CAPE
CUERPO ACADÉMICO DE PROBABILIDAD Y ESTADÍSTICA BUAP

Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

MODELOS GAMLSS: UNA ALTERNATIVA PARA LA MODELIZACIÓN DE LA ESPIROMETRÍA EN NIÑOS URUGUAYOS

PABLO PALAMARCHUK [a], RAMÓN ÁLVAREZ-VAZ [b]

[a]Licenciado en Estadística, Consultor Independiente
[b]ppalamarchuk87@gmail.com

[b]*Instituto de Estadística,Departamento de Métodos Cuantitativos, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo, Uruguay*
[a]ramon@iesta.edu.uy

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 Econmicas 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

Keywords: Ajuste de distribuciones, Espirometría, Modelos GAMLSS.

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 (FEV_1) y los flujos forzados denominados periféricos (FEF_{25}, FEV_{50}, FEV_{75} y FEF_{25-75}), que corresponden a porciones de la curva Flujo/Volúmen y representan los flujos de la 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) .

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

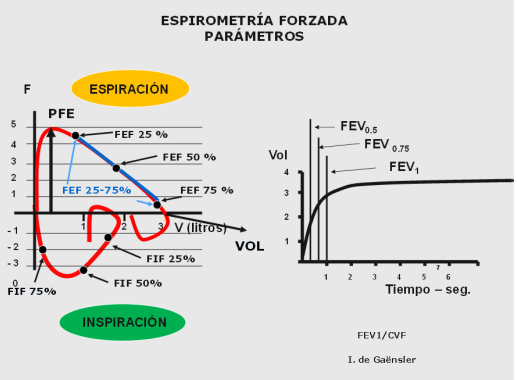

Fig. 1. Espirometría Forzada -Parámetros de curva flujo/volumen y volumen/tiempo

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 paises con características ambientales y climáticas distintas. Este trabajo se estructura en 3 partes. En la primera se introduce brevemente el problema que ya había sido desarrollado en el documento de trabajo [1]. En la sección 2 se resume muy brevemente la metodología que ya había sido adelantado en [2]. Luego, en la sección 3 se aborda la aplicación, donde se hace una descripción de los datos, 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 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. 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 [3] y por [4] 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 [5] [6], [7] y [8] como una forma de superar algunas de las limitaciones

asociadas con los modelos lineales generalizados (GLM) y modelos aditivos generalizados (GAM) [3] y [4], 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 "semiparamétrico" 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 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).

2.1.1. 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 se presentan, son caracterizados como parámetros de forma (asimetría y kurtosis).

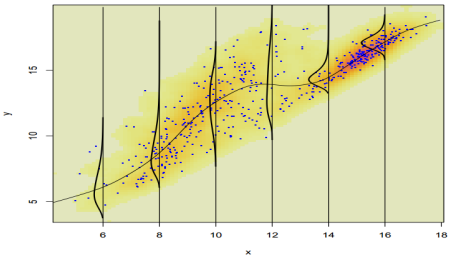

Fig. 2. Supuestos del modelo de regresión GAMLSS (Fuente: *A flexible regression approach using GAMLSS in R, Rigby y Stasinopoulos, 2010*)

Sea $\mathbf{Y}^\top = (Y_1,Y_2,...,Y_n)$ el vector de largo n de la variable de respuesta. [7] 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.

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k\beta_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\gamma_{jk} \qquad (1)$$

llevado al caso

$$g_1(\mu) = \eta_1 = \mathbf{X}_1\beta_1 + \textstyle\sum_{j=1}^{J_1} \mathbf{Z}_{j1}\gamma_{j1}$$

$$g_2(\sigma) = \eta_2 = \mathbf{X}_2\beta_2 + \textstyle\sum_{j=1}^{J_2} \mathbf{Z}_{j2}\gamma_{j2}$$

$$g_3(\nu) = \eta_3 = \mathbf{X}_3\beta_3 + \textstyle\sum_{j=1}^{J_3} \mathbf{Z}_{j3}\gamma_{j3}$$

$$g_4(\tau) = \eta_4 = \mathbf{X}_4\beta_4 + \textstyle\sum_{j=1}^{J_4} \mathbf{Z}_{j4}\gamma_{j4}$$

donde μ,σ,ν,τ, y, para $k = 1,2,3,4$, θ_k y η_k son vectores de largo n, $\beta_k^\top = (\beta_{1k},\beta_{2k},...,\beta_{J_k'k})$ 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 $q_{jk} \times q_{jk}$

$\mathbf{G}_{jk} = \mathbf{G_{jk}}(\lambda_{jk})$, la cual puede depender de un vector de hiperparámetros λ_{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 $\mathbf{G_{jk}^{-1}}$.

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

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k\beta_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \qquad (2)$$

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

$$g_1(\theta_k) = \eta_k = \mathbf{X}_k\beta_k \qquad (3)$$

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:

$$g_k(\theta_k) = \eta_k = h_k(\mathbf{X}_k, \beta_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \qquad (4)$$

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*:

$$g_k(\theta_k) = \eta_k = h_k(\mathbf{X}_k, \beta_k). \qquad (5)$$

Si, adicionalmente, $h_k(\mathbf{X}_k, \beta_k) = \mathbf{X}_k^{\mathsf{T}\beta_k}$ para $i = 1,2,...,n$ 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 $h_k(\mathbf{X}_k, \beta_k)$ 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 paramétros β_k y los parámetros 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

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

donde $\ell(\beta,\gamma) = \sum_{i=1}^{n} log f_Y(y_i|\theta^i) = \sum_{i=1}^{n} log f_Y(y_i|\mu_i,\sigma_i,\nu_i,\tau_i)$ 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 (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 $\ell(\beta)$.

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.

3. Aplicación

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.

Fig. 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 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 $\pm$0.1kg) y se midió su talla (estatura) descalzos mediante un pediómetro digital Sohenle 5001 (error $\pm$0.5cm) 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 [9] a través de la UI (interfáz de usuario) RStudio [10] utilizando las librerías *readr* [11], *tibble* [12], *tidyr* [13], *dplyr* [14], *ggplot2* [15], *gamlss* [6], *ICSNP* [16] y *MASS* [17].

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.

Tabla 1. Descripción de las variables espirométricas del estudio.

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.
Alérgicos	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.

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 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,con penalización SBC, determinar si dicha inclusión es estadísticamente importante.

4. Modelización para FEV_1

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

$$\mu = a_\mu + pb(Talla, edf_{\mu T}) + pb(Edad, edf_{\mu E}) \quad (7)$$

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:

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

El término a_σ es una constante

Asimetría ν y curtosis τ El parámetro de asimetría, ν, se opta por modelizarlo de la forma:

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

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

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

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 FEV_1, con la inclusión de la variable Sexo dentro del mismo, en los parámetros de distribución que sean necesarios.

Tabla 2. Coeficientes de la regresión lineal del modelo

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

- **Modelo GAMLSS por sexo**: Separar a los niños normales por sexo y luego ajustar un modelo para las variablesFEV_1 para los niños y otro para las niñas.

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 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.

4.1. 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.

Tabla 3. Ecuaciones de regresión para FEV_1

Sexo	Ecuación de regresión	SBC
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	$\mu = -2.572 + pb(Talla, 3.65)$ $log(\sigma) = -1.868$ $\nu = 0.799$ $log(\tau) = 0.355$	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.

5. Conclusiones y pasos a futuro

Se llevó a cabo un estudio entre los niños normales y con antecedentes patológicos con el objetivo de ver sus diferencias, y 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.

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.

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 [18] lo hicieron con datos de niños y jovenes entre 4 y 19 años, con modelos lineales partidos a través de la talla. [19] y [20] 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. y se espera a futuro poder emplearlo en otros trabajos del'ambiot sd ela salud (Nutrición) y delas Ciencias Sociales en General. Por otra parte tiene varias herramientas para chequear posibles inadecuaciones en los modelos, como los gráficos de gusano y los cuantiles residuales, entre otros .

Agradecimientos

.

Referencias

[1] R. Álvarez-Vaz, P. Palamarchuk, and E. Riaño, "Elaboración de patrones espirométricos normales en niños uruguayos mediante modelos gam y gamlss:parte 1-identificación de la distribución de la variable de respuesta," Instituto de Estad?ística, Facultadad de Ciencias Económicas y de Administraciónn, Universidad de la República, Uruguay., Serie Documentos de Trabajo. DT (16/3), Nov. 2016. [Online]. Available: https://iesta.fcea.udelar.edu.uy/wp-content/uploads/2020/09/ddt_03_16.pdf

[2] ——, "Elaboración de patrones espirométricos en niños uruguayos mediante modelos gam y gamlss: Parte 2-modelización de cvf y fev por talla edad y sexo." Instituto de Estad?ística, Facultadad de Ciencias Económicas y de Administraciónn, Universidad de la República, Uruguay., Serie Documentos de Trabajo. DT (17/2)., Dec. 2017. [Online]. Available: https://iesta.fcea.udelar.edu.uy/wp-content/uploads/2020/09/ddt_03_16.pdf

[3] J. Nelder and R. Wedderburn, *Generalized Linear Models*. Chapman & Hall/CRC, 1972.

[4] T. Hastie and R. Tibshirani, *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1986, vol. 1, no. 3. [Online]. Available: http://www.jstor.org/stable/2245459

[5] R. Rigby and D. Stasinopoulos, "The GAMLSS project a flexible approach to statistical modelling," 2001. [Online]. Available: http://www.gamlss.org/wp-content/uploads/2013/01/paper044.pdf

[6] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape,(with discussion)," *Applied Statistics*, vol. 54, pp. 507–554, 2005.

[7] D. Stasinopoulos and R. Rigby, "Generalized additive models for location scale and shape (gamlss) in r," *Journal of Statistical Software*, vol. 23, no. 7, 2007.

[8] K. Akanztiliotou, R. Rigby, and D. Stasinopoulos, "The R implementation of Generalized Additive Models for Location, Scale and Shape," 01 2002.

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[10] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA, 2016. [Online]. Available: http://www.rstudio.com/

[11] H. Wickham, J. Hester, and R. Francois, *readr: Read Rectangular Text Data*, 2017, r package version 1.1.1. [Online]. Available: https://CRAN.R-project.org/package=readr

[12] K. Müller and H. Wickham, *tibble: Simple Data Frames*, 2017, r package version 1.3.3. [Online]. Available: https://CRAN.R-project.org/package=tibble

[13] H. Wickham, *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2017, r package version 0.6.3. [Online]. Available: https://CRAN.R-project.org/package=tidyr

[14] H. Wickham, R. Francois, L. Henry, and K. Müller, *dplyr: A Grammar of Data Manipulation*, 2017, r package version 0.7.1. [Online]. Available: https://CRAN.R-project.org/package=dplyr

[15] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. [Online]. Available: http://ggplot2.org

[16] K. Nordhausen, S. Sirkia, H. Oja, and D. E. Tyler, *ICSNP: Tools for Multivariate Nonparametrics*, 2015, r package version 1.1-0. [Online]. Available: https://CRAN.R-project.org/package=ICSNP

[17] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[18] M. Rosenthal and S. Bain, "Lung function in white children aged 4 to 19 years," *Spirometry Thorax*, vol. 48, pp. 794–802, 1993.

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

[20] S. Stanojevic, A. Wade, and T. Cole, "Reference ranges for spirometry across all ages: a new approach," *American Journal of Respiratory and Critical Care Medicine*, vol. 177, no. 3, p. 253–260, 2007.