



UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

**Elaboración De Patrones Espirométricos en Niños
Uruguayos Mediante Modelos GAM Y GAMLSS:
Parte 2-Modelización de CVF y FEV por talla edad
y sexo.**

Ramón Álvarez-Vaz; Pablo Palamarchuck; Eugenia Riaño

Diciembre, 2017

Serie Documentos de Trabajo

DT (17/2) - ISSN : 1688-6453

Forma de citación sugerida para este documento:

Álvarez-Vaz, Ramón; Palamarchuck, Pablo; Riaño, Eugenia (2017).
Elaboración De Patrones Espirométricos en Niños Uruguayos Mediante Modelos GAM Y GAMLSS:
Parte 2-Modelización de CVF y FEV por talla edad y sexo. [en línea].
Serie Documentos de Trabajo, DT (17/2). Instituto de Estadística, Facultad
de Ciencias Económicas y de Administración, Universidad de la República,
Uruguay.

Elaboración De Patrones Espirométricos en Niños Uruguayos Mediante Modelos GAM Y GAMLSS: Parte 2-Modelización de CVF y FEV por talla edad y sexo.

Ramón Álvarez-Vaz ¹; Pablo Palamarchuk²; Eugenia Riaño ³

Departamento de Métodos Cuantitativos, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República

RESUMEN

En este trabajo se presenta la segunda parte de un estudio donde se construyen modelos para obtener curvas de referencia espirométricas en niños uruguayos, utilizando Modelos Aditivos Generalizados de Localización, Escala y Forma, para comparar los resultados con otros estudios internacionales. En la primera parte presentada en las jornadas académicas (JJAA2016) de acultad de Ciencias Económicas y de Administración, se identificaron las distribuciones de las variables que componen las variables de espirometría, trabajando con los datos disponibles hasta ese momento, identificando cuatro familias de distribuciones paramétricas como posibles alternativas para la modelización de las variables de respuesta. La espirometría refiere a un conjunto de variables que da cuenta de la capacidad pulmonar la cual varía de acuerdo al tamaño de los pulmones, teniendo una relación directa con la estatura. Pero también puede variar de acuerdo a la etnia y el sexo. 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 de escuelas públicas y privadas del Uruguay por un grupo de investigadores del Centro Hospitalario Pereira Rossell. Los resultados obtenidos en esta segunda parte del trabajo se comparan con otros estudios internacionales, señalando similitudes y diferencias, tanto en metodología como en diseño muestral. Se presentan las tablas percentilares, que pasan a ser valores de referencia a nivel nacional, para las variables espirométricas que surgen de los modelos estimados, que dependen esencialmente de la talla para niños y niñas

Palabras clave: Ajuste de distribuciones, Espirometría, Modelos GAMLSS, Percentiles, Remuestreo.

CÓDIGOS JEL: C31, C51

Clasificación MSC2010: 62P10 ,62J02

¹ *email:* ramon@iesta.edu.uy, ORCID: 0000-0002-2505-4238

² *email:* ppalamarchuk87@gmail.com

³ *email:* eugenia@iesta.edu.uy, ORCID: 0000-0003-3123-2165

Ramón Álvarez-Vaz¹; Pablo Palamarchuk²; Eugenia Riaño³

Departamento de Métodos Cuantitativos, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República

ABSTRACT

In this paper we present the second part of a study where we constructed models to obtain spirometric reference curves in Uruguayan children, using Generalized Additive Models for Location, Scale and Form, to compare the results with other international studies. The results for the first part of the study were presented in a workshop at the Facultad de Ciencias Económicas y de Administración, in October 2016 (JJAA2016). The distributions of the variables that compose the spirometry variables were identified, working with the data available at this moment, identifying four families of parametric distributions as possible alternatives for modeling the output variables. Spirometry refers to a set of variables that accounts for lung capacity which varies according to the size of the lungs, having a direct relationship with height. But it can also vary according to ethnicity and sex. For this reason it is necessary to estimate correct models for Uruguayan children to be able to make a comparison within the same environmental, climatological and geographical conditions. The data used for this purpose comes from a sample of public and private schools in Uruguay worked by a group of researchers from the Pereira Rossell Hospital Center. The results obtained in this second part of the work were compared with other international studies, pointing out similarities and differences, both in methodology and in sample design. The percentile tables presented, become reference values at national level, for the spirometric variables that arise from the estimated models, those that depend essentially on the size boys and girls.

Key words: Adjustment of distributions, GAMLSS models, Percentiles, Resampling, Spirometry.

JEL CODES: C31, C51

Mathematics Subject Classification MSC2010: 62P10 ,62J02

¹*email:* ramon@iesta.edu.uy, ORCID: 0000-0002-2505-4238

²*email:* ppalamarchuk87@gmail.com

³*email:* eugenia@iesta.edu.uy, ORCID: 0000-0003-3451-8249

1. Introducción

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

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 (FEV₁) y los flujos forzados denominados periféricos (FEF₂₅, FEF₅₀, FEF₇₅ y FEF₂₅₋₇₅), que corresponden a porciones de la curva Flujo/Volumen y representan los flujos de la vía aérea más pequeña.

Además se mide el Índice de Gaënsler - el cual se determina con la espiración forzada expresada como un ratio FEV₁/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) .

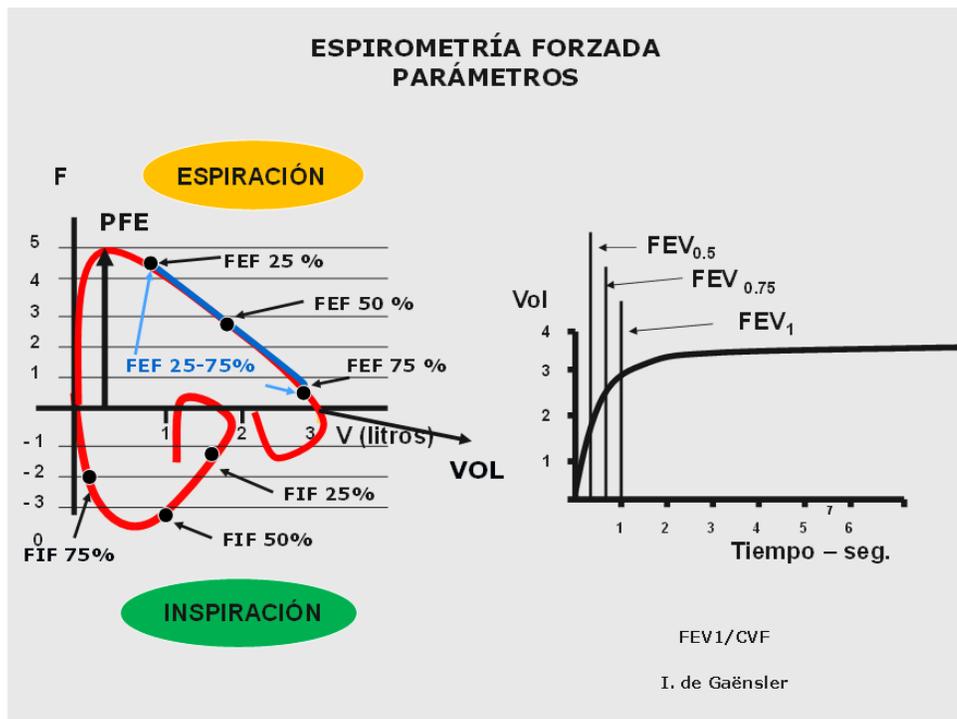


Figura 1: Espirometría Forzada -Parámetros de curva flujo/volumen y volumen/tiempo

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 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).

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.

El objetivo principal del presente trabajo consiste en encontrar curvas de referencia de parámetros espirométricos 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

países con características ambientales y climáticas distintas. Este trabajo se estructura en 3 partes. En la primera se introduce brevemente la el problema que ya había sido desarrollado en el documento de trabajo (Álvarez-Vaz *et al.*, 2016). En la sección 2 se resume muy brevemente la metodología que ya había sido adelantado en (Álvarez-Vaz *et al.*, 2016). Luego, en la sección 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 ajuste de distribuciones a las variables espirométricas, para culminar con la modelización de las mismas. En la última sección se presentan las principales conclusiones y consideraciones a futuro.

2. Metodología

El análisis de regresión es una de las técnicas estadísticas 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. Los usuarios de los modelos de regresión lineal estándar, pronto encuentran que los supuestos clásicos sobre normalidad, varianza constante de los errores y linealidad de la relación entre la variable de respuesta y las explicativas, 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.

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 (Akanztliotou *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).

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.

2.1. 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 modelo 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-paramétrico” 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 remplazado por una distribución general, incluyendo distribuciones continuas y discretas con posible asimetría y/o kurtosis. 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).

2.1.1. El modelo GAMLSS

Un modelo GAMLSS asume que, para $i = 1, 2, \dots, n$, observaciones independientes de la variable de respuesta Y_i , ésta tiene función de densidad $f_Y(y_i|\theta^i)$ 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 $i = 1, 2, \dots, n$, 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 se presentan, son

caracterizados como parámetros de forma (asimetría y kurtosis).

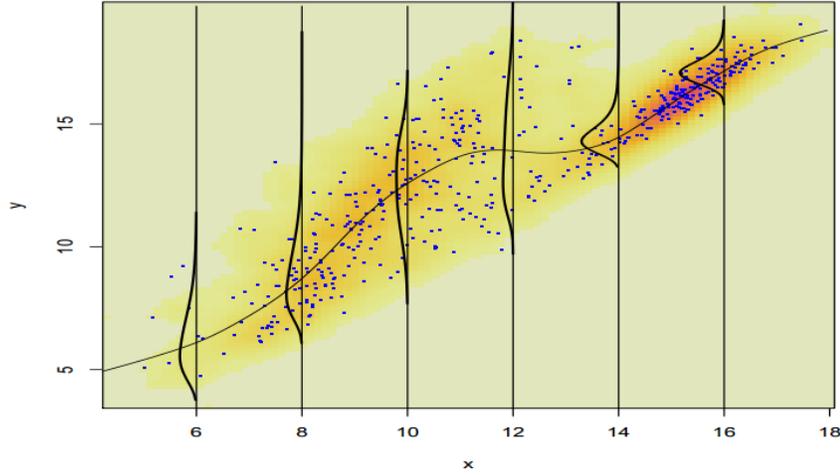


Figura 2: Supuestos del modelo de regresión GAMLSS (Fuente: *A flexible regression approach using GAMLSS in R*, Rigby y Stasinopoulos, 2010)

Sea $\mathbf{Y}^T = (Y_1, Y_2, \dots, Y_n)$ 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 $g_k(\cdot)$ una función de enlace monótona conocida que relaciona el parámetro de distribución θ_k al predictor η_k .

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

llevado al caso

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

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

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

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

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 , $\boldsymbol{\beta}_k^T = (\beta_{1k}, \beta_{2k}, \dots, \beta_{J_k k})$ 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 $N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-1})$, donde \mathbf{G}_{jk}^{-1} es la matriz inversa (generalizada) de una matriz simétrica de $q_{jk} \times q_{jk}$ $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$, 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 $\exp(-\frac{1}{2}\gamma_{jk}^T \mathbf{G}_{jk} \gamma_{jk})$, 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}_{jk}^{-1} .

El modelo (1) 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 (1), entonces tenemos la formulación *aditiva semi-paramétrica* de GAMLSS dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \quad (2)$$

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*

$$g_1(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k \quad (3)$$

El modelo (2) 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:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \quad (4)$$

Nos vamos a referir al modelo (4) 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, (4) se reduce a un modelo GAMLSS *paramétrico no-lineal*:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k). \quad (5)$$

Si, adicionalmente, $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) = \mathbf{X}_k^T \boldsymbol{\beta}_k$ para $i = 1, 2, \dots, n$ y $k = 1, 2, 3, 4$, entonces el modelo (5) se reduce a un modelo paramétrico lineal (3). 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 (3) o (5) como modelos GAMLSS paramétricos.

Los vectores de parámetros β_k y los parámetros aleatorios γ_{jk} , para $j = 1, 2, \dots, J_k$ 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

$$\ell_p(\beta, \gamma) = \ell(\beta, \gamma) - \frac{1}{2} \sum_{k=1}^p \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^T \mathbf{G}_{jk} \gamma_{jk} \quad (6)$$

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 β'_k s y los γ'_{jk} s, para $j = 1, 2, \dots, J_k$ y $k = 1, 2, 3, 4$. Para modelos GAMLSS paramétricos (3) o (5), $\ell_p(\beta, \gamma)$ se reduce a $\ell(\beta)$, y los β_k para $k = 1, 2, 3, 4$, son estimados maximizando la función de verosimilitud $\ell(\beta)$.

2.1.2. 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()` (Stasinopoulos y Rigby, 2007) 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.

2.2. Criterios de selección del modelo

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 los 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.

2.2.1. 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 el criterio de información de Akaike generalizado (GAIC) ver ecuación (7). El modelo con el menor valor de $GAIC(k)$, para un valor de k elegido, es seleccionado.

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

$$GAIC(k) = -2 \sum_{i=1}^n \log \left[f(y_i | \hat{\theta}_i) \right] + k \cdot df \quad (7)$$

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.

2.2.2. 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 logarítmico 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 las que 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 *deviance* global (GD). La mejor función de enlace resulta en la GD más baja.

2.2.3. 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 θ_i para $i = 1, 2, 3, 4$, donde $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) = (\mu, \sigma, \nu, \tau)$. 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 continuas. 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. En el ejemplo anterior la interacción de los factores f_1 y f_2 afecta 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 afecta 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 mismos. Por ejemplo, si el número de variables explicativas continuas 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 continuas, como 50, hay $2^{50} = 1,13 \times 10^{15}$ combinaciones diferentes que no todas pueden ser ajustadas, por lo que se debe implementar otra estrategia.

Hay muchas funciones dentro de la librería **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).

2.2.4. 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 hiperparámetro) 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 ser fijados o estimarse a partir de 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.

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 $i = 1, 2, \dots, n$. 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 hessiana \mathbf{H} . El problema con los *residuos crudos* 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 desviación* $r_i^d = \text{sign}(y_i - \hat{\mu}_i) / \sqrt{d_i}$ donde $d_i = -2 \log(L_i^c / L_i^s)$ o los *residuos de Pearson* $r_i^P = (y_i - \hat{\mu}_i) / \text{se}(\hat{\mu}_i)$ a menudo son utilizados. Desafortunadamente, los residuos de desviación 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).

Hay varias funciones implementadas que usan los cuantiles normalizados (aleatorizados) 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 `Q.stats()` de los estadísticos Q (Royston y Wright, 2000), para detectar si los residuos son “significativamente” diferentes a una distribución normal en su media, varianza, asimetría 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 continua)

3. Aplicación

Con respecto al documento de trabajo (Álvarez-Vaz *et al.*, 2016) luego de la depuración de los datos en la figura 3 se ve como quedan los datos en las diferentes etapas

Ante la imposibilidad de realizar un estudio aleatorizado de todas las escuelas públicas y privadas del país, se seleccionó una muestra por conveniencia, incluyendo zonas en donde existe ascendencia indígena (Tacuarembó) y de distintos niveles de contaminación ambiental.

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 primer 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.

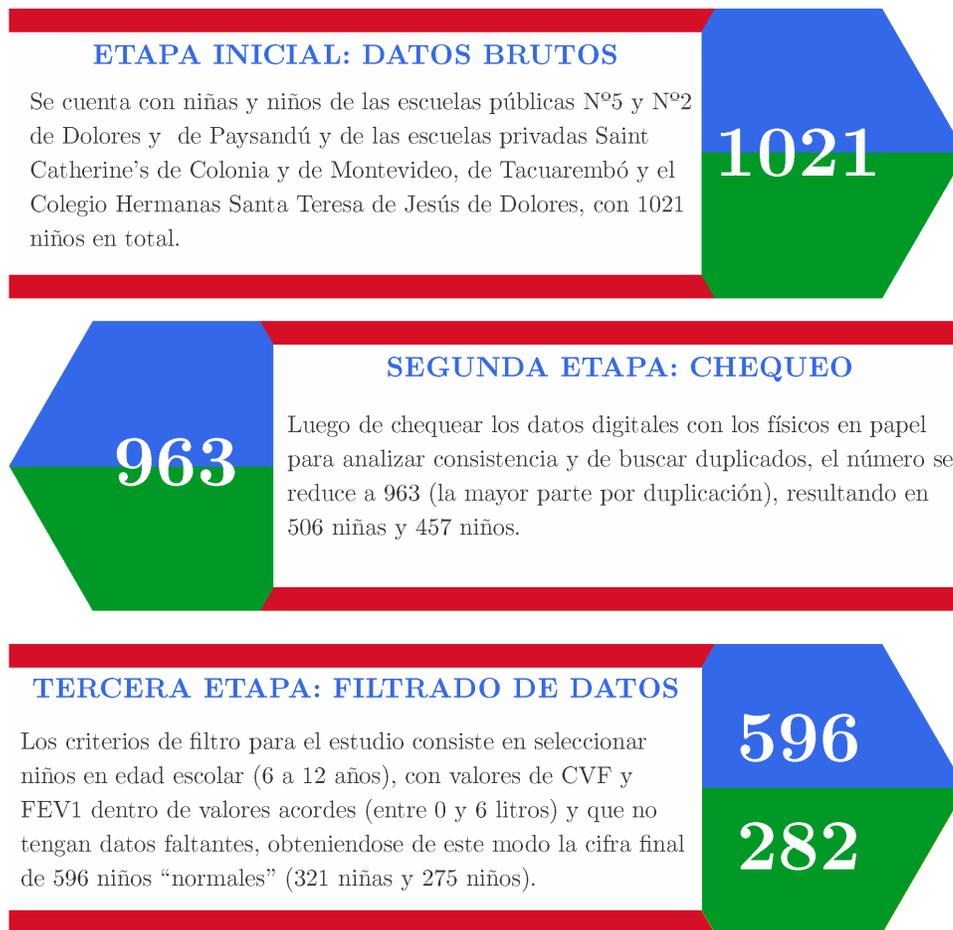


Figura 3: Etapas de depuración del conjunto de datos, donde se muestra la cantidad de observaciones implicadas y la descripción de las mismas

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 (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).

Los equipos médicos estaban constituidos 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 $\pm 0,1\text{kg}$) y se midió su talla (estatura) descalzos mediante un pediómetro digital Sohenle 5001 (error $\pm 0,5\text{cm}$) 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 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.

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).

4. Variables dentro del estudio

Las variables comprendidas en este estudio se listan a continuación en la Tabla 2.

A continuación se presentan las medidas de resumen para las variables del estudio, tanto para las variables de respuesta como para las variables explicativas. En la Tabla 3 se muestra el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, el máximo, la media y la varianza de las variables antropométricas continuas. En la Tabla 4 se muestra la distribución de frecuencias de las variables categóricas. En la Tabla 5 se muestran los mismos valores que en la Tabla 3 para las variables de respuesta de los parámetros de espirométricos.

Tabla 2: Descripción de las variables espirométricas del estudio.

Variable	Tipo de variable	Descripción
Edad	Continúa	Edad del niño expresada en años al momento del estudio.
Talla	Continúa	Talla del niño al momento de estudio. Expresada en centímetros.
Peso	Continúa	El peso expresado en kilogramos.
Sexo	Categoría Nominal	El sexo del niño, con valores F para femenino y M para masculino.
Alérgicos	Categoría Nominal	Variable que refiere a antecedentes patológicos.
ContFab	Categoría Nominal	Presencia de contaminación ambiental por actividades industriales .
Fuman	Categoría Nominal	Si en la casa hay alguien que fuma, ya sea madre, padre, abuelos u otros.
Escuela	Categoría Nominal	Escuela a la cual pertenece el niño. Puede considerarse también como una variable geográfica.
CVF	Continúa	Capacidad Vital Forzada, expresada en litros (L).
FEV1	Continúa	Volumen Espiratorio Forzado en el primer segundo (FEV_1), expresado en litros (L).
FEF2575	Continúa	Flujo Espiratorio Forzado medido en la mitad de la espiración (FEF_{25-75}) ó mesoflujo.
PFE	Continúa	Pico de Flujo Espirométrico. Se mide en litros por minuto (L/min).
DT (17/3)-Instituto de Estadística Gaensler	Continúa	Álvarez-Vaz, R.; Palamarchuk, P.; Riaño, E. La relación FEV_1/CVF , también conocido como Índice de Gaënsler.

Tabla 3: Medidas de localización y dispersión para las variables antropométricas continuas

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 4: Frecuencias absolutas de las variables categóricas.

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

Tabla 5: Medidas de resumen para las variables espirométricas

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

4.1. Relaciones entre variables antropométricas

Como primera observación (ver Figura 4), se puede decir que la relación entre la Talla y el Peso no es necesariamente lineal. También parece no haber simetría respecto de la línea azul, que representa la media local, es decir, es decir que podría existir homocedasticidad.

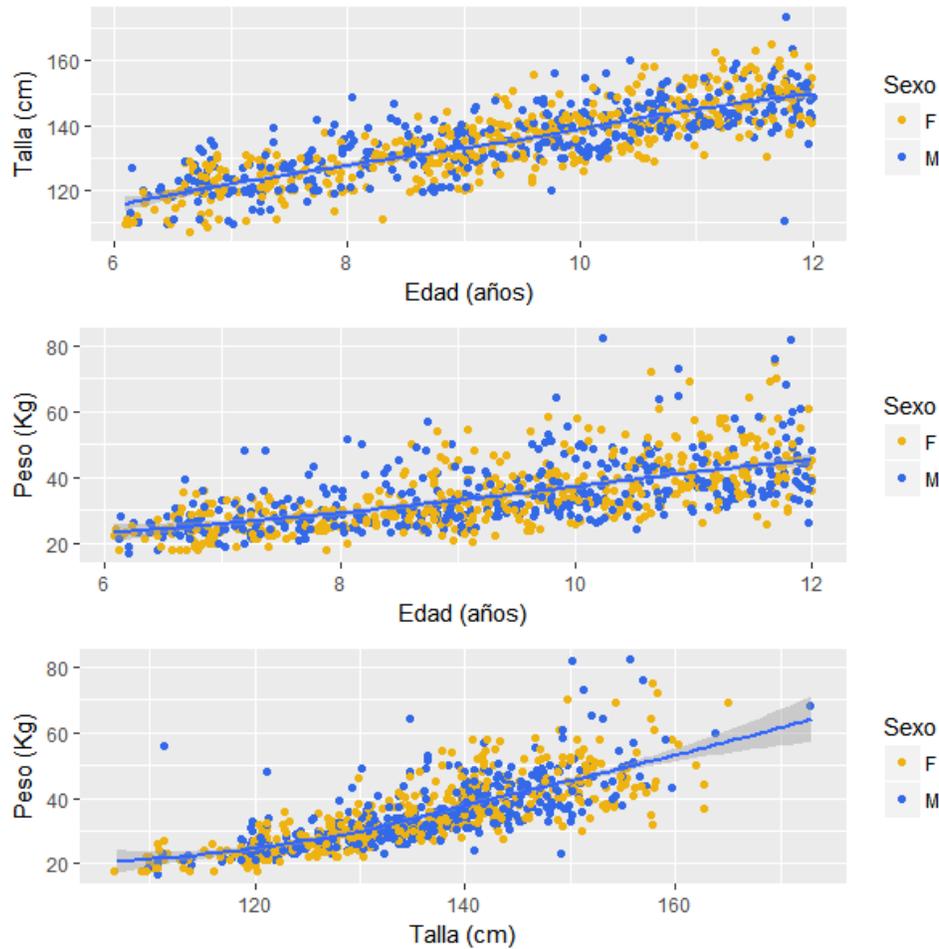


Figura 4: Gráfico de dispersión entre las variables Edad, Talla y Peso, coloreado por Sexo, donde F refiere a femenino y M a masculino: (arriba) Edad y Talla; (centro) Edad y Peso; (abajo) Talla y Peso.

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 considerarse lineal, con una menor dispersión que en los gráficos anteriores, aunque igual con cierta heterocedasticidad respecto a la media local.

4.1.1. Relación entre Volumen Espiratorio Forzado en el primer segundo (FEV₁) y las variables antropométricas.

Análogamente, se presentan los gráficos que relacionan la variable (FEV₁) con las variables explicativas Talla, Edad y Peso.

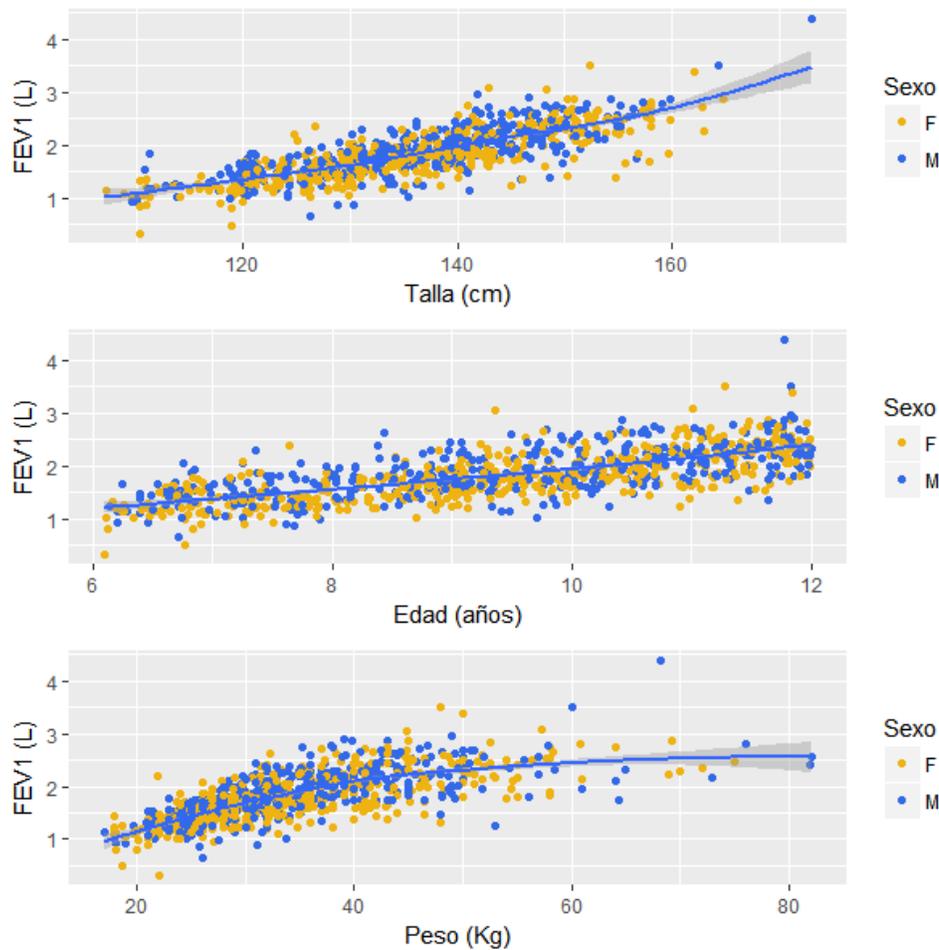


Figura 5: FEV₁ en relación a: (arriba) Talla; (centro) Edad y (abajo) Peso.

El gráfico central de la Figura 5 da muestras de una relación lineal entre el FEV₁ 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 presenta una concavidad negativa para la variable Peso.

5. 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 6 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 observa con detenimiento en la Figura 6, 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.

Tabla 6: 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

La Figura 7 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 solo diferencias en la forma de la densidad, sino la existencia de un desplazamiento en media, donde los niños alérgicos presentan valores menores que los niños normales.

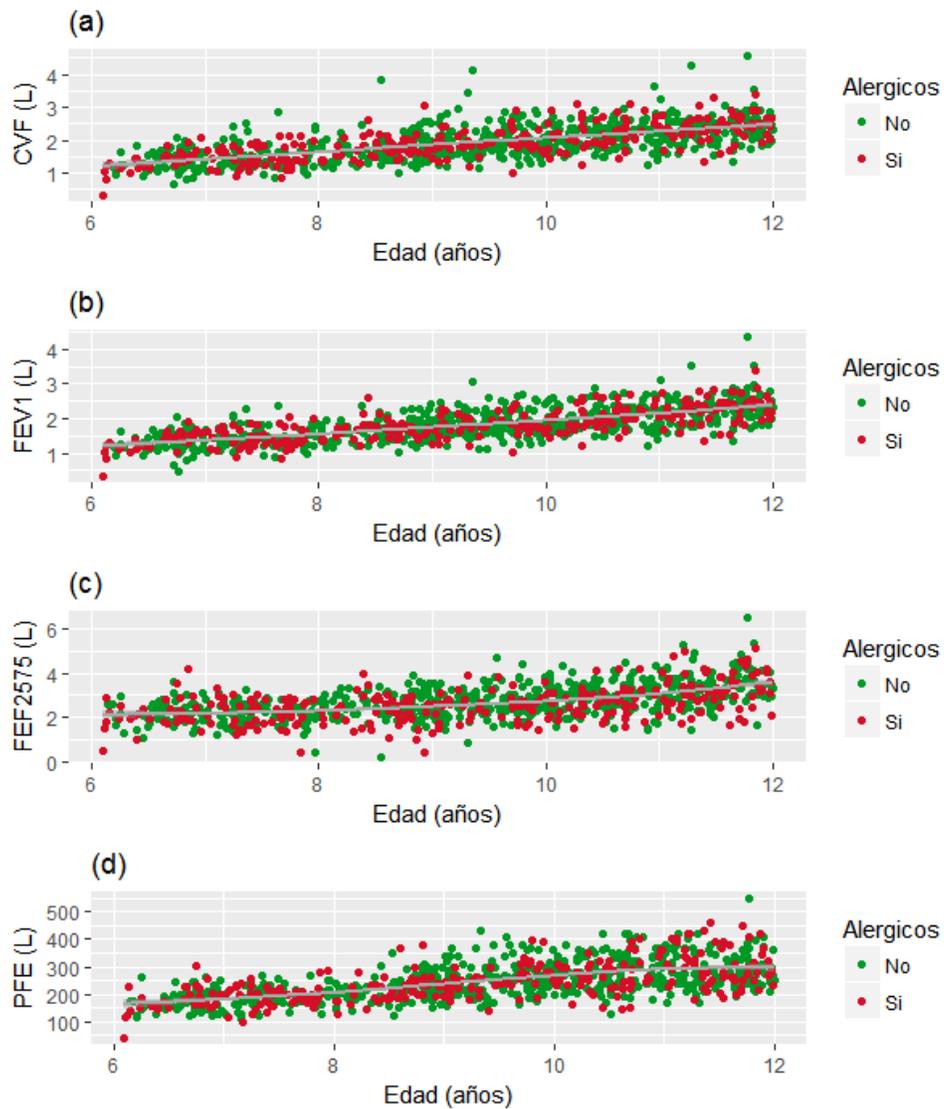


Figura 6: Gráfico de (a) CVF, (b) FEV₁, (c) FEF₂₅₋₇₅ y (d) PFE en función de la edad, donde se distinguen entre niños alérgicos (rojos) y normales (verdes).

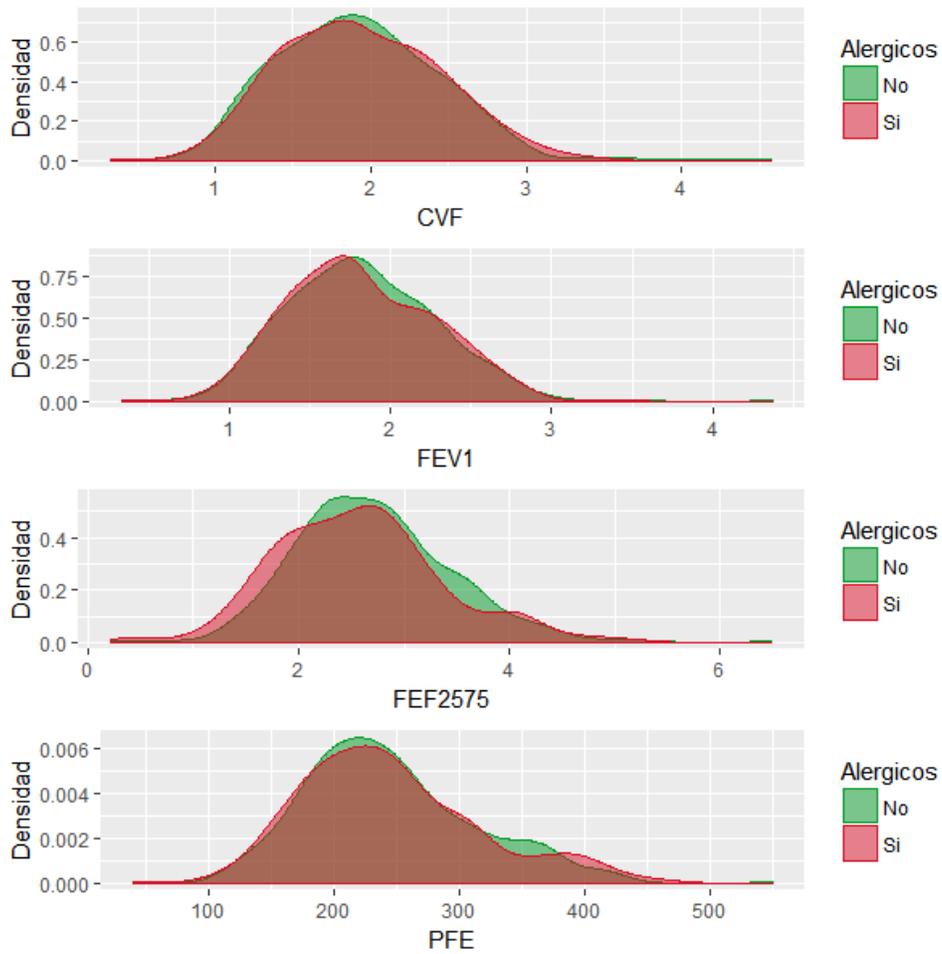


Figura 7: Gráfico comparativo de densidades para los parámetros CVF, FEV₁, FEF₂₅₋₇₅ y PFE de los niños alérgicos (rojo) y normales (verde).

Tabla 7: Matriz de varianzas y covarianzas de los parámetros espirométricos CVF, FEV₁, FEF₂₅₇₅ y PFE para el grupo de los niños normales (arriba) y los niños alérgicos (abajo).

Grupo	Variable	CVF	FEV ₁	FEF ₂₅₇₅	PFE
normales	CVF	0.28	0.23	0.18	18.39
	FEV ₁		0.22	0.22	17.29
	FEF ₂₅₇₅			0.55	23.33
	PFE				4590.50
alérgicos	CVF	0.26	0.22	0.19	22.06
	FEV ₁		0.21	0.25	20.87
	FEF ₂₅₇₅			0.63	27.72
	PFE				5144.64

Tabla 8: Resultado de la prueba T^2 de Hotelling para las variables CVF, FEV₁, FEF₂₅₋₇₅ y PFE entre niños con antecedentes patología respiratoria (alérgicos) y sin antecedentes de patología (normales).

Prueba de Hotelling T^2

Variables: CVF, FEV1, FEF2575, PFE

Estadístico de prueba	2.48
Gr. Libertad Numerador	4
Gr. Libertad Denimonador	873
P-valor	0.042

El p-valor de la prueba ($p\text{-valor} < 0.05$) en la Tabla 8 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 continua utilizando solo a los niños normales, al igual que otros autores, para así también poder comparar con los distintos resultados de estudios internacionales.

6. Estado nutricional

A continuación se presentan los valores de Z-score de Talla e Índice 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 (?), 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.

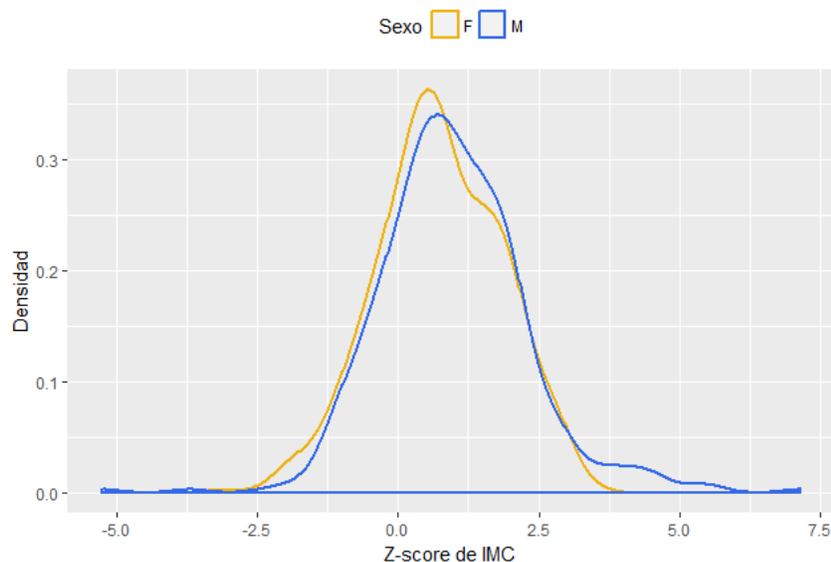


Figura 8: Densidades de los Z-score del IMC por edad de niñas (naranja) y niños(azul).

Se puede observar en la Figura 8 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 9.

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

Tabla 9: IMC por Edad

En cuanto a la Talla, en la Tabla ?? 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.

En resumen, con la información obtenida con el programa Anthro, se puede decir que los niños en términos de la Talla, son comparables a nivel internacional, mientras que el Peso, a través del IMC, presentan una tendencia al sobrepeso.

Si bien hasta el momento se trabajó con todas las variables componentes de la espirometría en este trabajo solo se presentan los resultados que corresponden a FEV_1

7. Distribución de FEV_1

La meta es encontrar familias de distribuciones para las variables de respuesta 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. Las diferencias entre ambas ya se habían presentado en (Álvarez-Vaz *et al.*, 2016) 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. Se establece un número de iteraciones en primera instancia (`iter`).
2. Luego se fija un valor de tamaño de muestra del total de datos (`nS`).
3. Se crea un *data frame* para guardar los resultados de la rutina, donde se registra 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`.
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`.

7.1. Resultados para FEV₁

Primero se muestra los resultados correspondientes al conjunto de todos los niños normales. La Tabla 10 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.

En la Figura 9 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 asimetría hacia la izquierda.

7.1.1. Resultados para niñas

A continuación, en la Tabla 11 se muestran los resultados obtenidos del proceso iterativo para el conjunto de datos de niños normales de sexo femenino de la variable FEV₁.

Tabla 10: Resultado de la prueba de robustez para la variable FEV_1 de niños normales.

Familia	k	Frec.Rel	GAIC medio	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$\hat{\tau}$
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
SHASH _o	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
SHASH _{o2}	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 11: Resultado de la prueba de robustez para la variable FEV_1 de niños normales de sexo femenino.

Familia	k	Frec.Rel	GAIC medio	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$\hat{\tau}$
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
SHASH _{o2}	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
SHASH _o	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

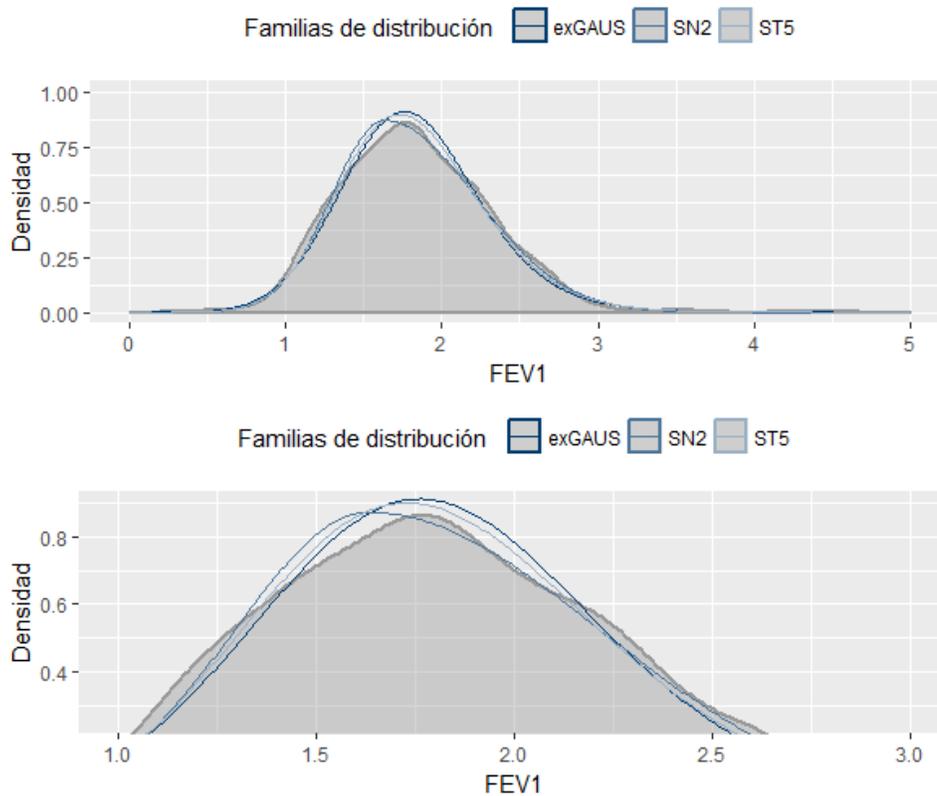


Figura 9: Densidades ajustadas para FEV_1 : *exponential Gaussian* (exGAUS); *skewed Normal type 2* (SN2); *skewed t type 5* (ST5). El gráfico inferior es un zoom de la zona central.

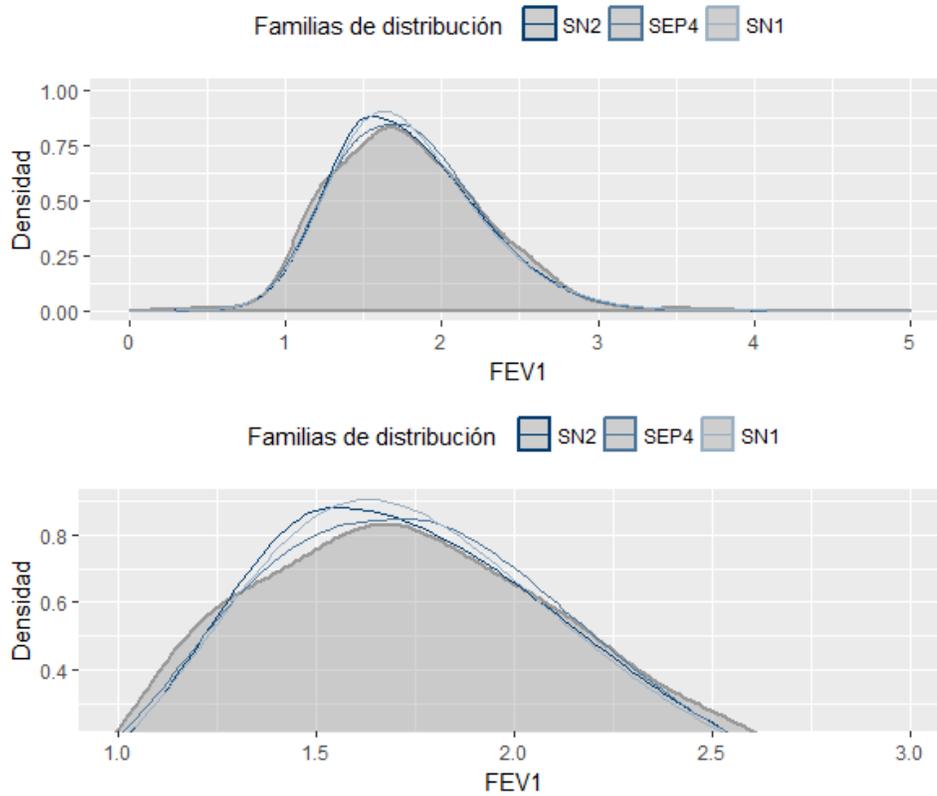


Figura 10: Densidades ajustadas para FEV_1 de niños de sexo femenino: *skewed Normal type 2* (SN2); *skewed power Exponential type 4* (SEP4); *skewed Normal type 1* (SN1). El gráfico inferior es un ampliación de la zona central.

En la Figura 10 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 adecuado.

7.1.2. Resultados para niños

A continuación, en la Tabla 12 se muestran los resultados obtenidos del proceso iterativo para el conjunto de datos de niños normales de sexo masculino de la variable FEV₁.

Tabla 12: Resultado de la prueba de robustez para la variable FEV₁ de niños normales de sexo masculino.

Familia	k	Frec.Rel	GAIC medio	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$\hat{\tau}$
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

Se observa en la Figura 11 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 desplazamiento hacia la derecha y la familia de distribución SN2 presenta un corrimiento de la media a la izquierda.

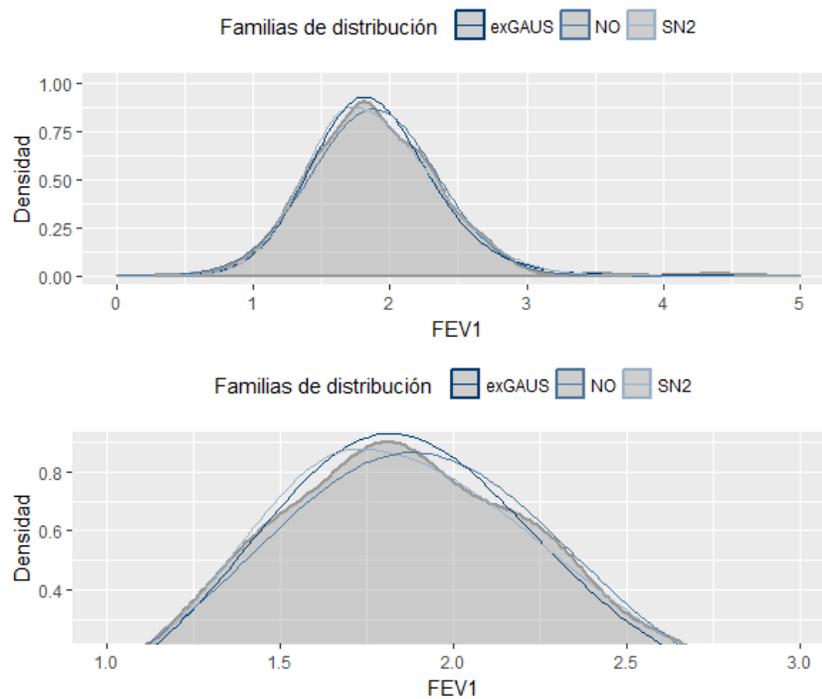


Figura 11: Densidades ajustadas para FEV_1 de niños de sexo masculino: *exponential Gaussian* (exGAUS); *Normal* (NO); *skewed Normal type 2* (SN2). El gráfico inferior es un ampliación de la zona central.

8. Modelización para FEV_1

La estrategia que se adoptó para la modelización de las variables espirométricas de las cuales se da cuenta los resultados para 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), que es utilizada por otros autores para éste fin.

Los modelos, 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)` (spline penalizado) 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:

$$\mu = a_\mu + pb(Talla, edf_{\mu T}) + pb(Edad, edf_{\mu E}) \quad (8)$$

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 para **Talla** y **Edad** respectivamente. El término a_μ es una constante.

Luego, según la situación, se incluirá la variable **Sexo** como factor regresor.

Variabilidad σ

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

$$\log\sigma = a_\sigma + pb(Talla, edf_{\sigma T}) + pb(Edad, edf_{\sigma E}) \quad (9)$$

El término a_σ es una constante

Asimetría ν y curtosis τ

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

$$\nu = a_\nu + pb(Talla, edf_{\nu T}) + pb(Edad, edf_{\nu E}) \quad (10)$$

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

$$\log\tau = a_\tau + pb(Talla, edf_{\tau T}) + pb(Edad, edf_{\tau E}) \quad (11)$$

donde los términos a_ν y a_τ son constantes.

Con la idea de comparar los modelos, se decide por seguir la siguiente estructura:

- **Modelo GAMLSS con variable Sexo:** La idea es hacer un modelo GAMLSS para las variable de FEV₁, 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 FEV₁ para los niños y otro para las niñas.

8.1. Descripción de los diferentes modelos estimados

Antes de dar paso a la presentación de la modelización de FEV₁ en sus distintos escenarios, se presenta la nomenclatura de los modelos dentro de los mismos.

Los distintos modelos siguen una estructura de la forma *m_variable_abc*.

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

variable hace referencia a cual de las variables de respuesta es modelizada;

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.

8.2. Modelos GAMLSS global para FEV₁

Tabla 13: Desarrollo de los modelos GAMLSS para FEV₁, 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.

(r)3-7 Modelo	Distribución	Talla	Predictor lineal para μ			Sexo	df	<i>deviance</i>	SBC
			$edf_{\mu T}$	Edad	$edf_{\mu E}$				
<i>m_fev_001</i>	NO	pb()	7.47	-	-	-	8.47	152.48	204.74
<i>m_fev_002</i>	exGAUS	pb()	7.50	-	-	-	9.5	152.03	210.63
<i>m_fev_003</i>	BCPE	pb()	3.32	-	-	-	6.3	116.71	155.70
<i>m_fev_013</i>	BCPE	pb()	3.72	-	-	Si	7.72	95.01	142.62
<i>m_fev_023</i>	BCPE	pb()	3.83	pb()	2.01	Si	8.85	86.73	141.38

El modelo final *m_fev_023*, resultante de la función `stepGAICAll.A()`, con familia de distribución BCPE, es de la forma:

$$\mu = a_{\mu} + pb(\text{Talla}, 3,83) + pb(\text{Edad}, 2,01) + \text{Sexo}; \quad \log(\sigma) = a_{\sigma}; \quad \nu = a_{\nu}; \quad \log(\tau) = a_{\tau} \quad (12)$$

con un SBC=141.38 y 8.85 grados de libertad.

En este caso sólo un parámetro de la distribución fue modelizado, 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 FEV₁ 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.

En la Figura 12 se puede ver como la variable Talla tiene un aporte mayor en el valor del FEV₁ que la variable Edad, que es casi nulo.

Tabla 14: Coeficientes de la regresión lineal del modelo m_fev_023

	Estimación	Error Std.	t	p
<i>Parámetro de localización</i>				
función enlace μ : identidad				
coeficientes μ				
Constante	-2.327	0.036	-64.742	< 0.001
pb(Talla)	0.028	6.035e-04	46.378	< 0.001
SexoM	0.107	0.022	4.791	< 0.01
pb(Edad)	0.034	0.010	3.753	< 0.01
<i>Parámetro de escala</i>				
función enlace σ : log				
coeficientes σ				
Constante	-1.904	0.045	-42.56	< 0.001
<i>Parámetro de asimetría</i>				
función enlace ν : identidad				
coeficientes ν				
Constante	1.234	0.312	3.957	< 0.001
<i>Parámetro de curtosis</i>				
función enlace τ : log				
coeficientes τ				
Constante	0.360	0.093	3.867	< 0.001

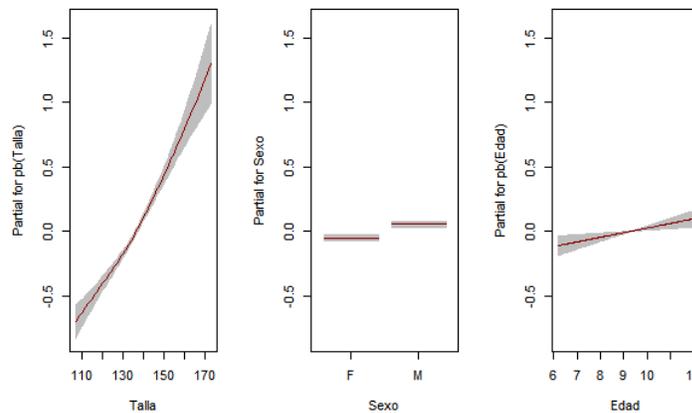


Figura 12: Efecto de las variables Talla, Edad y Sexo sobre el predictor lineal de μ .

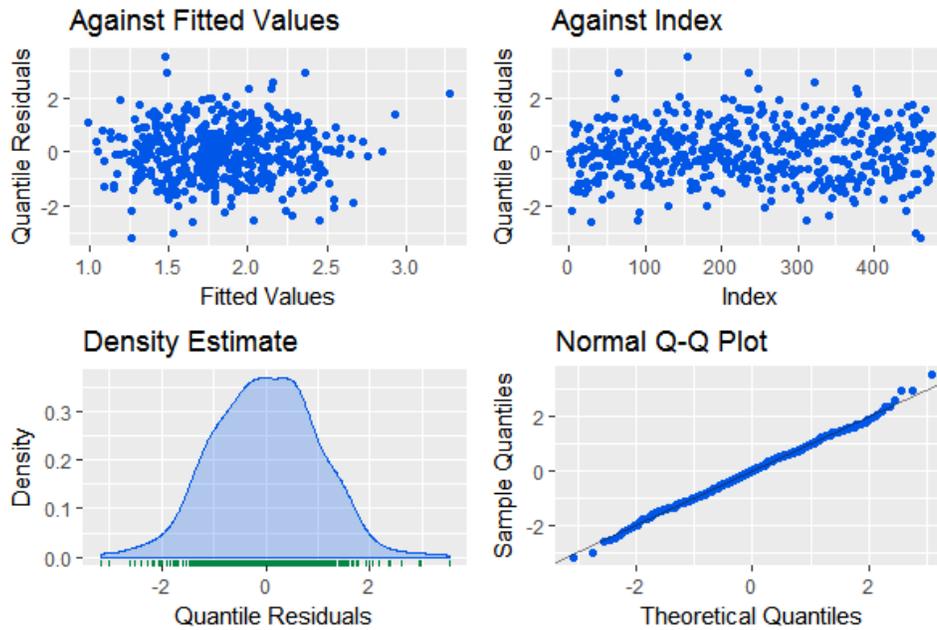


Figura 13: Gráfico de los residuos del modelo m_fev_023.

Tabla 15: Cuantiles residuales de los errores para el modelo m_fev_023.

Resumen de los Cuantiles Residuales	
media	-0.0017
varianza	1.0004
coef. de asimetría	0.013
coef. de curtosis	3.19
coef. de correlación de Filliben	0.998

En la Figura 13 se puede observar que la densidad estimada de los residuos del modelo `m_fev_023` se asemejan a una distribución normal. Además, los cuantiles residuales tienen un coeficiente de correlación de Filliben de 0.998 (Tabla 15), lo que es indicador de alta normalidad (la normalidad tiene un valor de 1). La Figura 14 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.

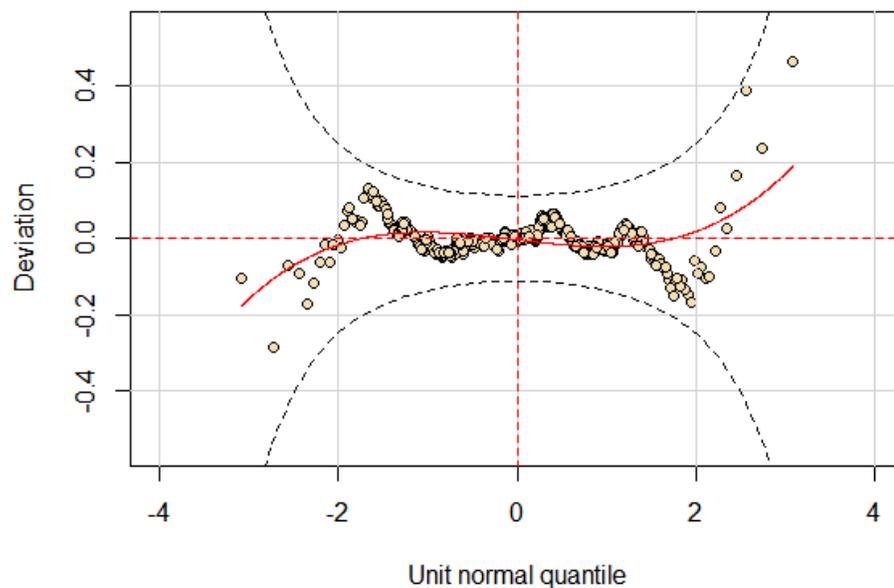


Figura 14: Worm plot del modelo `m_fev_023`.

8.3. Modelos GAMLSS por sexo para FEV_1

Para los modelos estimados por sexo (es decir trabajando con datos separadas por sexo) solo se presentan los modelos finales.

8.3.1. Niñas

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

$$\mu = a_\mu + pb(\text{Talla}, 3,6524) \log(\sigma) = a_\sigma \quad \nu = a_\nu \quad \log(\tau) = a_\tau \quad (13)$$

con un SBC=83.21 y 6.65 grados de libertad.

8.3.2. Niños

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

$$\mu = a_\mu + pb(\text{Talla}, 3,66) \quad \log(\sigma) = a_\sigma \quad \nu = a_\nu \quad \log(\tau) = a_\tau \quad (14)$$

con un SBC=72.02 y 6.66 grados de libertad.

Tabla 16: Ecuaciones de regresión para FEV₁

Parámetro espirométrico	Sexo	Ecuación de regresión	SBC
FEV ₁ (L)	Global	$\mu = -2,327 + pb(\text{Talla}, 3,83) + pb(\text{Edad}, 2,01) + 0,107 \times \text{Sexo}$ $\log(\sigma) = -1,904$ $\nu = 1,234$ $\log(\tau) = 0,360$	141,38
	Niñas	$\mu = -2,572 + pb(\text{Talla}, 3,65)$ $\log(\sigma) = -1,868$ $\nu = 0,799$ $\log(\tau) = 0,355$	83,21
	Niños	$\mu = -2,775 + pb(\text{Talla}, 3,66)$ $\log(\sigma) = -1,95$ $\nu = 1,230$ $\log(\tau) = 0,573$	72,02

Nota: todos los modelos presentados tienen una distribución BCPE.

9. Conclusiones

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 T² multivariada con los distintos parámetros espirométricos CVF, FEV₁, FEF₂₅₋₇₅ 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 FEF₂₅₋₇₅ (lo cual es un hallazgo novedoso y sorprendente) 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.

La estructura de la modelización del parámetro 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 idea 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 partieron de modelos que incluyeran términos de suavizado en función de la talla, dado que esta variable es la que tiene mayor peso en relación al tipo de unidades 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, músculos, 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 incluido dentro de la construcción de los modelos para poner a prueba si su inclusión era estadísticamente significativa. Cada parámetro de distribución se quiso modelizar con la estructura $g(\theta_k) = a_\theta + pb(Talla) + pb(Edad)$, para cada $k = 1, 2, 3, 4$, donde $g(\cdot)$ es la función de enlace y $pb(\cdot)$ 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 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 de las funciones en la librería, 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, 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, el más cercano es el 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 jóvenes 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 técnica adecuada 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 categórica. Posee varias herramientas para chequear posibles inadecuaciones en los modelos, como los gráficos de gusano y los cuantiles residuales.

Referencias Bibliográficas

- Akantziliotou, K., Rigby, R., y Stasinopoulos, D. (2002). The R implementation of Generalized Additive Models for Location, Scale and Shape.
- Álvarez-Vaz, R., Palamarchuck, P., y Riaño, E. (2016). Elaboración de patrones espirométricos normales en niños uruguayos mediante modelos gam y gamlss: Parte 1-identificación de la distribución de la variable de respuesta. *Documentos de Trabajo -Serie DT IESTA* (16 / 3), *Universidad de la República. Facultad de Ciencias Económicas y de Administración*.
- 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., Stanojevic, S., y Stocks, J. (2008). Age-and size-related reference ranges: A case study of spirometry through childhood and adulthood. *Statistics in Medicine*.
- Dunn, P. K. y Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.
- Hastie, T. y Tibshirani, R. (1986). *Generalized Additive Models*, volumen 1. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.
- 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.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- 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 Wright, E. (2000). Goodness of fit statistics for age-specific reference intervals. *Statistics in Medicine*.
- DT (17/3)-Instituto de Estadística Álvarez-Vaz, R.; Palamarchuck, P.; Riaño, E.

- RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.
- 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):253â260.
- 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 edición. ISBN 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.

A. Tablas normales de referencia

A continuación se presentan las tablas con los valores estimado de los percentiles 5, 50 y 95 para las variable FEV₁.

(r)2-7 Talla	Percentiles niñas			Percentiles niños		
	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 17: Estimación de los percentiles 5, 50 y 95 para la variable FEV₁ de niñas (m_fev_103) y niños (m_fev_203) a través de la variable Talla.

Instituto de Estadística

Documentos de Trabajo



Eduardo Acevedo 1139. CP 11200 Montevideo, Uruguay

Teléfonos y fax: (598) 2410 2564 - 2418 7381

Correo: ddt@iesta.edu.uy

www.iesta.edu.uy

Área Publicaciones

Diciembre, 2017

DT (17/2)