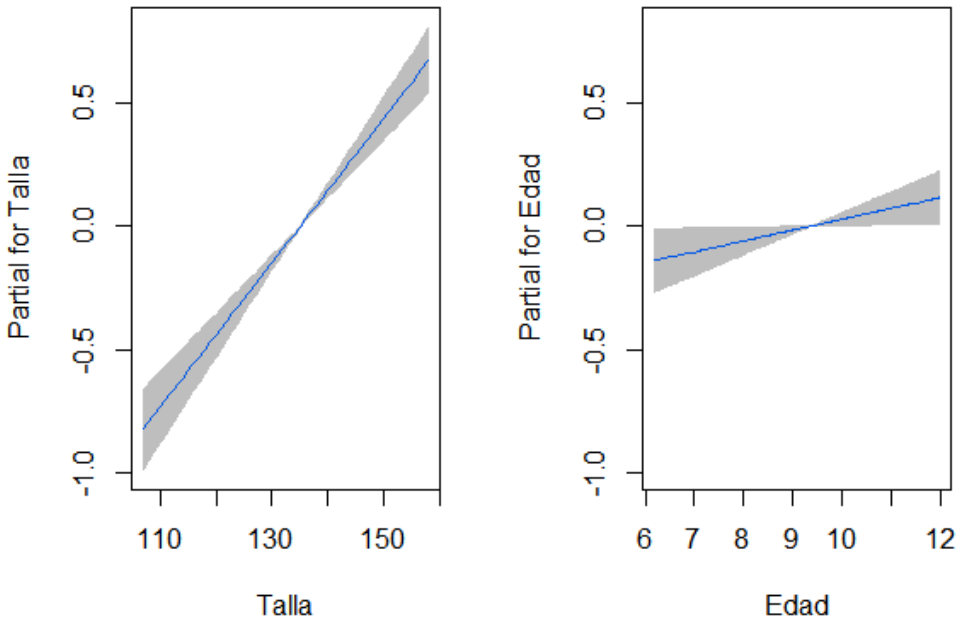

Figura 3.25: *Efecto de las variables Talla y Edad sobre el predictor lineal de μ.*

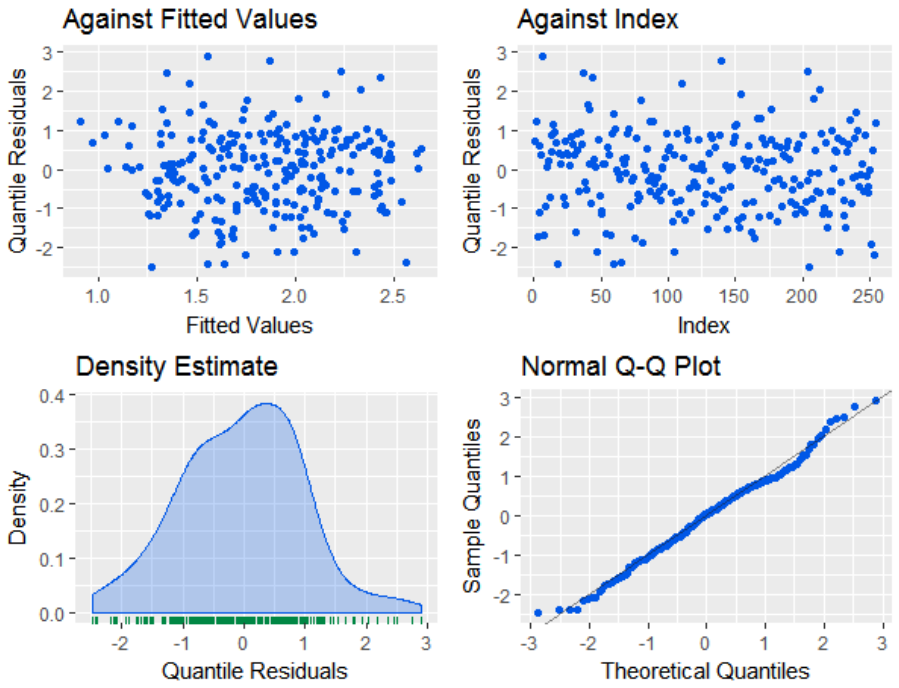

Figura 3.26: *Gráfico de los Cuantiles Residuales del modelo m_cvf_123.*

Resumen de los Cuantiles Residuales	
media	-0.0375327
varianza	1.001122
coef. de asimetría	0.01713106
coef. de curtosis	3.039941
coef. de correlación de Filliben	0.9966144

Tabla 3.26: *Cuantiles residuales de los errores para el modelo m_cvf_123.*

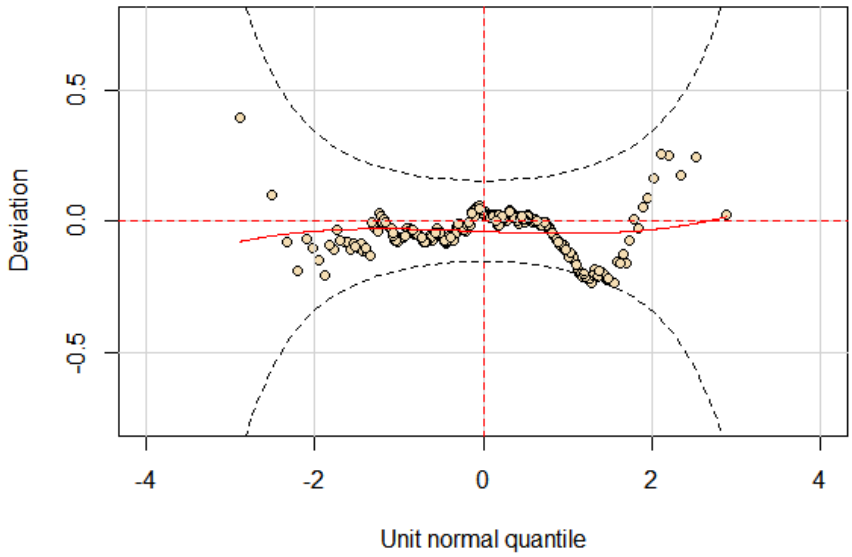

Figura 3.27: *Worm plot del modelo m_cvf_123.*

plot de la Figura 3.27 se observa que no hay falta de ajuste, al mismo tiempo que el coeficiente de correlación de Filliben de la Tabla 3.26 sugiere normalidad.

Niños

Análogamente se desarrolla el modelo de la variable CVF para los niños.

Mirando la Tabla 3.27 se da nota de que si bien tiene mayor *deviance*, el modelo m_cvf_203 presenta menor SBC y utiliza 5 grados de libertad. A su vez, se aprecia que la inclusión de la variable Edad en los modelos, con familias de distribución

		Predictor lineal para μ						
Modelo	Distribución	Talla	$edf_{\mu T}$	Edad	$edf_{\mu E}$	df	*deviance*	SBC
m_cvf_201	NO	`pb()`	2.0016	-	-	3	128.53	144.72
m_cvf_202	exGAUS	`pb()`	2.0011	-	-	4	128.69	150.27
m_cvf_203	BCPE	`pb()`	2.0016	-	-	5	115.02	142.00
m_cvf_212	exGAUS	`pb()`	2.0012	`pb()`	2.0036	5	128.24	155.23
m_cvf_213	BCPE	`pb()`	2.0016	`pb()`	2.0042	6	114.96	147.35

Tabla 3.27: *Desarrollo de los modelos GAMLSS para CVF de las niños, con predictor lineal para la mediana μ, los grados de libertad del ajuste, y los valores de deviance y SBC, donde cada fila es un modelo separado.*

exGAUS y BCPE, aumenta el valor del SBC. También se observa que los grados de libertad efectivos de la función de suavizado `pb()`, tanto para la variable Talla como para Edad, son levemente superiores a 2, lo que indica una relación casi-lineal.

	df	SBC	LRT	Pr(Chi)
Eliminación de un solo término para μ				
ninguno		147.35		
pb(Talla)	1.0034	238.76	96.823	< 2.2e-16
pb(Edad)	1.0042	142.00	0.067	0.7978

Tabla 3.28: *Test de eliminación de variables en el parámetro de localización del modelo m_cvf_213.*

En la Tabla 3.28 se puede ver que el término pb(Edad) del modelo puede no se incluído, ya que su p-valor en la prueba de eliminación, Pr(Chi)=0.79 es mayor a 0.05, al mismo tiempo, en la Tabla 3.27 se puede ver que disminuye el SBC aproximadamente en 5 unidades con su eliminación.

A continuación se presentan las fórmulas de los predictores lineales en detalle de los distintos parámetros de distribución. A través de la función `stepGAICAll.A()` (con penalización $k = log(n)$), se encuentra que el modelo que mejor ajusta es el modelo m_cvf_203, con un término de suavizado dependiente de la Talla para el parámetro de localización (μ).

El modelo final m_cvf_203, con familia de distribución BCPE es de la forma:

$$\mu = a_\mu + pb(\mathsf{Talla}, 2{,}0016)$$

$$log(\sigma) = a_\sigma$$

$$log(\nu) = a_\nu$$

con un SBC=142 y 5 grados de libertad.

	Estimación	**Error Std.**	**t**	**p**
Parámetro de localización				
función enlace μ: identidad				
coeficientes μ				
Intercepto	-3.0920623	0.0862416	-35.85	< 2.2e-16
pb(Talla)	0.0376770	0.0006911	54.52	< 2e-16
Parámetro de escala				
función enlace σ: log				
coeficientes σ				
Intercepto	-1.81584	0.05892	-30.82	< 2.2e-16
Parámetro de asimetría				
función enlace ν: log				
coeficientes ν				
Intercepto	1.0099	0.3769	2.679	7.95e-03
Parámetro de curtosis				
función enlace ν: log				
coeficientes ν				
Intercepto	0.4302	0.1507	2.855	4.72e-03

Tabla 3.29: *Coeficientes de la regresión lineal del modelo m_cvf_203*

El coeficiente del término de suavizado correspondiente en la Tabla 3.29 es de 0.037. Dado que los grados de libertad de tal término de suavizado es prácticamente 2, la relación es lineal, por lo que se puede decir que al aumentar en un centímetro la Talla, el CVF aumenta en 0.037 unidades, o 37mL.

Observando la Figura 3.28 se puede ver que el QQ-plot es adecuado. A su vez, en la Figura 3.29 se puede ver que no hay mal ajuste de los datos. El coeficiente de correlación de Filliben de los Cuantiles Residuales es de 0.998, lo que indica un alto grado de normalidad en ellos (Tabla 3.30).

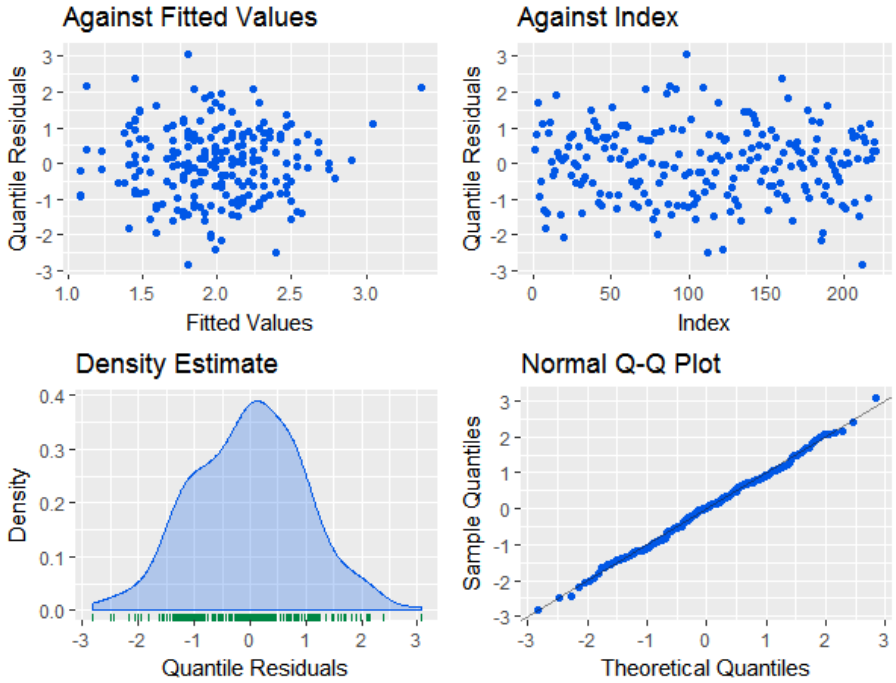

Figura 3.28: *Gráfico de los Cuantiles Residuales del modelo m_cvf_203.*

Resumen de los Cuantiles Residuales	
media	-0.01310141
varianza	1.00445
coef. de asimetría	0.005341516
coef. de curtosis	2.960239
coef. de correlación de Filliben	0.9982216

Tabla 3.30: *Cuantiles residuales de los errores para el modelo m_cvf_203.*

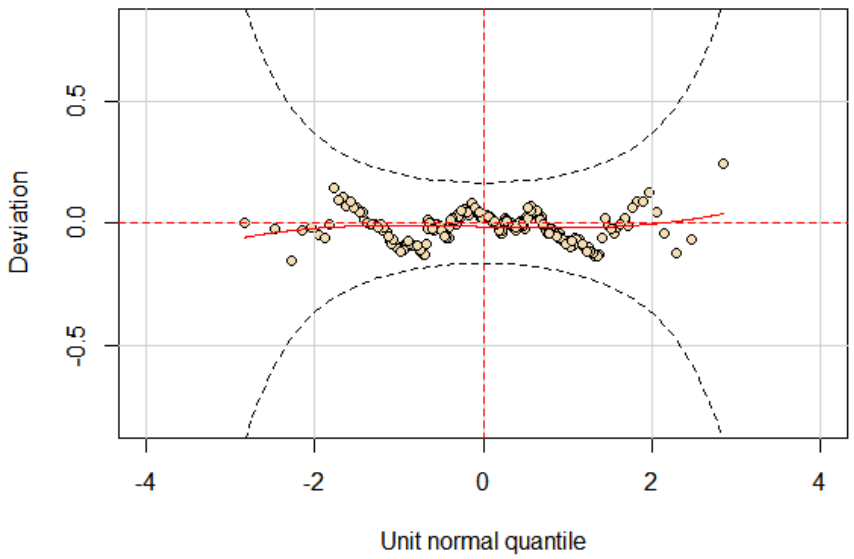

Figura 3.29: *Worm plot del modelo m_cvf_203.*

Variable de respuesta: FEV_1

Niñas

		Predictor lineal para μ						
Modelo	Distribución	Talla	$edf_{\mu T}$	Edad	$edf_{\mu E}$	df	*deviance*	SBC
m_fev_101	NO	`pb()`	3.0074	-	-	4	64.31	86.54
m_fev_102	SN2	`pb()`	2.0034	-	-	4	65.57	87.77
m_fev_103	BCPE	`pb()`	3.6524	-	-	6.65	46.38	83.21
m_fev_113	BCPE	`pb()`	3.5691	`pb()`	2.0072	7.66	39.64	82.09

Tabla 3.31: *Desarrollo de los modelos GAMLSS para FEV_1 de las niñas, con predictor lineal para la mediana μ, los grados de libertad del ajuste, y los valores de deviance y SBC, donde cada fila es un modelo separado.*

Se observa en la Tabla 3.31 que el modelo con mejor ajuste a los datos es uno con familia de distribución BCPE, ya que tiene menor SBC. La inclusión de un término de suavizado casi lineal con la variable Edad ($edf_{\mu e} = 2{,}0072$) reduce el SBC en 1.18 unidades.

Mirando los resultados de la prueba de eliminación de cada término aditivo de

	df	SBC	LRT	Pr(Chi)
Eliminaciones de un solo término para μ				
ninguno		82.147		
pb(Talla)	2.6614	162.592	95.203	$<$ 2.2e-16
pb(Edad)	1.0135	83.264	6.738	9.658e-03

Tabla 3.32: *Test de eliminación de variables en el modelo m_fev_113.*

la Tabla 3.32 se observa que la inclusión de la variable Edad es estadísticamente significativa, ya que el p-valor de la prueba (Pr(Chi)=9.6e-03) es menor a 0.05. Aún así, se opta por quedarse con un modelo que solo contiene la Talla a modo de poder presentar los resultados a través de curvas percentilares.

El modelo final m_fev_103, con familia de distribución BCPE es de la forma:

$$\mu = a_\mu + pb(\mathsf{Talla}, 3{,}6524)$$

$$log(\sigma) = a_\sigma$$

$$\nu = a_\nu$$

$$log(\tau) = a_\tau$$

con un SBC=83.21 y 6.65 grados de libertad.

En la Figura 3.30 se puede ver como cambia la función de suavizado al variar la variable Talla. Se puede obervar una leve curva con concavidad positiva. El la parte central de los valores de la Talla, entre 125 cm y 150 cm aproximadamente, su comportamiento es lineal, mientras que en los extremos se comporta diferente.

En la Figura 3.32 se puede ver que no hay indicios de un ajuste inadecuado, pero si se puede ver que hay ciertos puntos que tienen un desvío mayor al resto, lo cual se evidencia en el QQ-plot de la Figura 3.31, donde se observa que hay puntos que se salen de la linea identidad sobre la derecha del gráfico. La densidad estimada de los residuos se asemejan a los de una distribución normal en términos generales, lo cual también nos dice el coeficiente de Filliben con un valor de 0.997 (Tabla 3.34).

	Estimación	**Error Std.**	**t**	**p**
Parámetro de localización				
función enlace μ: identidad				
coeficientes μ				
Intercepto	-2.572752	0.194833	-13.21	$<$ 2e-16
pb(Talla)	0.032235	0.001486	21.69	$<$ 2e-16
Parámetro de escala				
función enlace σ: log				
coeficientes σ				
Intercepto	-1.8683	0.0543	-34.41	$<$ 2e-16
Parámetro de asimetría				
función enlace ν: identidad				
coeficientes ν				
Intercepto	0.7996	0.3246	2.464	1.44e-02
Parámetro de curtósis				
función enlace τ: log				
coeficientes τ				
Intercepto	0.3550	0.1163	3.053	2.51e-03

Tabla 3.33: *Coeficientes de la regresión lineal del modelo m_fev_103. Nota: El modelo con menor SBC contiene a la variable Edad en el parámetro μ y ν, pero en términos de predicción no existen cambios significativos, por lo que se decide utilizar el modelo sólo con la variable Talla.*

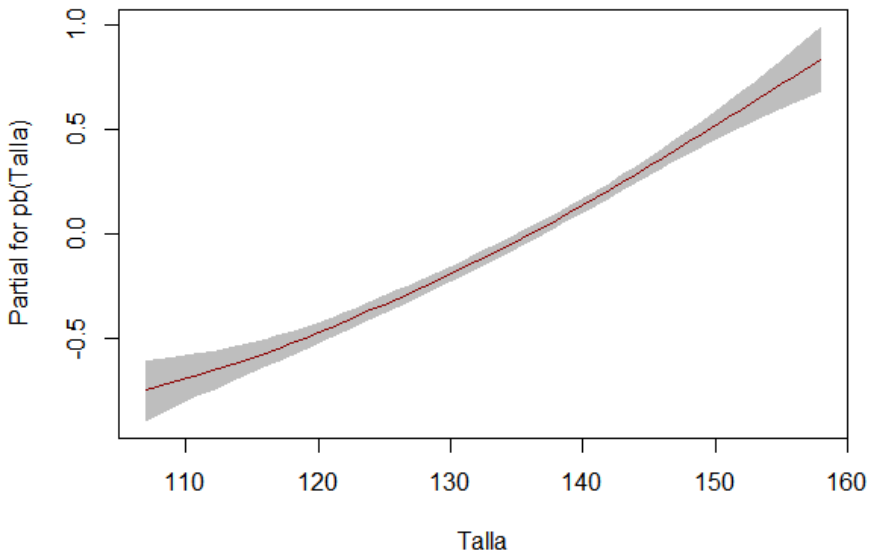

Figura 3.30: *Efecto de la variable Talla en el término de suavizado para el modelo m_fev_103.*

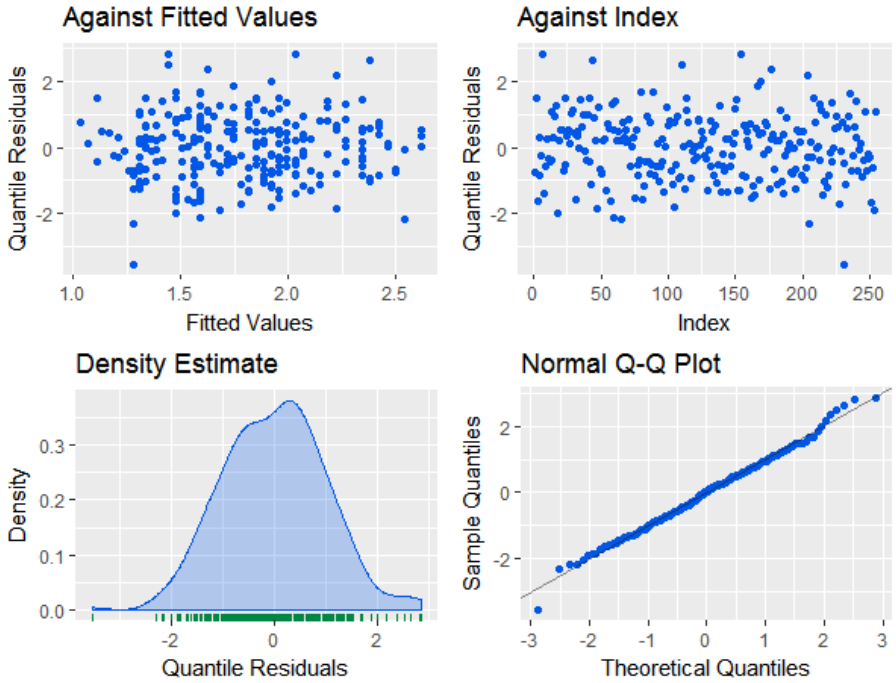

Figura 3.31: *Gráfico de los Cuantiles Residuales del modelo m_fev_103.*

Resumen de los Cuantiles Residuales	
media	-0.01041859
varianza	1.00167
coef. de asimetría	0.03358931
coef. de curtosis	3.243744
coef. de correlación de Filliben	0.9972792

Tabla 3.34: *Cuantiles residuales de los errores para el modelo m_fev_103.*

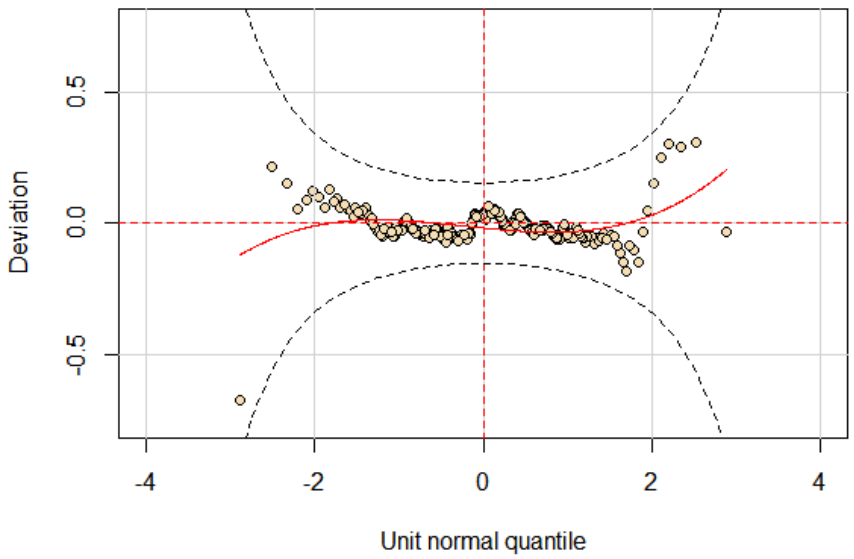

Figura 3.32: *Worm plot del modelo m_fev_103.*

Niños

		Predictor lineal para μ						
Modelo	Distribución	Talla	$edf_{\mu T}$	Edad	$edf_{\mu E}$	df	*deviance*	SBC
m_fev_201	NO	`pb()`	5.467234	-	-	6.467234	33.5521	68.4339
m_fev_202	exGAUS	`pb()`	5.347792	-	-	7.347792	33.2261	72.8573
m_fev_203	BCPE	`pb()`	3.669627	-	-	6.669627	36.0489	72.0224
m_fev_211	NO	`pb()`	5.508823	`pb()`	2.005373	7.514	32.6116	73.1403
m_fev_213	BCPE	`pb()`	3.876297	`pb()`	2.005869	7.88	33.997	76.5104

Tabla 3.35: *Desarrollo de los modelos GAMLSS para FEV_1 de las niños, con predictor lineal para la mediana μ, los grados de libertad del ajuste, y los valores de deviance y SBC, donde cada fila es un modelo separado.*

En la Tabla 3.35 se observa que el modelo con distribución normal es el que ajusta mejor los datos, con un SBC=68.43 y con 6.47 grados de libertad. El que le sigue es el modelo m_fev_203, con un SBC=72.02 y 6.67 grados de libertad. También se ve que la inclusión de un término de suavizado con la variable Edad, en ambos casos aumenta el SBC.

La Tabla 3.36 muestra que la adición de un término de suavizado de la variable Edad no es necesaria, ya que el p-valor de la prueba de eliminación es mayor a 0.05.

	df	SBC	LRT	Pr(Chi)
Eliminaciones de un solo término para μ				
ninguno		76.510		
pb(Talla)	2.8788	176.945	115.962	< 2.2e-16
pb(Edad)	1.2125	72.022	2.052	0.1935

Tabla 3.36: *Test de eliminación de variables en el modelo m_fev_213.*

El modelo final m_fev_203, con familia de distribución BCPE es de la forma:

$$\mu = a_\mu + pb(\mathsf{Talla}, 3{,}66)$$

$$log(\sigma) = a_\sigma$$

$$\nu = a_\nu$$

$$log(\tau) = a_\tau$$

con un SBC=72.02 y 6.66 grados de libertad.

	Estimación	**Error Std.**	**t**	**p**
Parámetro de localización				
función enlace μ: identidad				
coeficientes μ				
Intercepto	-2.775246	0.207739	-13.36	< 2e-16
pb(Talla)	0.034525	0.001575	21.92	< 2e-16
Parámetro de escala				
función enlace σ: log				
coeficientes σ				
Intercepto	-1.95034	0.05298	-36.82	< 2e-16
Parámetro de asimetría				
función enlace ν: identidad				
coeficientes ν				
Intercepto	1.230	0.384	3.202	1.57e-03
Parámetro de curtósis				
función enlace τ: log				
coeficientes τ				
Intercepto	0.5729	0.1450	3.951	1.06e-04

Tabla 3.37: *Coeficientes de la regresión lineal del modelo m_fev_203*

En la Tabla 3.37 se muestran los coeficientes estimados del modelo con sus errores estándar, junto a el valor de la prueba t y su p-valor correspondiente.

En la Figura 3.33 se observa que a partir de los 140 cm, el efecto de la Talla sobre

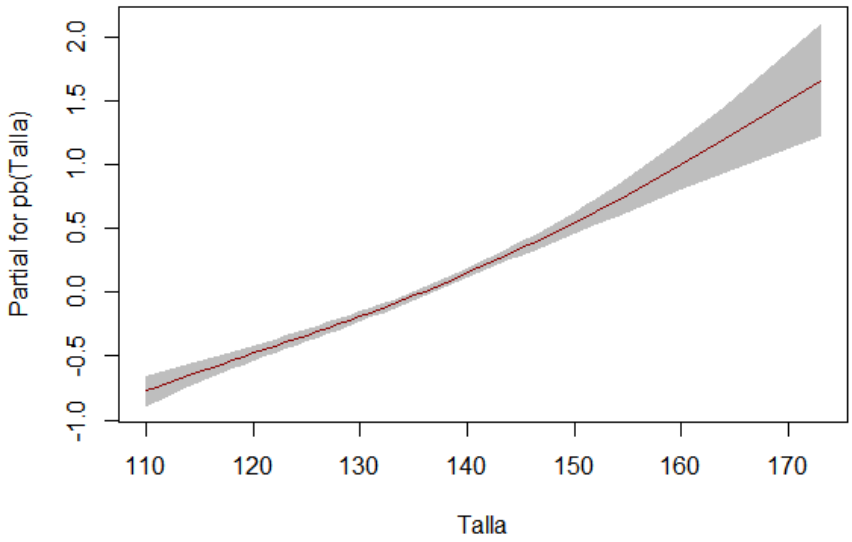

Figura 3.33: *Efecto de la variable* Talla *en el término de suavizado para el modelo m_fev_203.*

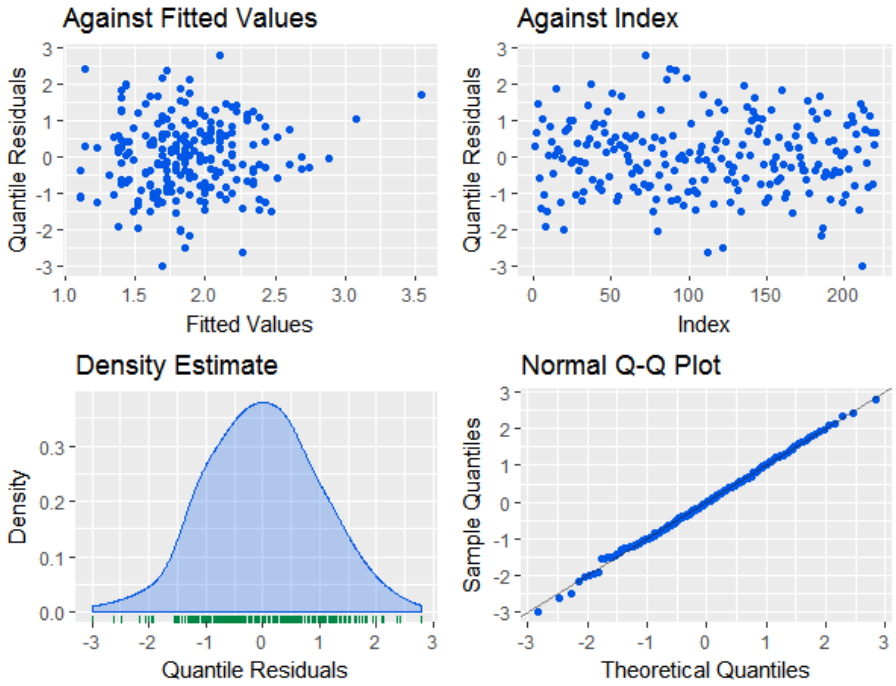

Figura 3.34: *Gráfico de los Cuantiles Residuales del modelo m_fev_203*

el término de suavizado se incrementa.

La Figura 3.35 muestra el *Worm plot* del modelo m_fev_203, donde se observa que

Resumen de los Cuantiles Residuales	
media	0.002812501
varianza	1.004405
coef. de asimetría	0.001862636
coef. de curtosis	2.983888
coef. de correlación de Filliben	0.998949

Tabla 3.38: *Cuantiles residuales de los errores para el modelo m_fev_203.*

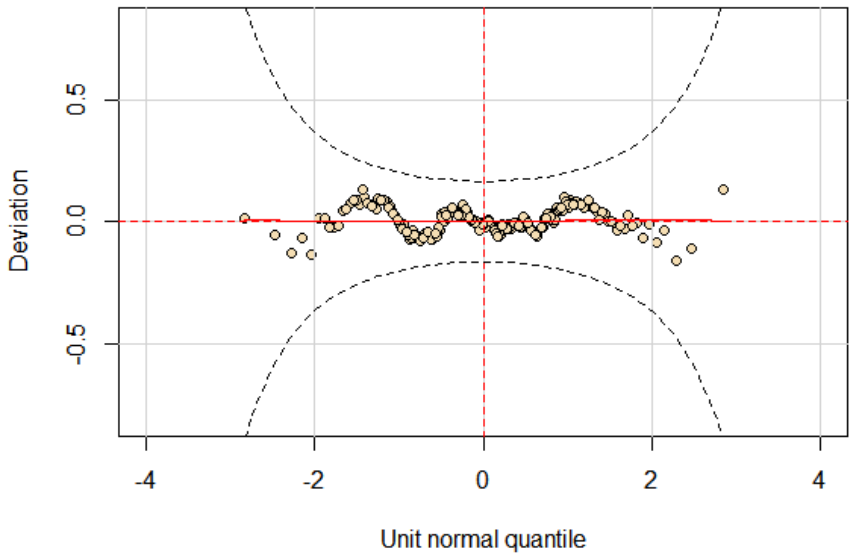

Figura 3.35: *Worm plot del modelo m_fev_203.*

el ajuste es adecuado, lo que también nos indica el QQ-plot de la Figura 3.34. La densidad estimada de los Cuantiles Residuales es similar a la de una distribución normal. El coeficiente de correlación de Filliben, con un valor de 0.998, indica una alta normalidad.

Parámetro espirométrico	**Sexo**	**Ecuación de regresión**	**SBC**
CVF (L)	Global	$\mu = -2{,}848 + pb(Talla, 2{,}0041) + 0{,}127 \times Sexo$ $log(\sigma) = -1{,}726$ $\nu = -0{,}288 + 1{,}619 \times Sexo$ $log(\tau) = 0{,}159$	313,75
	Niñas	$\mu = -2{,}521 + 0{,}029 \times Talla + 0{,}044 \times Edad$ $log(\sigma) = -1{,}699$ $\nu = -0{,}588$ $log(\tau) = 0{,}2891$	182,69
	Niños	$\mu = -3{,}092 + pb(Talla, 2{,}0016)$ $log(\sigma) = -1{,}815$ $\nu = 1{,}0099$ $log(\tau) = 0{,}4302$	142,00
FEV_1 (L)	Global	$\mu = -2{,}327 + pb(Talla, 3{,}83) + pb(Edad, 2{,}01) + 0{,}107 \times Sexo$ $log(\sigma) = -1{,}904$ $\nu = 1{,}234$ $log(\tau) = 0{,}360$	141,38
	Niñas	$\mu = -2{,}572 + pb(Talla, 3{,}6524)$ $log(\sigma) = -1{,}868$ $\nu = 0{,}799$ $log(\tau) = 0{,}355$	83,21
	Niños	$\mu = -2{,}775 + pb(Talla, 3{,}66)$ $log(\sigma) = -1{,}95$ $\nu = 1{,}230$ $log(\tau) = 0{,}573$	72,02

Nota: todos los modelos presentados tienen una distribución BCPE.

Tabla 3.39: *Ecuaciones de regresión para los parámetros espirométricos CVF y FEV_1*

Capítulo 4

Resultados

Discusión

La Tabla 3.39 presenta un resumen de todos los modelos finales para CVF y FEV_1, donde se explicita la ecuación para cada parámetro de distribución (μ, σ, ν, τ), tanto para niñas y niños por separado como todos juntos de forma global.

(Quanjer *et al.*, 2012), con un estudio multi étnico de la *Global Lung Iniciative* (GLI), con personas entre los 3 y 95 años, proponen las siguientes ecuaciones utilizando el método LMS, con distribución Box-Cox-Cole-Green (BCCG), con un tamaño de muestra de aproximadamente 30000 personas de sexo masculino y 40000 de sexo femenino. *Grupo* indica la etnicidad, donde caucásico es la referencia y aparecen afroamericanos, mexicanos, latinoamericanos, indios-pakistaníes, norasiáticos del este, sudasiáticos del este, norafricanos, Irán, Omán y otros, agrupados en caucásicos, africanos-americanos, norasiáticos del este y sudasiáticos del este. Se incluyen datos para adultos de la ciudad de Montevideo.

$$log(Y) = a + b \times log(Talla) + c \times log(Edad) + spline(Edad) + d \times grupo$$

$$log(CoV) = a + b \times log(Edad) + spline(Edad)$$

donde CoV es el *coeficiente de variación*, definido como $CoV = 100 \times SD/predicción$.

(Cole y Stanojevic, 2009) en un estudio realizado con personas entre 4 y 80 años, proponen la siguiente ecuación para el parámetro FEV_1 en personas del género masculino, con 1621 datos de distintos centros (EEUU, Bélgica, Inglaterra y Canadá). El modelo fue realizado con el enfoque GAMLSS, utilizando una distribución BCPE.

$$log(\mu) = -11{,}4 + 2{,}45 \times log(Talla) + cs(logEdad, 7) + centro$$

$$log(\sigma) = 1{,}7 + -0{,}95 \times log(Talla) + 0{,}30 \times log(Edad) + centro$$

$$\nu = 1{,}50$$

$$\tau = 2$$

(Rosenthal y Bain, 1993), en un estudio de una población de niños y jóvenes entre 4 y 19 años, propone un modelo lineal dónde la media μ se relaciona con la Talla como $\mu = A + B \times Talla$, separando por sexo, y con ecuaciones partidas según el valor de Talla (162.5 para los niños y 152.5 para las niñas). En la Tabla 4.1 se presentan los valores estimados de los coeficientes de dicho estudio.

(Meng-Chiao Tsai, 2010), en un estudio sobre una población de niños entre 6 y 11 años de Taiwan, con un tamaño de muestra de 309, proponen ecuaciones con modelos lineales con la Talla como única variable regresora.

	Sexo	Talla	A	B
CVF (L)	M	$< 162{,}6$	-3.619	0.0429
	M	$> 162{,}5$	-7.038	0.0678
	F	$< 152{,}6$	-3.311	0.03918
	F	$> 152{,}5$	-3.881	0.04512
FEV_1 (L)	M	$< 162{,}6$	-2.780	0.03425
	M	$> 162{,}5$	-5.108	0.0521
	F	$< 152{,}6$	-2.734	0.03316
	F	$> 152{,}5$	-3.680	0.04112

Tabla 4.1: *Coeficientes del modelo lineal para los parámetros CVF y FEV_1 de Rosenthal, Fuente:Lung function in white children aged 4 to 19 years: I-Spirometry*

	Sexo	Ecuación de regresión
CVF (L)	Niños	$-2{,}743 + 0{,}0337 \times Talla$
	Niñas	$-2{,}643 + 0{,}0323 \times Talla$
	Global	$-2{,}690 + 0{,}0330 \times Talla$
FEV_1 (L)	Niños	$-2{,}596 + 0{,}0317 \times Talla$
	Niñas	$-2{,}527 + 0{,}0306 \times Talla$
	Global	$-2{,}559 + 0{,}0311 \times Talla$

Tabla 4.2: *Ecuaciones de regresión de los parámetros CVF y FEV_1 en relación a la Talla. Fuente: Spirometric reference equations fot healthy children aged 6 to 11 years in Taiwan - Meng-Chiao - 2010*

Si bien el conjunto de datos del estudio no explota de gran manera la potencia de los modelos GAMLSS, se muestra que puede ser usado para construir rangos de referencia espirométricos dependientes de la talla, el sexo, y en algunos casos de la edad. También logra sobreponerse a los supuestos clásicos de la regresión lineal múltiple de linealidad, aditividad, homocedasticidad y normalidad de los errores. Ejemplificando, la mediana μ de la distribución para el FEV_1 en niños se encontró que varía de forma no lineal con la talla 3.4; en CVF, se encontró que la asimetría de la distribución ν era influída por el sexo, al igual que la mediana μ 3.17.

La no inclusión de la variable Edad en algunos de los modelos, se debe a que en ciertos casos dicha inclusión aumentaba el GAIC(k=log(n)), o SBC, por lo que el algoritmo de la función `stepGAICAll.A()` no la incluía. En los casos de que el algoritmo incluyera a la variable Edad ya sea lineal o como un término de suavizado, a través de la función `drop1()` se realizaba un test de Chi-cuadrado para analizar la prueba de inclusión y se verificaba cuanto mejoraba la predicción en ese caso. Al no presentar una mejora, se optó por no incluirla.

La utilización de gráficos de curvas en la medicina clínica es una práctica común ya que facilita la interpretación. Por ésto se intentó mantener modelos que contengan sólo la variable Talla, no sólo porque se ha mostrado y demostrado en la literatura referente a la función respiratoria que la talla es el parámetro más determinante, sino también para poder incluir curvas univariadas de los percentiles 5, 50 y 95 a modo de agilizar la interpretación y poder situar al niño rápidamente.

Se parte de considerar tres modelos para CVF y FEV_1, globales y separados por sexo, con distribución normal. Estos se comparan con los que resultan de la prueba de robustez (con mayor frecuencia en el proceso iterativo) y además con la distribución BCPE. Esto se debe a poder comparar entre un modelo de regresión simétrico a través de la distribución normal (el más utilizado en general), y luego ver los cambios en la forma de la distribución a través de la utilización de distribuciones que permitan

controlar la asimetría y curtosis, y en este aspecto comparar entre la utilización de distribuciones procedentes de los distintos procesos iterativos mostrados y la distribución Box-Cox Power Exponential (BCPE), que es utilizada por otros autores en este tipo de estudios, tales como Cole (Cole y Stanojevic, 2009) y Stanojevic (Stanojevic *et al.*, 2007).

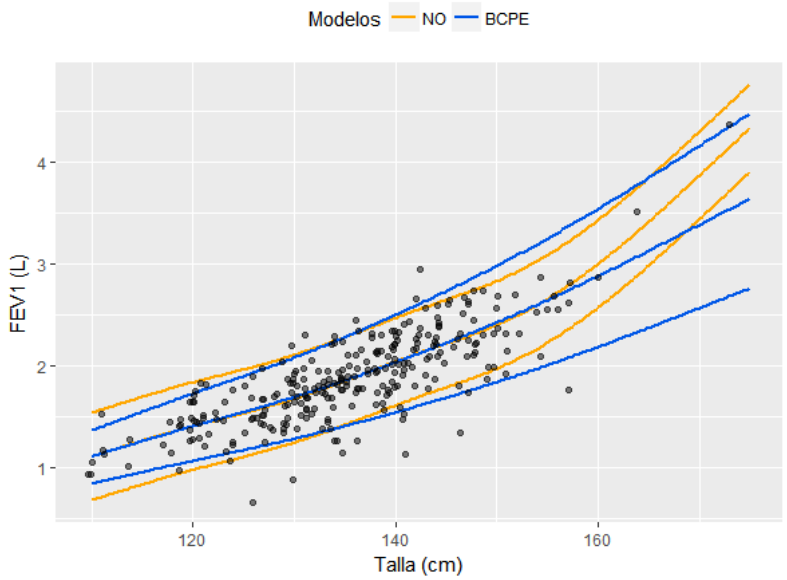

Figura 4.1: *Gráfico comparativo de predicción de curvas percentilares (5 %, 50 % y 95 %) entre el modelo m_fev_201 (naranja) y el modelo m_fev_203 (azul).*

Para entender los cuidados que se deben tener en la utilización de estos modelos, se puede mirar el caso de la modelización de la variable FEV_1 para los niños. En principio, si tomara como guía el criterio de selección de menor GAIC únicamente, se inclinaría por escoger el modelo m_fev_201, pero en la Figura 4.1 se puede ver que dicho modelo cambia de manera brusca la pendiente en valores de Talla mayores a 150 cm, mientras que el modelo m_fev_203 tiene recorrido más suave. Esto se debe a que el modelo con distribución normal, es simétrico respecto a la media y sólo cuenta con dos parámetros (media y desvío estándar), pero al aparecer puntos que podrían considerarse atípicos, éstos hacen un efecto de palanca que tira para arriba la curva. Es por esa razón que se opta por utilizar el modelo m_fev_203, ya que la distribución BCPE es más flexible al contar con cuatro parámetros, resultando en una curva más suave.

Se tenían mayores expectativas de poder captar más suavidad y no tanta linealidad, quizas por un tema de la selección de los conjuntos de entrenamiento, o por el rango reducido de edad, no se logró. Sin embargo, los modelos de la variable FEV_1 presentan esta característica de curvas suaves, logrando incluir términos de suavizado en función de la talla.

Capítulo 5

Conclusiones

El objetivo principal del presente trabajo consistía en encontrar curvas de referencia de parámetros espriométricos en niños uruguayos, con datos procedentes de investigadores del Centro Hospitalario Pereira Rossell, con la idea de contar con referencias locales, ya que hasta el momento se han venido utilizando valores de referencia procedentes de otros paises con características ambientales y climáticas distintas.

Para ese objetivo se ha utilizado el marco teórico de los Modelos Aditivos Generalizados de Localización, Escala y Forma (GAMLSS), logrando contruir modelos sencillos dentro de las complejidades que éstos presentan. A su vez se experimentó con las distintas funciones de ajuste de densidades contenidas dentro de la librería **gamlss**, desarrollándose un proceso iterativo para estudiar la robustez de las mismas, ya que las distribuciones incluídas se construyen a través de transformaciones de distribuciones más conocidas, lo que puede llevar a resultados muy similares o con mínimas diferencias.

Se mostró la construcción de los modelos GAMLSS, donde partiendo de los modelos lineales, se fueron levantando los supuestos clásicos de normalidad y homocedasti-

cidad que son los que a menudo no se sostienen en la realidad, y como sobreponerse a esas dificultades, pasando a los modelos lineales generalizados (GLM). Luego se introdujeron los modelos aditivos generalizados (GAM) y como utilizar funciones de suavizado dentro de los mismos, y pasando por los modelos mixtos lineales generalizados (GLMM) y los modelos mixtos aditivos generalizados (GAMM), para luego llegar a los GAMLSS. A su vez se mostraron todas las posibilidades que este tipo de modelos tiene, los cuales permite modelizar los parámetros de localización, escala, asimetría y curtosis como funciones de variables regresoras. Además, se presentó el proceso de ajuste de modelos de regresión paramétricos, los criterios de selección del modelo para los cuatro componentes (distribución, funciones de enlace, términos aditivos y parámetros de suavizado). También se presentaron las distintas herramientas de diagnóstico de los modelos, los cuales incluyen los cuantiles residuales normalizados y los gráficos de gusano, entre otros.

Se llevó a cabo un estudio entre los niños normales y con antecedentes patológicos con el objetivo de ver sus diferencias, llevando a cabo una prueba de Hotelling T^2 multivariada con los distintos parámetros espirométricos CVF, FEV_1, FEF_{25-75} y PFE, con un supuesto de multi-normalidad para cada grupo. Se tuvo que proceder de esta forma debido a que los cuatro parámetros se obtienen de la misma maniobra o ejercicio. Se encontró que los niños normales y los niños con antecedentes patológicos difieren en el parámetro FEF_{25-75} (lo cual es un hallazgo novedoso y sorpresivo) lo que no acepta la idea de que provengan de la misma población. Esto llevó a la utilización de sólo los datos correspondientes a los niños normales a efectos de construir los modelos correspondientes, y al mismo tiempo hacer que el estudio sea comparable con otros estudios internacionales en los cuales eran ajustados con datos de niños normales solamente. Esta necesidad de comprobar estadísticamente dicha situación se debió a que los valores de los parámetros CVF y FEV_1 en ambos grupos eran similares, por lo que en un principio se pensó en tomarlos como un sólo grupo.

Previo a la modelización de los parámetros espirométricos CVF y FEV_1, se realizó un proceso iterativo utilizando una función de ajuste de densidades incluída en la librería **gamlss** con tamaños de muestra del 80 % del total, con el fin de encontrar una distribución para utilizarla en los modelos, al menos con la intención de encontrar una distribución de partida. Para la elección de esta distribución, se tomó en cuenta la frecuencia relativa de ajuste resultante del proceso iterativo, seleccionando aquella con un valor más alto (proceso por voto). Los resultados fueron variados, pero una característica que han tenido en común, es la necesidad de incluir, al menos, un parámetro de asimetría, además de los de localización y escala. Se hizo una comparación entre la función para ajustar densidades del paquete **gamlss** con el del paquete **MASS**.

La estructura de la modelización de los parámetros CVF y FEV_1 presente en este trabajo se realizó separando los datos en un conjunto para ajustar los modelos, también llamado de *entrenamiento* y otro para validación del mismo, con una relación de 0.8 y 0.2 respectivamente. Se ajustaron modelos globales y para cada sexo por separado. Con la intención de hacer una comparación entre los modelos lineales y los GAMLSS, primero se ajustó un modelo con una distribución normal. Además se hizo lo mismo con la distribución con mayor frecuencia relativa en cada caso correspondiente, que permitía la modelización de la asimetría. También se hicieron ajustes con una distribución Box-Cox Power Exponential, BCPE, que permite la modelización de un cuarto parámetro referente a la curtosis, donde a su vez ha sido utilizada por otros estudios para este tipo de aplicaciones.

Se partió de modelos que solo incluyeran términos de suavizado en función de la talla, dado que esta variable es la que tiene mayor peso en relación al tipo de unida-

des de las variables en cuestión, que refieren a volúmenes, expresados en litros (L), o litros por unidad de tiempo (L/1s). Se contaba también con las variables edad y peso. Se ha encontrado en estudios previos (Stanojevic *et al.*, 2007) que el peso no tiene relación con los volúmenes pulmonares, ya que un alto valor en el peso no implica necesariamente un alto volumen pulmonar (puede provenir del tamaño de los huesos, musculos, o grasa corporal). En cuanto a la edad, al trabajar con rangos de 6 a 12 años, hay una fuerte relación con la talla, ya que se encuentra dentro de un período de crecimiento del niño. Sin embargo se ha incluído dentro de la construcción de los modelos para poner a prueba si su inclusión era estadísticamente significativa. Cada parámetro de distribución se quizo modelizar con la estructura $g(\theta_k) = a_\theta + pb(Talla) + pb(Edad)$, para cada $k = 1, 2, 3, 4$, donde $g(.)$ es la función de enlace y $pb()$ son splines penalizados, con sus grados de suavizado correspondientes.

Se encontró que la utilización de la distribución BCPE ajusta mejor los datos, tomando como criterio de ajuste el SBC (o BIC), tanto en los modelos globales como en los específicos para cada sexo. Esto hace pensar que en este tipo de modelos, el ajuste de una densidad univariada no se relaciona con la forma que puede llegar a tener a través de una regresión, es decir, puede que en un escenario no se logre captar la curtosis, pero a la hora de una regresión, esta si sea necesaria.

Se lograron realizar tablas y gráficos de las curvas de los percentiles 5, 50 y 95 a través de los modelos en relación a la talla para cada sexo por separado, debido a, por un lado, una simplificación en la presentación, y por otro a una limitante funcional dentro del paquete, ya que la estimación de los percentiles solo era posible para modelos con una sola variable regresora. Se encontró que los niños tienen valores más altos en los percentiles 5 y 50, tanto en CVF como en FEV_1, no así en el percentil 95. Sin embargo, los más importantes son el 5 y el 50, ya que valores por

debajo del percentil 5, pueden indicar patología.

En comparación con otros estudios, la comparación más cercana es con el estudio elaborado por Meng-Chao, en un estudio en niños de 6 a 11 años de edad en Taiwan, donde los modelos resultantes son modelos lineales con la variable talla. El resto de los estudios cuentan con un rango de edades más amplio, donde (Rosenthal y Bain, 1993) lo hicieron con datos de niños y jovenes entre 4 y 19 años, con modelos lineales partidos a través de la talla. Cole *et al.* (2008) y Stanojevic *et al.* (2007) tienen rangos de edades que van de los 4 a los 80 años, lo cual hace que la edad aporte más información.

Los modelos GAMLSS resultaron ser una buena técnica para abordar este tipo de problemas debido en gran parte a su flexibilidad. El paquete **gamlss** contiene funciones con grandes capacidades, ya sea para ajustar densidades "raras", como para construir modelos de regresión complejos, permitiendo la inclusión de funciones de suavizado de una o más variables, efectos aleatorios o variables del tipo factor. Posee varias herramientas para chequear posibles inadecuancias en los modelos, como los gráficos de gusano y los cuantiles residuales.

Bibliografía

Akanztiliotou, K., Rigby, R., y Stasinopoulos, D. (2002). The R implementation of Generalized Additive Models for Location, Scale and Shape.

Boor, C. (2001). *A practical guide to splines : with 32 figures*. Springer, New York.

Breslow, N. E. y Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Buuren, S. v. y Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20(8):1259–1277.

Cole, T. y Stanojevic, S. (2009). Age and size related reference ranges a case study of spirometry through childhood and adulthood. *Statistics in Medicine*, 28(5):880–898.

Cole, T., Stanojevic, S., y Stocks, J. (2008). Age-and size-related reference ranges: A case study of spirometry through childhood and adulthood. *Statistics in Medicine*.

de Onis, M., Onyango, A., Borghi, E., Siyam, A., Nishida, C., y Siekmann, J. (2007). Development of a who growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, 85:661–668.

Dunn, P. K. y Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Eilers, P. H. C. y Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121.

Green, P. J. (2000). *Nonparametric regression and generalized linear models : a roughness penalty approach*. Chapman & Hall/CRC, Boca Raton.

Hastie, T. y Tibshirani, R. (1986). *Generalized Additive Models*, volumen 1. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Hastie, T. y Tibshirani, R. (1993a). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.

Hastie, T. y Tibshirani, R. (1993b). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.

James, A. T. (1954). Normal multivariate analysis and the orthogonal group. *Ann. Math. Statist*, 25(1):40–75.

Lange, K. (1999). *Numerical Analysis for statisticians*. Springer.

Lin, X. y Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):381–400.

Meng-Chiao Tsai, M.-J. J. (2010). Spirometric reference equation for healthy children aged 6 to 11 in taiwan. *J Chin Med Assoc*, 73(1):21–28.

Müller, K. y Wickham, H. (2017). *tibble: Simple Data Frames*. R package version 1.3.3.

Nelder, J. y Wedderburn, R. (1972). *Generalized Linear Models*. Chapman & Hall/CRC.

Nordhausen, K., Sirkia, S., Oja, H., y Tyler, D. E. (2015). *ICSNP: Tools for Multivariate Nonparametrics*. R package version 1.1-0.

Pinheiro, J. y Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer, New York.

Quanjer, P., Stanojevic, S., y Cole, T. (2012). Multiethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *The European Respiratory Journal*, 40(6):1324–1343.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numer. Math.*, 10(3):177–183.

Rigby, R. y Stasinopoulos, D. (2001). The GAMLSS project a flexible approach to statistical modelling.

Rigby, R. A. y Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

Rosenthal, M. y Bain, S. (1993). Lung function in white children aged 4 to 19 years. *Spirometry Thorax*, 48:794–802.

Royston, P. y Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3):429–467.

Royston, P. y Wright, E. (2000). Goodness of fit statistics for age-specific reference intervals. *Statistics in Medicine*.

RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Ruppert, D., Wand, M. P., y Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Spriggs, E. (1978). The history of Spirometry. *Br f Dis Chest*, 72(165).

Stanojevic, S., Wade, A., y Cole, T. (2007). Reference ranges for spirometry across all ages: a new approach. *American Journal of Respiratory and Critical Care Medicine*, 177(3):253–260.

Stasinopoulos, D. y Rigby, R. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7).

Venables, W. N. y Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth ediciISBN 0-387-95457-0.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.6.3.

Wickham, H., Francois, R., Henry, L., y Müller, K. (2017a). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.1.

Wickham, H., Hester, J., y Francois, R. (2017b). *readr: Read Rectangular Text Data*. R package version 1.1.1.

Lista de Abreviaturas

(AIC) Akaike information criterion

(ANOVA) Analysis of variance

(ANACOVA) Analysis of covariances

(ATS) American Thoracic Society

(BCPE) Box Cox Power Exponential

(BIC) Bayesian information criterion

(DG) Desvianza Global

(DF) Degree of freedom

(EF) Exponential Family

(EMC) Estimador Mínimos Cuadrados

(EMV) Estimador Máximo Verosímil

(exGAUS) exponential Gaussian

(FCCEEyA) Facultad de Ciencias Económicas y de Administración

(FEF) Flujo Espiratorio Forzado

(FEV) Forced Expiratory Volume

(FEV1) Forced Expiratory Volume in 1 second

(FIF) Flujo Inspiratorio Forzado

(GAIC) Generalized Akaike Information Criterion

(GAM) Generalized Additive Models

(GAMM) Generalized Additive Mixed Models

(GAMLSS) Generalized Additive Models for Localization, Scale and Shape

(GCV) Generalized Cross Validation

(GD) Global Deviance

(GLI) Global Lung Iniciative

(GLM) Generalized Linear Models

(GLMM) Generalized Linear Mixed Models

(IESTA) Instituto de estadística

(LRT) Likelihood Ratio Test

(MV) Máxima verosimilitud

(NO) Distribución normal

(PFE) Pico de Flujo Espiratorio

(QVP) Quasi Verosimilitud Penalizada

(REML) Restricted Maximum Likelihood

(RS) Rigby Stasinopoulos algorithm

(RSS) Residual Sum Squares

(SBC) Schwartz Bayesian Criterion

(SCE) Suma Cuadrado de los Errores

(SEP4) Skewed Exponential Power type 4

(SN2) Skewed Normal type 2

(TGD) Test Global Deviance

(UdelaR) Universidad de la república

(VGD) Validation Global Deviance

Apéndice A

Apéndice Estadístico

Funciones de ajuste de densidades

Función `fitDist` (gamlss)

La función usa `gamlssML()` para ajustar todas las funciones de distribución paramétricas relevantes de la gamlss.family a un vector de datos. El modelo final es uno que es seleccionado por el criterio GAIC (Generalized Akaike Information Criterion) con penalización k, que se obtiene agregando una penalización fija a la desvianza global (Global Deviance) (k=2 corresponde al Akaike's Information Criterion, k=log(n) al Schwarz Bayesian Criterion).

```
fitDist(y, k=2, type="realAll", try.gamlss=FALSE, extra=NULL,
data=NULL, . . .)
```

`...` para incorporar argumentos extra para `gamlssML()` o `gamlss()`.

`try.gamlss` indica que si el algoritmo no converge, o tiene algún problema con la función `gamlssML()`, que utilice la función `gamlss()` para el ajuste.

Esta es una función para ajustar una familia de distribución gamlss a un conjunto de datos usando un algoritmo de maximización no lineal en R. Esto es relevante solo cuando no hay variables explicativas. De hecho, utiliza la función interna `MLE()` que es una copia de la función `mle()` del paquete **stat4**. La función `gamlssML()` puede ser más rápida para datos grandes que la función equivalente `gamlss()` que es diseñada para modelos de regresión.

Función `fitdistr` (MASS)

Realiza el ajuste por máxima verosimilitud de distribuciones univariadas, y permite mantener fijos los parámetros que se deseen.

```
fitdistr(x, densfun, start, . . .)
```

`...` parámetros adicionales, tanto para `densfun` o para `optim`. En particular, puede ser usado para especificar bordes via *lower* o *upper* o ambos.

`densfun`: las distribuciones que admite son las siguientes: beta, cauchy, chi-cuadrado, exponencial, F, gamma, geométrica, log-Normal, logística, binomial negativa, normal, poisson, t y weibull.

`start`: una lista con nombre que da los parámetros a ser optimizados con sus valores iniciales. Para algunas funciones puede ser omitido, pero para otras debe aparecer.

Para las funciones normal, log-Normal, Geométrica, Exponencial y Poisson, se usa la forma cerrada de MLE, y no debería usarse la lista con los parámetros.

Para el resto de las distribuciones, se usa la función `optim()` para optimizar la función de log-verosimilitud. Los errores estándar estimados son tomados de la matriz de información observada, calculada por aproximación numérica. Para problemas unidimensionales, se usa el método de Nelder-Mead, y para multidimensionales el

método BFGS, a no ser que los argumentos `lower` o `upper` sean suministrados (cuando L-BFGS-B es usado) o `method` es explicitado.

Devuelve la estimación de los parámetros, los errores estándar estimados, la matriz estimada de varianza y covarianza, y el log-verosimilitud.

Comparación entre `fitDist()` y `fitdistr()`

Para poder ver las diferencias entre la función `fitDist()` del paquete **gamlss** y la función `fitdistr()` del paquete **MASS** se presenta como queda para el caso de la variable CVF, se aplica cada función a la variable CVF, al conjunto de datos que contiene a todos los niños (tanto niños alérgicos como normales) para poder visualizarlas.

Se empieza por la función incluída en el paquete **MASS**.

```
distCVF=fitdistr(datos$CVF, "lognormal")
```

Con esto, se almacena en el objeto `distCVF` las estimaciones de la media y del desvío estándar, los desvíos de las estimaciones, la matriz de covarianzas y la verosimilitud de la distribución deseada, en este caso log-Normal, usando los datos de la variable CVF para todos los niños.

A continuación se muestra como se utiliza la función `fisDist()`, que está incluída en el paquete **gamlss**.

```
fDcvf=fitDist(CVF, type="realline", try.gamlss = TRUE,
data=datos)
```

En el objeto `fDcvf` se almacena, entre otras cosas, las estimaciones de cada parámetro de la familia de distribución ajustada, sus desvíos estándar, los valores de *deviance* global, AIC y SBC, y un resumen de los cuantiles residuales, para comparar

el ajuste con una distribución normal estandarizada.

La Figura A.1 muestra la densidad empírica no paramétrica de la variable de respuesta CVF, la densidad de la distribución log-Normal, cuyos parámetros fueron estimados con la función `fitdistr()` del paquete **MASS** y la densidad de la familia de distribución skewed t tipo 5 con parámetros estimados con la función `fitDist()`.

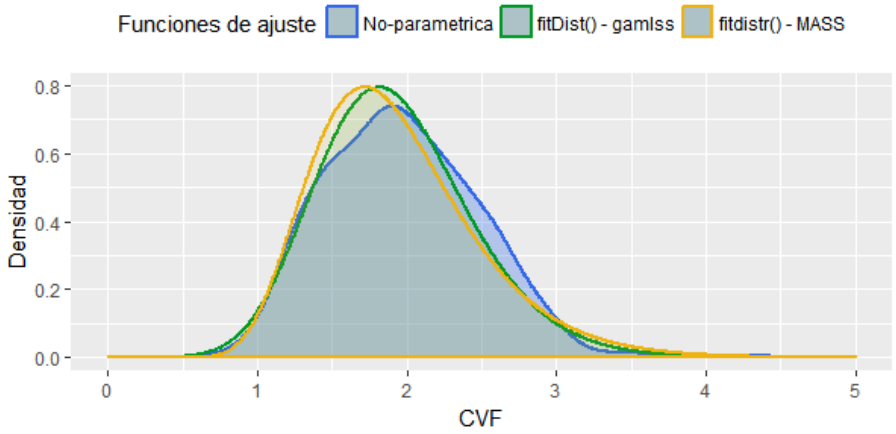

Figura A.1: *Comparación de las densidades ajustadas con las funciones `fitdistr()` y `fitDist()` en contraste con la densidad estimada para CVF con método no paramétrico.*

La función de densidad de la distribución skewed t type 5 (ST5), propuesta por Jones y Faddy en el 2003 (pág. 162), es de la forma:

$$f_Y\left(y \mid \mu, \sigma, \nu, \tau\right) = \frac{c}{\sigma}\left[1 + \frac{z}{\left(a+b+z^2\right)^{1/2}}\right]^{a+1/2}\left[1 + \frac{z}{\left(a+b+z^2\right)^{1/2}}\right]^{b+1/2}$$

para $\infty \leq y \leq \infty$, donde μ y $\nu \in \mathbb{R}$, σ y $\tau > 0$, y donde $z = \left(y-\mu\right)/\sigma$, $c = \left[2^{a+b-1}\left(a+b\right)^{1/2} B(a,b)\right]^{-1}$ y $\nu = \left(a-b\right)/\left[ab\left(a+b\right)\right]^{1/2}$ y $\tau = 2/\left(a+b\right)$. Aquí $E\left(Y\right) = \mu + \sigma E\left(Z\right)$ con $E\left(Z\right) = \left(a-b\right)\left(a+b\right)^{1/2}\Gamma\left(a-1/2\right)\Gamma\left(a-1/2\right)/\left[2\Gamma\left(a\right)\Gamma\left(b\right)\right]$ y $Var\left(Y\right) = \sigma^2 Var\left(Z\right)$, con $E\left(Z^2\right) = \left(a+b\right)\left[\left(a-b\right)^2 + a + b - 2\right]/\left[4\left(a-1\right)\left(b-1\right)\right]$.

Comparativamente, se puede ver que la función del paquete **gamlss**, al ser más flexible, ajusta mejor la distribución que la función de la librería **MASS**. Esto se debe a las limitaciones que ésta última tiene, que solo admite ciertas familias de distribución.

Funciones de diagnóstico de modelo

La función `plot.gamlss()`

El nombre de la función es `plot.gamlss()`, pero ya que es una función de método en **R**, se la puede llamar usando sólo `plot()`, siempre que su primer argumento sea un objeto `gamlss` ajustado. La función produce cuatro gráficos para analizar los cuantiles residuales, definidos anteriormente, de un objeto `gamlss`. La aleatorización tiene lugar cuando la variable de respuesta es discreta o mixta y también para datos censurados o en intervalos. Los cuatro gráficos son:

- residuos contra los valores ajustados del parámetro μ.
- residuos contra un índice o una variable especificada.
- la densidad de los residuos estimada por núcleo.
- un gráfico QQ-normal de los residuos

Al mismo tiempo que la función muestra los gráficos, provee un resumen de los cuantiles residuales, en el cual se encuentra la media, la varianza, el coeficiente de asimetría, el coeficiente de curtosis y el *coeficiente de correlación de Filliben*. Para comparar estas medidas con las de una distribución normal estandarizada, se debe contrastar con los valores (0,1,0,3,1), respectivamente.

Coeficiente de correlación de Filliben

Esta prueba utiliza el coeficiente de correlación r entre las observaciones ordenadas X_i y los correspondientes cuantiles ajustados M_i determinado por las posiciones p_i para cada X_i.

Se asume que las observaciones podrían haber sido extraídas de la distribución ajus-

```
## ******************************************************************
##        Summary of the Quantile Residuals
##                             mean   =  0.0009096
##                         variance   =  1.002
##                 coef. of skewness  =  -0.008444
##                 coef. of kurtosis  =  2.993
## Filliben correlation coefficient   =  0.9993
## ******************************************************************
```

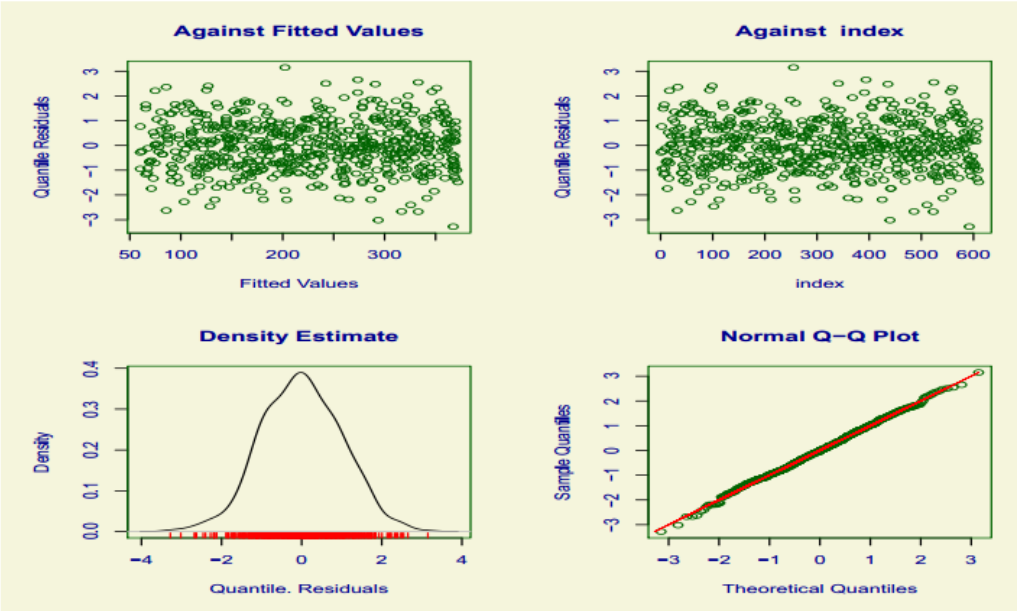

Figura A.2: *Ejemplo de la gráfica de la función* `plot.gamlss()` *(abajo) junto al resumen de los cuantiles residuales (arriba). Fuente: Flexible Regression and Smoothing The GAMLSS packages in R, Stasinopoulos - 2015*

tada si el valor de r es cercano a 1. Esencialmente, r mide la linealidad del gráfico de probabilidad, proporcionando una evaluación cuantitativa del ajuste.

El coeficiente de correlación r viene dado por

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(M_i - \bar{M})^2}} \tag{A.2.1}$$

Donde $\bar{X}$ y $\bar{M}$ denotan los valores medios de las observaciones X_i y los cuantiles ajustados M_i, respectivamente, y n es el tamaño de la muestra. En (A.2.1), Filliben utilizó la estimación de la mediana estadística para M_i como se muestra a continuación,

$$M_i = \phi^{-1}(m_i) \tag{A.2.2}$$

donde $\phi^{-1}(.)$ es la inversa de la función de distribución acumulada de una normal estándar y m_i es el valor de la mediana propuesto por

$$\begin{aligned} m_i &= 1 - (0{,}5)^{1/n}, i = 1; \\ m_i &= (i - 0{,}3175)/(n + 0{,}365), i = 2, ..., n-1; \\ m_i &= (0{,}5)^{1/n}, i = n \end{aligned} \tag{A.2.3}$$

Gráficos de gusano

(Buuren y Fredriks, 2001) introdujeron los *worm plots* de los residuos para identificar las regiones (intervalos) de una variable explicativa dentro de la cual el modelo no ajusta adecuadamente con los datos (llamada "violación del modelo"). La función `wp()` de la librería **gamlss** contiene una implementación en R que provee uno o *varios worm* plots para modelos GAMLSS ajustados. Esta es una herramienta de diagnóstico para analizar los residuos para diferentes rangos (por defecto no superpuestos) de una o dos variables explicativas. Son QQ-plots sin tendencia y su

nombre proviene de la forma de su gráfico (Figura A.3)

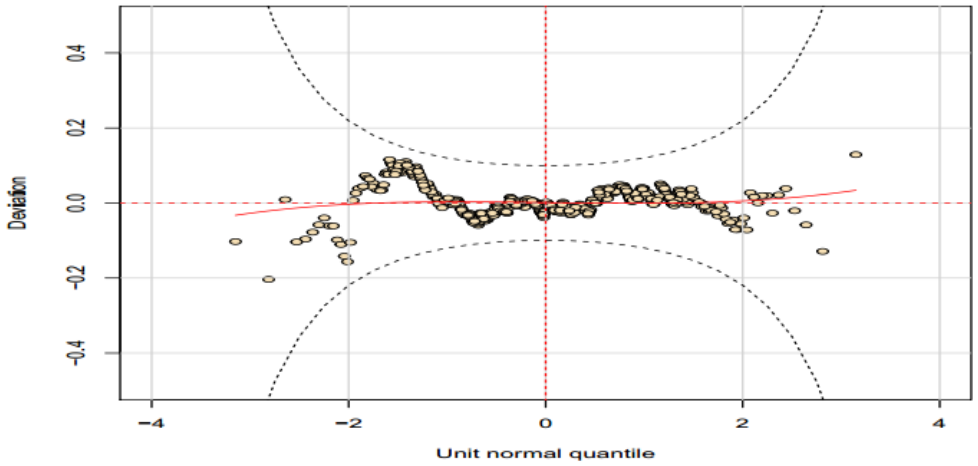

Figura A.3: *Ejemplo de un worm plot. Fuente: Flexible regression and smoothing. The GAMLSS packages in R. Stasinopoulos, Rigby - 2015.*

Hay varias características a destacar en la Figura A.3:

- Los puntos dorados (o el gusano) del gráfico: Estos puntos muestran que tanto se alejan los residuos de sus valores esperados, representados en el gráfico por la línea horizontal punteada roja.

- Las regiones de confianza del 95 % para los puntos dadas por las dos curvas elípticas en el centro de la figura: Si el modelo es correcto, esperamos que aproximadamente el 95 % de los puntos estén entre las dos curvas elípticas y el 5 % fuera. Un mayor porcentaje de los puntos fuera de las dos curvas elípticas indica que la distribución ajustada (o los términos ajustados) del modelo son inadecuados para explicar la variable de respuesta.

- La curva roja ajustada a los datos: Esta curva es un ajuste cúbico a los puntos del gusano. La forma de este ajuste cúbico refleja diferentes deficiencias en el modelo. Estos se describen en la Tabla A.1 y se ilustran en la Figura A.4.

El punto importante aquí es que las formas cuadráticas y cúbicas en un worm

plot indican la presencia de asimetría y curtosis respectivamente en los residuos. En cuanto a la Figura A.3, ya que todas las observaciones caen en la región de "aceptación", dentro de las dos curvas elípticas, y no se detecta una forma específica en los puntos, el modelo general parece tener un buen ajuste.

Forma del gusano (o su curva ajustada)	Residuos	Variable de respuesta
Nivel: sobre el origen	media muy grande	parámetro de localización muy chico
Nivel: debajo del origen	media muy chica	parámetro de localización muy grande
Línea: pendiente positiva	varianza muy grande	parámetro de escala muy chico
Línea: pendiente negativa	varianza muy chica	parámetro de escala muy grande
Forma de U	asimetría positiva	asimetría muy baja
Forma de U invertida	asimetría negativa	asimetría muy alta
Forma de S positiva	lepto curtosis	curtosis muy chica
Forma de S negativa	platy curtosis	curtosis muy grande

Tabla A.1: *Las diferentes formas para el worm plot de los residuos (primera columna) y la deficiencia correspondiente en los residuos (segunda columna) y la deficiencia en la distribución de respuesta variable (tercera columna).Fuente: Flexible regression and smothing. The GAMLSS packages in R. Stasinopoulos, Rigby - 2015.*

Funciones de suavizado

Spline cúbico

Las funciones de spline cúbico `cs()` y `scs()` están basados en la función `smooth.spline()` de R y puede ser usada para un suavizado univariado. Los splines cúbicos han tenido una extensa cobertura en la literatura, como por ejemplo (Reinsch, 1967), (Green, 2000) y (Hastie y Tibshirani, 1986). Estos asumen en el modelo (2.1.14) que las funciones $h(t)$ son dos veces diferenciables continuamente y maximizamos una log-verosimilitud penalizada, dada por l sujeto a términos de penalización de la forma $\lambda \int_{-\infty}^{\infty} [h''(t)]^2 dt$. La solución para maximizar las funciones $h(t)$ son todas splines cúbicos naturales, y por lo tanto pueden expresarse como combinaciones lineales de

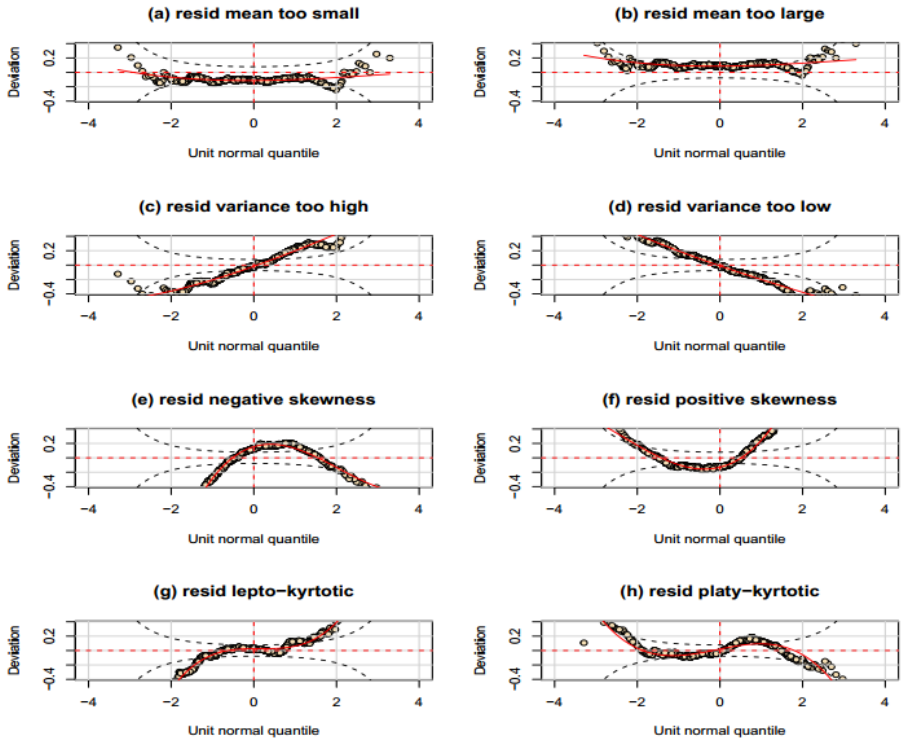

Figura A.4: *Diferentes tipos de fallas del modelo indicadas por el worm plot: i) las figuras (a) y (b) indican un fallo para un ajuste correcto del parámetro de localización, con puntos que caen por debajo y por encima de la línea punteada horizontal (roja). ii) las gráficas (c) y (d) indican un fallo para ajustar correctamente el parámetro de escala. iii) las gráficas (e) y (f) indican un fallo para modelar la asimetría en los datos correctamente y iv) las gráficas (g) y (h) indican fallo para modelar la curtosis. Fuente: Flexible regression and smothing. The GAMLSS packages in R. Stasinopoulos, Rigby - 2015.*

sus splines cúbicos base naturales (Boor, 2001). En las funciones `cs()` y `codescs()` cada valor distinto de x es un nodo. Estas difieren en como han sido implementadas y deberían producir resultados idénticos.

Coeficientes variables

Tanto `vc()` como `pvc()` son funciones de coeficientes variables. Los términos de coeficientes variables fueron introducidos por (Hastie y Tibshirani, 1993a) para acomodar un tipo especial de interacción entre variables explicativas. Esta interacción toma la forma de $\beta(r)x$, el coeficiente lineal de la variable explicativa x cambia suavemente de acuerdo a otra variable explicativa r. En algunas aplicaciónes, r puede ser el tiempo. En general, r debería ser una variable continua, mientras que x puede ser tanto contínua como categórica. En la implementación de la función `vc()`, x tiene que ser una variable contínua o una variable categórica, de dos niveles; 0 y 1. En la función `pvc()`, que usa B-splines penalizados, x puede ser un factor con más de dos niveles.

Splines penalizados

Las funciones `pb(), ps(), cy(), tp(), pvc()` se basan todas en B-splines penalizados. Los splines penalizados fueron introducidos por (Eilers y Marx, 1996). Los splines penalizados (o P-splines) son polinomios por piezas definidos por funciones bases de B-splines en la variable explicativa donde los coeficientes de las funciones de base son penalizados para garantizar suficiente suavidad, véase Eilers y Marx (1996).

Más precisamente, consideremos el modelo $\theta = Z(x)\gamma$ donde θ puede ser cualquier parámetro de una distribución de un modelo GAMLSS, $Z(x)$ es una matriz de diseño de $n \times q$ base para la variable explicativa x definida en q nodos dentro del rango de

x, y γ es un vector de coeficientes de $q \times 1$ que tiene ciertas restricciones estocásticas impuestas por el hecho de que $D_\theta \sim N_{q-r}(0, \lambda^{-1}I)$. D es una matriz $(q-r) \times q$ que da las r-ésimas diferencias del vector q-dimensional γ. Entonces para definir un spline penalizado necesitamos:

1. el número de nodos q en el eje x [`ps.intervals`].

2. el grado del polinomio a trozos usado en la base B-spline para poder definir $\mathbf{X}$, definido por el argumento `degree`.

3. r, el orden de diferencias de la matriz D indicando el tipo de penalización impuesta en los coeficientes de las funciones base de los B-splines, definido por el argumento `order`.

4. la cantidad de suavizado requerida definida por los grados de libertad deseados, definidos por el argumento `df` o alternativamente por el parámetro de suavizado mediante el argumento `lambda`.

Polinimos locales `loess`

La función `lo()` permite al usuario utilizar un ajuste `loess` en una formulación gamlss. Un ajuste `loess` es una curva polinómica determinada por una o más variables explicativas (contínuas), las cuales son ajustadas localmente.

Polinomios fraccionales

La función `fp()` es una implementación de los polinomios fraccionales introducidos por (Royston y Altman, 1994). Las funciones involucradas en `fp()` y `bfp()` se basan vagamente en la función `fracpoly()` dada por (Hastie y Tibshirani, 1993b). La función `bfp()` genera la matriz de diseño correcta para ajustar un polinimio de

potencia del tipo $b_0 + b_1 x^{p_1} + b_2 x^{p_2} + ... + b_k^{p_k}$. Para las potencias $p_1, p_2, ..., p_k$ dadas, a través del argumento `powers` en `bfp()`, la función puede ser usada para ajustar polinomios de potencia de la misma manera como las funciones `poly()` o `bs()` del paquete **splines** son usadas para ajustar polinomios ortogonales o por trozos respectivamente. La función `fp()` trabaja como un término aditivo de suavizado dentro de **gamlss**. Se utiliza para ajustar los mejores polinomios fraccionarios entre un conjunto específico de valores de potencia.

Polinomios de potencia

La función de polinomios de potencia `pp()` es una función experimental y está diseñada para las situación en cual el modelo es de la forma $b_0 + b_1 x^{p_1} + b_2 x^{p_2}$, con potencias p_1, p_2 para ser estimadas no linealmente por los datos. Se deben establecer valores iniciales para los parámetros.

Términos no lineales

La función `nl()` se encuentra en el paquete agregado **gamlss.nl** designado para ajustar modelos paramétricos no lineales dentro del entorno GAMLSS. Proporciona una forma de ajustar términos no lineales junto con términos lineales o de suavizado en el mismo modelo.

Efectos aleatorios

La función `random()` permite que los valores ajustados para un predictor de factores (categóricas) sean reducidos hacia la media global, donde la cantidad de reducción depende del parámetro λ o de los grados de libertad equivalentes (df). La función `ra()` es similar a la función `random()`, pero su procedimiento de ajuste se basa

en los mínimos cuadrados aumentados, hecho que hace que `ra()` sea más general, pero también más lento que `random()`. La función de coeficiente aleatorio `rc()` es experimental. Notar que las funciones `random(), ra()` y `rc()` son usadas para estimar los efectos aletorios γ *dados* los hiperparámetros λ.

Apéndice B

Tablas Estadísticas

Curvas normales de referencia

En esta sección se presentan los resultados de las diferentes curvas percetilares para los parámetros CVF y FEV_1.

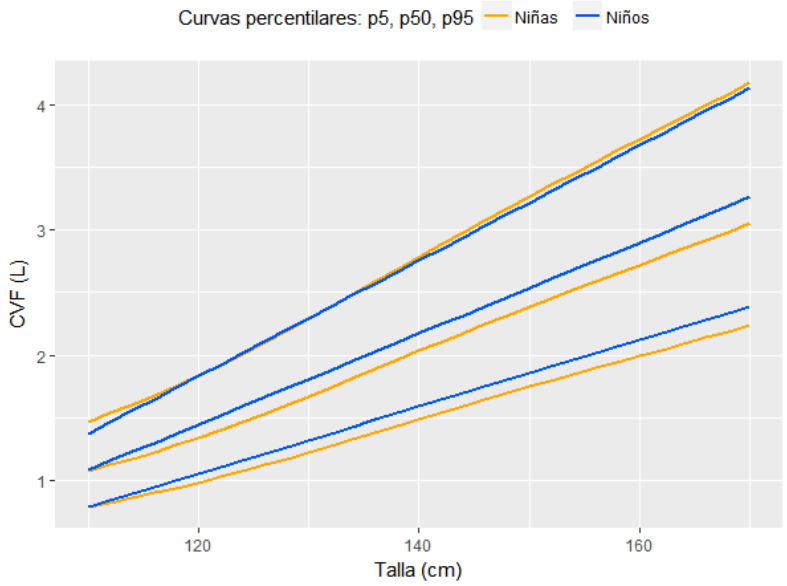

Figura B.1: *Comparación entre las curvas percentilares de niñas y niños del parámetro espirométrico CVF en relación a la variable regresora* Talla *(modelos m_cvf_103 y m_cvf_203).*

En la Figura B.1 se muestran las curvas percentilares de la variable CVF. En ella se observa que los niños tienen valores superiores a las niñas en casi la totalidad del recorrido entre los 110 y 160 cm, teniendo un percentil 95 similares entre ambos sexos.

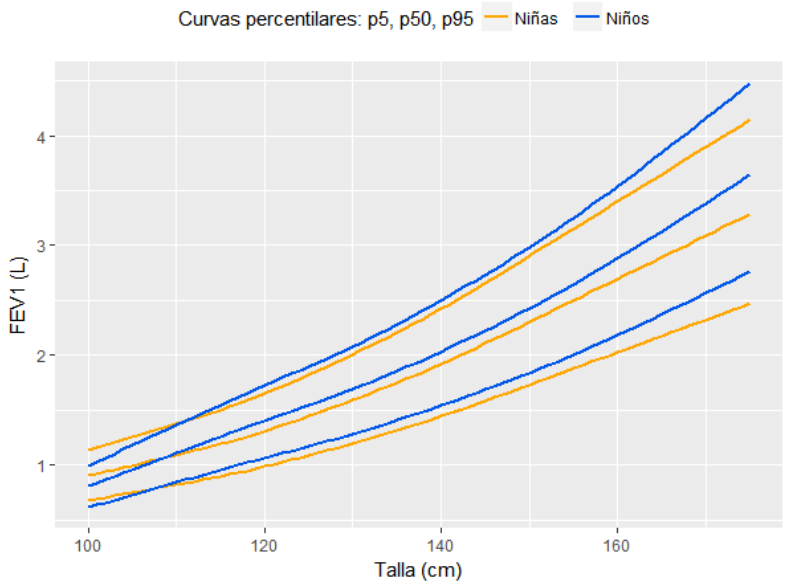

Figura B.2: *Comparación entre las curvas percentilares de niñas y niños del parámetro espirmétrico FEV_1 en relación a la variable regresora Talla (modelos m_fev_103 y m_fev_203).*

En la Figura B.2 se muestran las curvas percentilares de la variable FEV_1. En ella se observa que los niños tienen valores superiores a las niñas en gran parte del recorrido, en particular luego de los 120 cm de altura (Talla).

Tablas normales de referencia

A continuación se presentan las tablas con los valores estimado de los percentiles 5, 50 y 95 para las variable CVF y FEV_1.

	Percentiles niñas			Percentiles niños		
Talla	Percentil 5	Percentil 50	Percentil 95	Percentil 5	Percentil 50	Percentil 95
110.00	0.78	1.07	1.46	0.79	1.08	1.37
112.00	0.82	1.12	1.53	0.84	1.15	1.46
114.00	0.86	1.17	1.61	0.90	1.23	1.56
116.00	0.90	1.23	1.68	0.95	1.30	1.65
118.00	0.94	1.28	1.76	1.00	1.37	1.74
120.00	0.98	1.34	1.84	1.06	1.45	1.83
122.00	1.03	1.40	1.92	1.11	1.52	1.93
124.00	1.07	1.47	2.01	1.16	1.59	2.02
126.00	1.12	1.53	2.10	1.22	1.66	2.11
128.00	1.17	1.60	2.19	1.27	1.74	2.20
130.00	1.22	1.67	2.29	1.32	1.81	2.29
132.00	1.28	1.74	2.39	1.37	1.88	2.39
134.00	1.33	1.81	2.48	1.43	1.95	2.48
136.00	1.38	1.89	2.58	1.48	2.03	2.57
138.00	1.44	1.96	2.68	1.53	2.10	2.66
140.00	1.49	2.03	2.78	1.59	2.17	2.75
142.00	1.54	2.10	2.88	1.64	2.24	2.85
144.00	1.60	2.18	2.98	1.69	2.32	2.94
146.00	1.65	2.25	3.08	1.75	2.39	3.03
148.00	1.70	2.32	3.17	1.80	2.46	3.12
150.00	1.75	2.39	3.27	1.85	2.53	3.22
152.00	1.80	2.45	3.36	1.91	2.61	3.31
154.00	1.85	2.52	3.45	1.96	2.68	3.40
156.00	1.90	2.59	3.54	2.01	2.75	3.49
158.00	1.94	2.65	3.63	2.06	2.83	3.58
160.00	1.99	2.72	3.72	2.12	2.90	3.68
162.00	2.04	2.79	3.81	2.17	2.97	3.77
164.00	2.09	2.85	3.91	2.22	3.04	3.86
166.00	2.14	2.92	4.00	2.28	3.12	3.95
168.00	2.19	2.98	4.09	2.33	3.19	4.05
170.00	2.24	3.05	4.18	2.38	3.26	4.14

Tabla B.1: *Estimación de los percentiles 5, 50 y 95 para la variable CVF de niñas (m_cvf_103) y niños (m_cvf_203) a través de la variable* Talla.

	Percentiles niñas			Percentiles niños		
Talla	Percentil 5	Percentil 50	Percentil 95	Percentil 5	Percentil 50	Percentil 95
110.00	0.82	1.09	1.38	0.84	1.11	1.36
112.00	0.85	1.13	1.43	0.89	1.17	1.44
114.00	0.88	1.17	1.48	0.93	1.23	1.51
116.00	0.91	1.21	1.53	0.98	1.29	1.58
118.00	0.95	1.26	1.59	1.02	1.35	1.66
120.00	0.98	1.31	1.65	1.07	1.40	1.73
122.00	1.02	1.36	1.72	1.11	1.46	1.80
124.00	1.07	1.42	1.79	1.15	1.52	1.86
126.00	1.11	1.47	1.86	1.19	1.57	1.93
128.00	1.15	1.53	1.93	1.24	1.63	2.01
130.00	1.20	1.59	2.01	1.28	1.69	2.08
132.00	1.24	1.65	2.09	1.33	1.76	2.16
134.00	1.29	1.72	2.16	1.38	1.82	2.24
136.00	1.34	1.78	2.25	1.43	1.89	2.32
138.00	1.39	1.85	2.33	1.49	1.96	2.41
140.00	1.44	1.92	2.42	1.54	2.03	2.50
142.00	1.50	1.99	2.51	1.60	2.11	2.59
144.00	1.55	2.07	2.61	1.66	2.19	2.69
146.00	1.61	2.15	2.71	1.72	2.26	2.78
148.00	1.67	2.22	2.81	1.78	2.34	2.88
150.00	1.73	2.30	2.90	1.84	2.43	2.98
152.00	1.79	2.38	3.00	1.90	2.51	3.09
154.00	1.85	2.46	3.10	1.97	2.60	3.19
156.00	1.91	2.54	3.20	2.04	2.69	3.31
158.00	1.97	2.62	3.30	2.11	2.78	3.42
160.00	2.03	2.70	3.40	2.18	2.88	3.54
162.00	2.09	2.77	3.50	2.26	2.98	3.66
164.00	2.15	2.85	3.60	2.33	3.08	3.78
166.00	2.20	2.93	3.70	2.41	3.18	3.91
168.00	2.26	3.01	3.80	2.49	3.28	4.03
170.00	2.32	3.09	3.90	2.57	3.38	4.16

Tabla B.2: *Estimación de los percentiles 5, 50 y 95 para la variable FEV_1 de niñas (m_fev_103) y niños (m_fev_203) a través de la variable Talla.*

Tablas de IMC y Talla por Edades

A continuación se presentan las tablas de IMC y Talla por Edad resultantes del macro para R del programa Anthro de la OMS.

Tabla B.3: *Talla por edad*

Edad	N	% < -3 SD	LI	LS	% < -2 SD	LI	LS	% > +1 SD	LI	LS	% > +2 SD	LI	LS	% > +3 SD	LI	LS	Media	SD
6 a 12	878	0,1	0	0,4	1,5	0,6	2,3	19,2	16,6	21,9	4,2	2,8	5,6	0,7	0,1	1,3	0,11	1,06
6	97	0	0	0,5	1	0	3,6	23,7	14,7	32,7	6,2	0,9	11,5	0	0	0,5	0,19	1,12
7	118	0	0	0,4	2,5	0	5,8	20,3	12,7	28	5,1	0,7	9,5	0	0	0,4	0,06	1,05
8	143	0	0	0,3	0,7	0	2,4	21	14	28	4,2	0,6	7,8	1,4	0	3,7	0,2	1,08
9	177	0	0	0,3	1,7	0	3,9	14,1	8,7	19,5	4	0,8	7,1	1,1	0	3	0,05	1,03
10	169	0	0	0,3	0,6	0	2	20,7	14,3	27,1	3	0,1	5,8	0,6	0	2	0,16	1,01
11	173	0,6	0	2	2,3	0	4,8	18,5	12,4	24,6	4	0,8	7,3	0,6	0	2	0,06	1,07
12	1	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	-0,01	
Niños																		
6 a 12	412	0,2	0	0,8	1	0	2	19,7	15,7	23,6	5,6	3,2	7,9	1,2	0	2,4	0,19	1,08
6	41	0	0	1,2	0	0	1,2	26,8	12	41,6	12,2	1	23,4	0	0	1,2	0,39	1,2
7	60	0	0	0,8	1,7	0	5,7	20	9	31	8,3	0,5	16,2	0	0	0,8	0,08	1,13
8	67	0	0	0,7	0	0	0,7	26,9	15,5	38,2	6	0	12,4	3	0	7,8	0,35	1,21
9	79	0	0	0,6	1,3	0	4,4	15,2	6,6	23,7	3,8	0	8,6	1,3	0	4,4	0,08	0,99
10	83	0	0	0,6	0	0	0,6	19,3	10,2	28,4	3,6	0	8,2	1,2	0	4,2	0,28	0,94
11	81	1,2	0	4,3	2,5	0	6,5	14,8	6,5	23,2	3,7	0	8,4	1,2	0	4,3	0,05	1,09
12	1	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	-0,01	
Niñas																		
6 a 12	466	0	0	0,1	1,9	0,6	3,3	18,9	15,2	22,5	3	1,3	4,7	0,2	0	0,7	0,05	1,03
6	56	0	0	0,9	1,8	0	6,1	21,4	9,8	33,1	1,8	0	6,1	0	0	0,9	0,05	1,05
7	58	0	0	0,9	3,4	0	9	20,7	9,4	32	1,7	0	5,9	0	0	0,9	0,05	0,96
8	76	0	0	0,7	1,3	0	4,5	15,8	6,9	24,6	2,6	0	6,9	0	0	0,7	0,07	0,95
9	98	0	0	0,5	2	0	5,4	13,3	6	20,5	4,1	0	8,5	1	0	3,5	0,02	1,06
10	86	0	0	0,6	1,2	0	4	22,1	12,7	31,4	2,3	0	6,1	0	0	0,6	0,05	1,07
11	92	0	0	0,5	2,2	0	5,7	21,7	12,8	30,7	4,3	0	9,1	0	0	0,5	0,06	1,06

Tabla B.4: *Índice de Masa Corporal por Edad*

Edad	N	% < -3 SD	LI	LS	% < -2 SD	LI	LS	% > +1 SD	LI	LS	% > +2 SD	LI	LS	% > +3 SD	LI	LS	Media	SD
6 a 12	873	0,2	0	0,6	0,8	0,2	1,5	42,3	38,9	45,6	14,7	12,3	17,1	2,5	1,4	3,6	0,83	1,16
6	97	0	0	0,5	1	0	3,6	41,2	30,9	51,5	11,3	4,5	18,2	3,1	0	7,1	0,84	1,1
7	117	0	0	0,4	0	0	0,4	41	31,7	50,4	12	5,7	18,3	3,4	0	7,1	0,89	1,11
8	141	0	0	0,4	0	0	0,4	43,3	34,7	51,8	18,4	11,7	25,2	5	1	8,9	0,94	1,21
9	176	0	0	0,3	1,1	0	3	46	38,4	53,7	15,3	9,7	20,9	1,1	0	3	0,87	1,08
10	169	0,6	0	2	1,2	0	3,1	44,4	36,6	52,2	14,2	8,6	19,8	2,4	0	5	0,82	1,21
11	172	0,6	0	2	1,2	0	3,1	36,6	29,1	44,1	15,1	9,5	20,8	1,2	0	3,1	0,64	1,19
12	1	0	0	50	0	0	50	100	50	100	0	0	50	0	0	50	1,52	
Niños																		
6 a 12	407	0,2	0	0,8	0,5	0	1,3	46,2	41,2	51,2	15,5	11,8	19,1	4,4	2,3	6,5	0,94	1,19
6	41	0	0	1,2	0	0	1,2	39	22,9	55,2	9,8	0	20,1	4,9	0	12,7	0,9	1,02
7	59	0	0	0,8	0	0	0,8	49,2	35,5	62,8	16,9	6,5	27,4	6,8	0	14	1,15	1,18
8	65	0	0	0,8	0	0	0,8	44,6	31,8	57,5	18,5	8,3	28,7	9,2	1,4	17	1,03	1,32
9	78	0	0	0,6	1,3	0	4,4	52,6	40,8	64,3	17,9	8,8	27,1	1,3	0	4,4	1,04	1,06
10	83	1,2	0	4,2	1,2	0	4,2	50,6	39,2	62	15,7	7,2	24,1	3,6	0	8,2	0,91	1,29
11	80	0	0	0,6	0	0	0,6	37,5	26,3	48,7	12,5	4,6	20,4	2,5	0	6,5	0,65	1,18
12	1	0	0	50	0	0	50	100	50	100	0	0	50	0	0	50	1,52	
Niñas																		
6 a 12	466	0,2	0	0,7	1,1	0	2,1	38,8	34,3	43,4	13,9	10,7	17,2	0,9	0	1,8	0,73	1,11
6	56	0	0	0,9	1,8	0	6,1	42,9	29	56,7	12,5	2,9	22,1	1,8	0	6,1	0,8	1,16
7	58	0	0	0,9	0	0	0,9	32,8	19,8	45,7	6,9	0	14,3	0	0	0,9	0,62	0,97
8	76	0	0	0,7	0	0	0,7	42,1	30,3	53,9	18,4	9	27,8	1,3	0	4,5	0,86	1,1
9	98	0	0	0,5	1	0	3,5	40,8	30,6	51,1	13,3	6	20,5	1	0	3,5	0,74	1,09
10	86	0	0	0,6	1,2	0	4	38,4	27,5	49,2	12,8	5,2	20,4	1,2	0	4	0,73	1,13
11	92	1,1	0	3,7	2,2	0	5,7	35,9	25,5	46,2	17,4	9,1	25,7	0	0	0,5	0,64	1,21

Apéndice C

Código de R

```
#||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
#   Título: Elaboración de Patrones Espirométricos Normales
#           en niños uruguayos mediante modelos GAMLSS
#||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
#    1.Carga de base de datos y librerias
#-------------------------------------------------------------------------
# Librerias a utilizar:
library(readr)
library(tibble)
library(tidyr)
library(dplyr)
library(ggplot2)
library(gamlss)
library(ICSNP) #Test multivariados no paramétricos
library(MASS)
# Cargo la base de datos
datos <- read_delim("Database.csv",delim = ";")
parse_factor(datos$Alergicos,levels=c("No","Si"))
# Obtengo las variables necesarias, creo el índice de Gaënsler
# y filtro por Edad
datos <- datos %>%
  dplyr::select(Alergicos, Escuela, ContFab, Fuman, Sexo,
                EdadMeses, Edad, Talla,
                Peso, CVF, FEV1, FEF2575, PFE) %>%
  mutate(IGaensler=FEV1/CVF) %>%
  filter(!is.na(Talla) &
           !is.na(Peso) &
           !is.na(CVF) &
```

```
            !is.na(FEV1) &
            !is.na(FEF2575) &
            !is.na(PFE) &
            CVF > 0 & CVF < 6 &
            FEV1 > 0 & FEV1 < 6 &
            6 <= Edad & Edad <= 12)
#-------------------------------------------------------------
#    2. Descripción y exploración de datos
#-------------------------------------------------------------
# Relaciones entre Edad, Talla y Peso
gTallaPeso = ggplot(data=datos, aes(x=Talla, y=Peso))
p1 <- gTallaPeso + geom_jitter(aes(color=Sexo))+
  stat_smooth(se=TRUE,method = "loess")+
  xlab("Talla (cm)")+
  ylab("Peso (Kg)") + scale_color_manual(values = c(paleta[4],
                                                    paleta[2]))
# Para completar los gráficos, variar x e y en aes(),
# las etiquetas y el nombre del objeto p1 (p2 y p3)
#-------------------------------------------------------------
#    3. Estudio de grupo: Alérgicos y normales
#-------------------------------------------------------------
# Divido los datos entre los niños alérgicos y los normales
alergicos <- datos %>% filter(Alergicos=="Si")
normales <- setdiff(datos, alergicos)
# Test multivariado entre grupos: alérgicos y normales
HotellingsT2(normales[,10:13],alergicos[,10:13])
# La hipótesis nula H0) de este test es que la diferencia
# entre las medias es nula, con un p-valor<0.05,
# entonces rechazo H0)

# Matriz de varianzas y covarianzas de cada grupo
cov(normales[,10:13])
cov(alergicos[,10:13])
#-------------------------------------------------------------
#    4. Pruebas de robustez de los parámetros espirométricos
#-------------------------------------------------------------
# Separo los datos de los niños normales por Sexo:
nenas <- normales %>% filter(Sexo=="F")
nenes <- normales %>% filter(Sexo="M")
# Parámetro: CVF
#              a - Total
#              b - Niñas
#              c - Niños
# a -
```

```
# data frame donde se almacenaran los resultados
# de cada iteración
ResultsCVF=data_frame(Familia=col_character(),
                      AIC=col_double(),
                      mu=col_double(),
                      sigma=col_double(),
                      nu=col_double(),
                      tau=col_double())
# Fracción de muestra
fs <- 0.8
# Tamaño de muestra
nS <- floor(nrow(normales)*fs)
# Se define cantidad de iteraciones
iter<-1000
# Por cada iteración guarda en cada fila los valores
for (i in 1:iter){
  muestra=sample_n(normales, nS)
  AjDist=fitDist(CVF, type = "realline", try.gamlss = TRUE,
                 data=muestra)
  if (i==1){
    if (AjDist$df.fit == 4) {
      res <- data_frame(Familia=AjDist$family[1],
                        AIC=AjDist$aic,
                        mu=AjDist$mu,
                        sigma=AjDist$sigma,
                        nu=AjDist$nu,
                        tau=AjDist$tau)
      ResultsCVF <- res
    }else if (AjDist$df.fit == 3){
      res <- data_frame(Familia=AjDist$family[1],
                        AIC=AjDist$aic,
                        mu=AjDist$mu,
                        sigma=AjDist$sigma,
                        nu=AjDist$nu,
                        tau=NA)
      ResultsCVF <- res
    }else if (AjDist$df.fit == 2){
      res <- data_frame(Familia=AjDist$family[1],
                        AIC=AjDist$aic,
                        mu=AjDist$mu,
                        sigma=AjDist$sigma,
                        nu=NA,
                        tau=NA)
      ResultsCVF <- res
```

```
    }
  }else{
    if (AjDist$df.fit == 4) {
      res <- data_frame(Familia=AjDist$family[1],
                        AIC=AjDist$aic,
                        mu=AjDist$mu,
                        sigma=AjDist$sigma,
                        nu=AjDist$nu,
                        tau=AjDist$tau)
      ResultsCVF<-bind_rows(ResultsCVF,res)
    }else if (AjDist$df.fit == 3){
      res <- data_frame(Familia=AjDist$family[1],
                        AIC=AjDist$aic,
                        mu=AjDist$mu,
                        sigma=AjDist$sigma,
                        nu=AjDist$nu,
                        tau=NA)
      ResultsCVF<-bind_rows(ResultsCVF,res)
    }else if (AjDist$df.fit == 2){
      res <- data_frame(Familia=AjDist$family[1],
                        AIC=AjDist$aic,
                        mu=AjDist$mu,
                        sigma=AjDist$sigma,
                        nu=NA,
                        tau=NA)
      ResultsCVF<-bind_rows(ResultsCVF,res)
    }
  }
}
# frecuencia absoluta de cada familia
frec <- ResultsCVF %>%
  count(Familia) %>%
  arrange(Familia)
# Resumen de los resultados
resultsCVF <- ResultsCVF %>%
  group_by(Familia) %>%
  summarise(AIC_m=round(mean(AIC),2),
            media_mu=round(mean(mu),2),sd_mu=round(sd(mu),2),
            media_s=round(mean(sigma),2),sd_s=round(sd(sigma),2),
            media_nu=round(mean(nu),2),sd_nu=round(sd(nu),2),
            media_tau=round(mean(tau),2),sd_tau=round(sd(tau),2))
resultsCVF <- bind_cols(resultsCVF,frec[,2]/nrow(ResultsCVF))
resultsCVF <- rename(resultsCVF,fr.rel=n)
# muestra tabla de resumenes ordenada por frecuencia relativa
```

```
resultsCVF <- resultsCVF %>% arrange(desc(fr.rel)); resultsCVF
# gráfica la frecuencia relativa para CVF
ggresultsCVF = ggplot(data=resultsCVF, aes(Familia,fr.rel))
ggresultsCVF + geom_bar(stat="identity",fill=paleta[2]) +
  xlab("Familia de Distribución") +
  ylab("Frecuencia relativa")
# Se repite el procedimiento para FEV1, y además para
# cada sexo por separado en ambas variables (CVF y FEV1)
#-------------------------------------------------------------
#    5. Comparacion entre fitDist() y fitdistr()
#-------------------------------------------------------------
# Aplico las funciones a los datos
#    fitdistr() [MASS]
distCVF=fitdistr(datos$CVF, "lognormal")
# Extraigo los parámetros estimados
fd<-distCVF$estimate
#    fitDist() [gamlss]
fDcvf=fitDist(CVF, type="realline", try.gamlss = TRUE,
              data=datos)
fD<-data_frame(mu=fDcvf$mu,
               sigma=fDcvf$sigma,
               nu=fDcvf$nu,
               tau=fDcvf$tau,
               familia=fDcvf$family[1])
# Eje x
x <- seq(from=0,to=5,length.out = nrow(datos))
# Evalúa la densidad de ST5 sobre x
distST5 <- dST5(x, fD$mu, fD$sigma, fD$nu, fD$tau)
# Evalúa la densidad log-Normal sobre x
logN <- dlnorm(x,fd[1],fd[2])
# La densidad de CVF no paramétrica
denCVF<- density(datos$CVF,n=nrow(datos))
# Juntos las evaluaciones
Compara<-data_frame(x,distST5,logN,CVF=denCVF$y)
# Grafico las tres densidades en el mismo eje
ggComp <- ggplot(data=Compara,aes(x=x))
ggComp +
  geom_area(aes(y=CVF,colour="No-parametrica"),fill=paleta[2],
            alpha=0.3, size=1)+
  geom_area(aes(y=distST5, colour="fitDist() - gamlss"),
            fill=paleta[1], alpha=0.1, size=1)+
  geom_area(aes(y=logN, colour="fitdistr() - MASS"),
            fill=paleta[4], alpha=0.1, size=1)+
  scale_colour_manual(name="Funciones de ajuste",
```

```
                        breaks=c("No-parametrica",
                                 "fitDist() - gamlss",
                                 "fitdistr() - MASS"),
                        values = c("No-parametrica"=paleta[2],
                                   "fitDist() - gamlss"=paleta[1],
                                   "fitdistr() - MASS"=paleta[4]))+
  ylab("Densidad")+xlab("CVF")+
  theme(legend.position = "top")
#-------------------------------------------------------------------
#     6. Estado nutricional
#-------------------------------------------------------------------
datosEN <- datos %>% dplyr::select(Edad,Talla,Peso,Sexo)
datosEN$Age<-datosEN$Edad*12
head(datosEN)

who2007(FileLab = "Nutricional", FilePath = ".../Directorio.../who2007_R/",
        mydf= datosEN,sex = Sexo, age = Age, weight = Peso, height = Talla)
names(matz)
head(matz)
datosEN<-cbind(datosEN,matz[,6:13])

paleta <- c("#009925","#3369E8","#D50F25","#EEB211","#ffffff",
            "#332f2e")
densWHO <- ggplot(data=datosEN,aes(x=zbfa))
densWHO + geom_density(aes(color=Sexo),size=1)+
  ylab("Densidad")+xlab("Z-score de IMC")+
  scale_color_manual(values=c(paleta[4],paleta[2]))+
  theme(legend.position = "top")

summary(datosEN[,7:10])
#-------------------------------------------------------------------
#     7. Modelización CVF
#-------------------------------------------------------------------
# Separo los datos en dos conjuntos: de entrenamiento y
# de validación.
# Paso 1: Se establecen las fracciones para cada conjunto de datos
Tfr <- 0.8
Vfr <- 0.2
# Paso 2: Saco la muestra de entrenamiento y por diferencia la
# de validación
set.seed(1987) #para tener siempre la misma muestra
Train <- sample_n(normales, floor(nrow(normales)*Tfr))
Validation <- setdiff(normales,Train)
# Calculo de penalización SBC
```

```
SBC_t<-log(nrow(Train))
# Modelos iniciales
m_cvf_001 <- gamlss(CVF~pb(Talla), family = NO, data=Train)
m_cvf_002 <- gamlss(CVF~pb(Talla), family = SEP4, data=Train)
m_cvf_003 <- gamlss(CVF~pb(Talla), family = BCPE, data=Train)
# Grados efectivos de libertad de cada parámetro en cada modelo
edfAll(m_cvf_001)
# Comparación de AIC Generalizado entre modelos
GAIC(m_cvf_001,m_cvf_002,m_cvf_003,k=0) #Global deviance
GAIC(m_cvf_001,m_cvf_002,m_cvf_003,k=log(nrow(Train))) #SBC

m_cvf_013 <- gamlss(CVF~pb(Talla)+Sexo, family = BCPE,
                    data=Train) # Se agrega variable Sexo
m_cvf_023 <- gamlss(CVF~pb(Talla)+pb(Edad)+Sexo, family = BCPE,
                    data=Train) #Se agrega variable Edad

# Prueba de eliminación de términos para el modelo m_cvf_023
# con criterio SBC
drop1(m_cvf_023,k=SBC_t)
# Modelos con stepGAICAll.A (selección automática)
m_cvf_033 <- stepGAICAll.A(m_cvf_013,
                           scope = list(lower=~1,
                                        upper=~pb(Edad)+pb(Talla)+
                                          Edad+Talla+Sexo),
                           k=SBC_t,steps=10000)
# Para ver la fórmula del modelo en cada parámetro
m_cvf_033$call

# Construcción manual del modelo con selección automático para
#hacer prueba de eliminación de términos
m_cvf_033 <- gamlss(CVF~pb(Talla)+Sexo,
                    nu.formula = ~Sexo, data=Train, family=BCPE)
## Chequeo de los residuos
# Gráfico de gusano
wp(m_cvf_033)
# Cuantiles residuales
plot(m_cvf_033)
# Efecto de las variables sobre las funciones de suavizado
term.plot(m_cvf_033,parameter = "mu",pages=1,
          col.term = paleta[2])
term.plot(m_cvf_033,parameter = "nu",pages=1,
          col.term = paleta[2])
# Para obtener información sobre los términos de suavizado, se
# puede utilizar la función getSmo()
```

```
getSmo(m_fev_103)
# Brinda el tipo de función, los grados de libertad y el parámetro
# lambda de suavizado, en este caso para un P-spline

# Para comparar las curvas percentilares de distintos modelos, en
# este caso entre dos modelos anidados de la variable FEV1 para
# niñas con distribución BCPE
centiles.com(m_fev_123,m_fev_103, xvar=Train$Talla,
             cent = c(5,50,99),
             ylab="FEV1 (L)", xlab = "Talla (cm)")
# el argumento 'cent' establece los percentiles deseados
#-------------------------------------------------------------
#    8. Gráficos de Curvas percentilares
#-------------------------------------------------------------
# construyo grilla de valores de talla de 110cm a 170cm cada 2
# unidades
xTalla <- seq(from=110,to=170, by=2)
# CVF
# Predicción de los percentiles 5%, 50% y 95% de los modelos
#m_cvf_103 y m_cvf_203 en los valores de la grilla
cent.cvf.f <- centiles.pred(m_cvf_103, xname="Talla",
                            xvalues=xTalla,
                            cent=c(5,50,95),type = "centiles")
cent.cvf.m <- centiles.pred(m_cvf_203, xname="Talla",
                            xvalues=xTalla,
                            cent=c(5,50,95),type = "centiles")
# Se combinan en un data frame
cent.cvf <- cbind(cent.cvf.f,cent.cvf.m[,2:4])
# Cambio de nombres de las columnas
colnames(cent.cvf) <- c("Talla","f.c5","f.c50","f.c95","m.c5",
                        "m.c50","m.c95")
# Comparación de percentiles de CVF entre niñas y niños
cvfcomp <-ggplot(data=cent.cvf, aes(x=Talla))
cvfcomp + geom_line(aes(y=f.c5,colour="Niñas"),size=1)+
  geom_line(aes(y=f.c50,colour="Niñas"),size=1)+
  geom_line(aes(y=f.c95,colour="Niñas"),size=1)+
  geom_line(aes(y=m.c5,colour="Niños"),size=1)+
  geom_line(aes(y=m.c50,colour="Niños"),size=1)+
  geom_line(aes(y=m.c95,colour="Niños"),size=1)+
  ylab("CVF (L)")+xlab("Talla (cm)")+
  scale_colour_manual(name="Curvas percentilares: p5, p50, p95",
                      breaks = c("Niñas",
                                 "Niños"),
                      values = c("Niñas"=paleta[4],
```

```
                                    "Niños"=paleta[2]))+
  theme(legend.position = "top")
# Se repite el mismo procedimiento para la variable FEV1
```