Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado
Grado en Estadística

Modelos GAMLSS para la obtención de intervalos de referencia

Autor: Dª. Raquel Prado Santiago
Tutores: Dª. Itziar Fernández Martínez
Dª. Yolanda Larriba González

Año 2022-2023

Resumen

Las alteraciones en los niveles de creatinina pueden ser indicador de patologías o fallos en la función renal. Los análisis de orina son una de las pruebas más empleadas en la práctica clínica, por su carácter no invasivo, para recoger información sobre una gran variedad de biomarcadores, entre ellos la creatinina.

Se requiere de herramientas de diagnóstico precisas para evaluar los niveles y fluctuación de estos biomarcadores. Una de los procedimientos más utilizados por su simplicidad y fácil interpretación son los intervalos de referencia (IRs). Estos IRs proporcionan unos rangos globales de los valores de la medida que sirven como regla para determinar si los individuos están dentro de unos límites "normales". Sin embargo, estos límites no siempre son útiles ya que el sexo o la edad podrían influir en las variaciones de los biomarcadores. Como alternativa para la construcción de estos IRs, se han utilizado en la literatura los modelos GAMLSS (del inglés *Generalized Additive Models for Location, Scale and Shape*). Los GAMLSS son modelos aditivos semi-paramétricos muy flexibles que permiten construir IRs ajustados por las variables demográficas.

Este trabajo analiza los datos tomados de la encuesta The National Health and Nutrition Examination Survey (NHANES) sobre el estado de salud de la población americana y propone el uso de los modelos GAMLSS para la construcción de IRs ajustados por edad, sexo y raza para los niveles de creatinina en población pediátrica y adulta. Los resultados muestran que existe una clara dependencia de la edad, sexo y raza, tanto en población pediátrica como adulta, y en los diferentes parámetros de las distribuciones de interés. Esto pone de manifiesto la utilidad de los modelos GAMLSS para la construcción de IRs.

Palabras clave: Modelos Lineales, Modelo GAMLSS, Intervalo de Referencia, Creatinina, NHANES.

Abstract

Alterations in creatinine levels can indicate renal pathologies or failures. Urine analysis is one of the most commonly used tests in clinical practice due to its non-invasive nature, to gather information about a wide variety of biomarkers, including creatinine.

Accurate diagnostic tools are required to evaluate the levels and fluctuations of these biomarkers. One of the most commonly used procedures, due to its simplicity and easy interpretation, is the use of reference intervals (RIs). RIs provide overall ranges of measurement values that serve as a guideline to determine if individuals are within "normal" limits. However, these limits may not always be useful as factors such as sex or age could influence biomarker variations. As an alternative for constructing RIs, the literature has employed Generalized Additive Models for Location, Scale, and Shape (GAMLSS) models. GAMLSS models are highly flexible semi-parametric additive models that allow the construction of RIs adjusted for demographic variables.

This study analyzes data from The National Health and Nutrition Examination Survey (NHANES) on the health status of the American population and proposes the use of GAMLSS models to construct age, sex, and race adjusted RIs for creatinine levels in pediatric and adult populations. The results show a clear dependence on age, sex, and race in both pediatric and adult populations, as well as in the different parameters of the distributions of interest. This highlights the utility of GAMLSS models for constructing RIs.

Keywords: Linear Models, GAMLSS Model, Reference Interval, Creatinine, NHANES

Índice

Introducción

Los análisis clínicos (sangre, orina, heces,...) son una de las prácticas más habituales en medicina para el diagnóstico de patologías. El caso particular de los análisis de orina, destacan por ser pruebas no invasivas, y una excelente fuente de biomarcadores debido a que la orina se genera a partir de la sangre en el riñón con el objetivo de regular la cantidad y calidad del resto de fluidos corporales. Por ello, tanto alteraciones fisiológicas como enfermedades renales se reflejan en la composición y propiedades de la orina alterando los niveles de los biomarcadores. En particular, la concentración de creatinina en orina, producto de desecho resultado del proceso metabólico muscular, es un biomarcador de referencia de la función renal. Los niveles de creatina se han utilizado como indicadores de posibles patologías renales crónicas o de daño agudo de riñón [1]. Sin embargo, los niveles de creatinina, y en general de la gran mayoría de biomarcadores, varían dependiendo del sexo, de la edad, la raza o incluso la condición física.

Por tanto, resulta indispensable para la práctica clínica disponer de herramientas precisas de diagnóstico para evaluar los niveles y fluctuaciones de biomarcadores como la creatinina. Uno de los procedimientos más estándar y de uso más extendido en medicina, por su simplicidad a la hora de construirlo y su fácil interpretación, es la construcción de Intervalos de Referencia (IRs). Habitualmente un IR representa el intervalo entre dos percentiles extremos y simétricos, destacando entre los más utilizados en la práctica los IR del 95 % (percentil 2.5-97.5) y del 90 % (percentil 5.0-95.0) [2]. Los IRs permiten distinguir entre niveles "normales" y "anormales" de una medida fisiológica a partir de una muestra de pacientes sanos. Es decir, proporcionan una regla de diagnóstico que permite clasificar un valor como potencialmente patológico si una pequeña proporción de pacientes en la población sana tiene valores muy extremos (altos o bajos) en los niveles de la medida. Este hecho pone de manifiesto una de las principales dificultades en la construcción de los IRs: la selección de una muestra representativa de la población sana. También es importante resaltar que los valores de referencia globales que proporcionan los IRs no siempre son útiles, ya que características demográficas, como la edad y el sexo, influyen en las variaciones biológicas de los niveles de los biomarcadores [3].

Las limitaciones anteriores ponen de manifiesto la necesidad de una metodología alternativa más flexible que considere las características demográficas en la construcción de los IRs como los modelos GAMLSS (del inglés *Generalized Additive Models for Location, Scale and Shape*). Los GAMLSS son modelos de regresión semi paramétricos que permiten modelizar los parámetros de localización, escala y forma (asimetría y curtosis) de la distribución de la variable respuesta como funciones lineales, no lineales, o bien como funciones suaves no paramétricas de las variables explicativas. La flexibilidad de estos modelos también se debe a la gran variedad de familias de distribuciones a las que la distribución de la variable respuesta puede pertenecer. Los modelos GAMLSS han sido utilizados, entre otros, por la Organización Mundial de la Salud (OMS) para diseñar curvas de crecimiento de referencia demostrando ser un procedimiento adecuado para la construcción de IRs [4], [5].

El objetivo principal de este trabajo consiste en utilizar los modelos GAMLSS para la construcción de IRs ajustados por edad, sexo y raza en población pediátrica y adulta para los niveles de creatinina en orina a partir de los datos tomados de la encuesta The National Health and Nutrition Examination Survey (NHANES) que recoge anualmente datos sobre el estado de salud de la población americana.

Otros objetivos secundarios y menos aplicados de este trabajo son:

- Conocer y comparar las ventajas e inconvenientes de los principales métodos de construcción de IRs.
- Estudiar distintos modelos lineales, analizando sus principales diferencias y procedimientos de estimación y ajuste.
- Profundizar en el estudio de los modelos GAMLSS: principales característica, posibles distribuciones de la variable respuesta, selección y adecuación del modelo y cálculo de los percentiles para la construcción de los IRs.

La estructura de este trabajo consta de cuatro partes bien diferenciadas. En primer lugar se presenta una aproximación al concepto y las principales metodologías para la construcción de IRs, haciendo hincapié en las principales dificultades que deben tenerse en cuenta para que los IRs sean válidos. A continuación se describe el procedimiento estadístico propuesto en este trabajo para la construcción de IRs a partir de la estimación de los percentiles en los modelos GAMLSS. Para llegar hasta aquí se describen y comparan previamente distintos modelos lineales y se realiza una descripción completa de los modelos GAMLSS (estimación, ajuste, selección de regresores y adecuación del modelo). En tercer lugar se realiza un análisis descriptivo de los datos relativos a la creatinina tomados de la encuesta NHANES. A continuación, se selecciona el modelo GAMLSS más adecuado para cada estrato (población pediátrica y adulta) y se construyen los IRs ajustados por edad, sexo y raza, y se interpretan los resultados. Por último se incluye una discusión con las principales conclusiones de este trabajo y otras consideraciones a tener en cuenta como trabajo futuro.

Las asignaturas del Grado en Estadística que han tenido más relevancia a la hora de realizar este trabajo son las siguientes:

- Estadística Descriptiva: se presentan técnicas básicas de estadística descriptiva imprescindibles para el análisis de datos.
- Modelos probabilísticos: se describen los modelos probabilísticos básico y su utilidad en la práctica, sirviendo como base para modelos más complejos como los de este trabajo.
- Computación Estadística: se dan a conocer las principales técnicas computacionales para realizar análisis estadísticos de forma eficiente.
- Modelos Lineales, Regresión y ANOVA, Modelos Estadísticos Avanzados: se introducen los modelos lineales y los modelos GLM utilizados en este trabajo.
- Inferencia Estadística I, Inferencia Estadística II: se estudian los principales métodos inferenciales paramétricos y no paramétricos de estimación y contraste hipótesis, piedras angulares de este trabajo.

Intervalos de Referencia

2.1. Definición

En el campo de la medicina, y para una determinada medida fisiológica, un Intervalo de Referencia (IR) es el rango o intervalo de valores considerado "normal" entre personas sanas.

La necesidad de establecer IRs surgió con el desarrollo de la medicina moderna y la capacidad de realizar pruebas diagnósticas con mayor precisión. A medida que se adquiría más información sobre la fisiología humana y los procesos patológicos, se hizo evidente que era necesario establecer una escala de comparación para los resultados de las pruebas.

Son una de las herramientas más utilizada como apoyo a la hora de tomar decisiones médicas, por tanto es importante que los especialistas clínicos estén familiarizados con ellos y conozcan a la perfección la manera correcta de interpretarlos.

La historia de los IRs se remonta a la época en que se empezaron a desarrollar pruebas diagnósticas médicas. En la década de 1930, se establecieron los primeros valores normales para los análisis de sangre y orina, y desde entonces, se han desarrollado y actualizado continuamente.

En 1978, La Federación Internacional de Química Clínica y Medicina de Laboratorio (IFCC del inglés *International Federation of Clinical Chemistry and Laboratory*), publicó las primeras recomendaciones oficiales definidas por un panel de expertos sobre la teoría y generación de valores de referencia. En ellas, se define por primera vez el término IR [6].

Tras estas primeras recomendaciones, y basándose en ellas, otras sociedades científicas de distintos países publicaron sus propias recomendaciones. Años después, en 2005, con la finalidad del establecimiento de estándares globales, para apoyar a sus miembros a través de esfuerzos científicos y educativos, además de celebrar una serie de congresos, conferencias y reuniones, se creó el Comité de Intervalos de Referencia y Valores de Decisión Clínica (C-RIDL del inglés *Committee on Reference Intervals and Decision Limits*) con la ayuda de la IFCC.

Posteriormente, en 2010 el Instituto de Estándares de Laboratorio clínico publicó una guía definitiva para la definición, establecimiento y verificación de intervalos de referencia basándose en las recomendaciones originales mencionadas anteriormente. Este último documento ha sido muy utilizado para la obtención de IRs [7].

Los IRs se utilizan en diferentes contextos en la medicina y la salud, algunos de ellos son:

1. Diagnóstico clínico: sirven para ayudar a los médicos a determinar si los resultados de las pruebas de laboratorio de un paciente están dentro de los límites "normales" o no.

Si los resultados están fuera de los IRs, esto puede indicar una afección subyacente.
2. Monitoreo de enfermedades: se utilizan para monitorear la evolución de una enfermedad en un paciente con el fin de estudiar las medidas tomadas a lo largo del tiempo y poder realizar un seguimiento de la enfermedad.
3. Investigación clínica: se emplean en estudios clínicos para establecer grupos de comparación y para determinar si los tratamientos son efectivos.
4. Calibración de equipos: se usan para calibrar los equipos de laboratorio y garantizar que los resultados de las pruebas sean precisos y fiables.

Hay algunos malos usos comunes de los IRs que pueden llevar a errores en la interpretación de los resultados de las pruebas y, en consecuencia, a decisiones de atención médica equivocadas, como por ejemplo la prescripción de un tratamiento que no tiene en cuenta características importantes del paciente, como la edad o el sexo.

Para actuar de forma correcta, lo primero es revisar mediante un análisis exploratorio los datos que se vayan a utilizar para la construcción de los intervalos, teniendo en cuenta que los datos que estén fuertemente influidos por algún factor pueden representar un individuo enfermo, considerado *a priori* sano en la muestra. Después, se debe elegir un método de estimación adecuado, en el que es muy importante validar las asunciones que se deben verificar para la correcta aplicación del método elegido para la construcción de IRs. Otro paso a tener en cuenta, es el estudio sobre la posibilidad de que algunos factores clínicos puedan afectar a la estimación, como por ejemplo, la presencia de algún indicador de enfermedad hereditaria aunque no esté manifestada. Por último, se debe realizar una correcta interpretación de los resultados obtenidos.

Es muy importante utilizar IRs apropiados para la población específica que se vaya a estudiar y tener en cuenta que los IRs no son una línea divisoria entre "normal" y "anormal", sino un rango de valores en el que se encuentran la mayoría de los valores "normales". Por lo tanto, los resultados de las pruebas cerca de los límites de los IRs deben interpretarse con precaución.

A la hora de usar los IRs es importante considerar otros factores clínicos, como los antecedentes médicos, los síntomas y los resultados de otras pruebas. También hay que tener en cuenta que los IRs son una herramienta valiosa para ayudar a interpretar los resultados de las pruebas, pero no deben ser la única fuente de información utilizada para la toma de decisiones, ya que evidentemente muchos de estos biomarcadores van a verse afectados por las características propias de cada individuo.

Uno de los problemas que aparecen en la construcción de los IRs a la hora de definirlos son las variaciones intra e interindividuales de los resultados por factores como la edad, el sexo, la raza y la dieta, que son muy importantes a la hora de calcular, interpretar y comunicar los IRs [8].

Otro de los posibles problemas frecuentes sería la obtención de muestras representativas de todos los individuos sanos. Por una parte, no siempre está clara la definición de "individuo sano" y, por otra parte, el tamaño de la muestra requerido para conseguir estimaciones

fiables puede ser excesivamente grande.

Estos problemas pueden llevar a que se tomen decisiones incorrectas, lo que aumenta los riesgos innecesarios en la seguridad del paciente y aumenta los costes de las investigaciones.

2.2. Construcción de los IRs

En general los IRs se establecen como los valores entre los cuales están comprendidos el 90 % o 95 % de los valores centrales observados en una muestra representativa de la *población de referencia*.

Para el cálculo de estos intervalos primeramente hay que definir la *población de referencia*, es decir, el grupo de individuos que va a quedar representado por el IR, habitualmente la población sana [7]. La definición de la *población de referencia* no siempre es fácil, ya que es necesario tener en cuenta factores como la edad, el sexo, la raza, etc, que pueden influir en las medidas que se están analizando. En algunos casos incluso será necesario construir distintos IRs para estratos de la población según estos factores.

Por ejemplo, es común que los IRs de muchos marcadores sean diferentes para niños que para adultos, o que se apliquen sólo a subgrupos poblacionales, como en el caso de mujeres embarazadas.

Por tanto, para delimitar la población de referencia es muy importante establecer de forma clara los criterios de inclusión y exclusión, es decir, las características de la población que hacen a los individuos elegibles o no elegibles para participar en el estudio, respectivamente. De igual forma son importantes los criterios de exclusión para el filtrado de los datos que serán diferentes dependiendo del contexto que tenga la prueba que se quiere realizar, ya que puede ser importante eliminar los individuos con ciertas patologías concretas, enfermedades o bajo algún tratamiento con determinados fármacos.

Una vez se haya definido la población de referencia, hay que extraer una muestra de referencia. Cuando la población no es homogénea una técnica útil es el muestreo estratificado, en el que se divide la población de referencia en subgrupos o estratos y se extrae una muestra representativa de cada estrato.

Básicamente, los IRs biológicos se pueden estimar de dos formas, mediante el método paramétrico y el método no paramétrico.

- **El método paramétrico.**
  Este procedimiento asume que la distribución subyacente del marcador de interés (Y) es Normal, es decir que $Y \sim \mathcal{N}(\mu, \sigma)$.
  En este caso, la construcción del IRs se reduce a calcular un intervalo de confianza (IC) $100(1-\alpha)\,\%$ para μ : $[\overline{Y} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}]$, donde $\overline{Y}$ es la media muestral, S es la cuasidesviación típica muestral (estimador de σ), $z_{\alpha/2}$ es el cuantil $1-\alpha/2$ de una distribución Normal estándar y n es el tamaño de la muestra.
  El valor α se establece en 0.05 o 0.1 para obtener los IRs del 95 % o 90 %, respectivamente.
  En muchos casos, donde no es posible asumir la normalidad, una simple transfor-

mación de la variable de interés permite utilizar este método.

- **El método no paramétrico.**
  Este método tiene la ventaja de que no requiere de la asunción de normalidad, hipótesis que no se verifica en muchos de los casos.
  En este caso, el IR se construye como un intervalo de percentiles, es decir: $[y_{\alpha/2}, y_{1-\alpha/2}]$ donde y_α representa el percentil 100α de la variable Y.
  Como en el método paramétrico, el valor α se establece en 0.05 o 0.1 para los IRs del 95 % y 90 %, respectivamente.

2.3. Aspectos a tener en cuenta en la construcción de IRs

Antes de la construcción de los IRs es importante considerar varios aspectos que pueden influir en la validez de los resultados obtenidos.

2.3.1. Definición de Población de referencia

En cualquier estudio clínico, una observación determinada no tiene valor por si misma, sino que debe haber algún valor de referencia para poder interpretarla a través de la comparación. En este sentido, podría decirse que la interpretación de una observación concreta es un proceso de comparación [7].

Según el objetivo del estudio, la referencia puede ser distinta. Por ejemplo, podrían usarse unos valores de decisión establecidos a partir de información científica publicada en un momento concreto, o unos valores previos del mismo individuo que permitan interpretar el cambio. Cuando se definen los IRs objeto de este trabajo, los valores de comparación se establecen a partir de lo observado en un grupo de individuos sin patología que tienen unas características determinadas y bien definidas.

Aunque ya se han comentado en este trabajo algunos conceptos claves para establecer formalmente los IRs, será necesario definir claramente lo que es un individuo, una población y una muestra de referencia. La IFCC establece las siguientes definiciones:

- Un individuo de referencia es un individuo que puede ser seleccionado, utilizando unos criterios previamente definidos, para llevar a cabo la comparación.
- Una población de referencia está formada por todos los posibles individuos de referencia, que por lo general tendrá un tamaño desconocido.
- Una muestra de referencia es un número adecuado de individuos de referencia seleccionados para representar a la población de referencia.

A pesar de que estas definiciones son claras, en la práctica es necesario tener en cuenta ciertos aspectos:

- Fuente de los datos: es importante extraer los datos de una fuente fiable, bien documentada, que proporcione datos actualizados y precisos, que utilice una metodología confiable y sea transparente en su proceso de recopilación y presentación de los datos.
- Tamaño de la población: la muestra debe ser representativa de la población, cuyo tamaño, en principio desconocido, será grande. Lo habitual es disponer de un mínimo

de 750-1000 individuos por grupo para garantizar cierta robustez estadística, esto puede ser un problema en la práctica, tanto por el hecho de conseguir este número de individuos, como por el coste que ésto supone.

- Periodo de recopilación de datos: obtener muestras tan grandes requiere mucho tiempo de trabajo en la recogida de datos. Periodos de tiempo demasiado prolongados puede suponer que cambien las formas de obtener las medidas, esto incluye desde cambios en la tecnología, hasta cambios en los materiales que se necesitan para hacer las mediciones.
- Estratificación de los datos: típicamente, las variables analizadas se van a ver influenciadas entre otras, por algunas características de los individuos como son el sexo, la edad y la raza. Es importante verificar que no existen diferencias estadística o clínicamente relevantes entre ellos, de ser así, se deben establecer IRs basados en estos grupos debido a las implicaciones que pueden tener a la hora de tomar cualquier decisión basada en ellos. Para determinar si es necesaria una partición en grupos, podemos realizar un análisis descriptivo o inferencial y comprobar si existen diferencias, o llevar a cabo pruebas estadísticas más concretas.
- Criterios de exclusión: dependiendo del contexto clínico de la prueba, puede que sea importante la eliminación del estudio de ciertos individuos que presenten alguna patología concreta o que esté en tratamiento con algún fármaco determinado.
- Definición de "individuo sano": en muchos casos los IRs se obtendrán en un grupo de individuos considerados sanos, entendiendo, según la OMS, que la salud es un estado de completo bienestar físico, mental y social y no únicamente la ausencia de afección o enfermedad. Utilizando esta definición, el concepto de salud puede ser diferente según las culturas y los países.
  Desde un punto de vista práctico, podría entenderse la salud como la ausencia de signos de enfermedad o condiciones que se puedan relacionar específicamente con las mediciones que se están analizando [9].

2.3.2. Sesgos

Cuando se realiza el muestreo de los datos, hay que tener en cuenta la existencia de un factor muy importante, el sesgo, que puede tener un impacto significativo en la interpretación de los resultados. El sesgo también puede afectar a la precisión y la sensibilidad de las pruebas, lo que puede llevar a una mayor probabilidad de resultados falsos positivos o falsos negativos, ya que los IRs en estos casos, suelen ser demasiado amplios o estrechos.

Por ejemplo, supongamos que se realiza un estudio para determinar los niveles de creatinina en sangre u orina en diferentes grupos de edad, desde adolescentes hasta adultos mayores. El objetivo es investigar si existen diferencias significativas en los niveles de creatinina entre los grupos de edad. Durante la recolección de datos, se observa que la muestra de adolescentes incluye principalmente a atletas de alto rendimiento, mientras que la muestra de adultos mayores está compuesta en su mayoría por individuos sedentarios. Esta diferencia en la selección de la muestra puede introducir un sesgo en los resultados.

2.3.3. Presencia de *outliers*

Los valores atípicos o *outliers* son las observaciones cuya discordancia respecto a la mayoría de la muestra es excesiva en relación a la distribución de dicha mayoría. Su presencia puede tener un impacto significativo en la interpretación de los resultados de las pruebas diagnóstico y en la toma de decisiones [2].

Existen varias técnicas para corregir los efectos de los outliers en la estimación de los IRs:

- Eliminación de *outliers*: una forma de corregir los efectos de los *outliers* es eliminarlos de la muestra antes de calcular los IRs. Sin embargo, esto debe hacerse cuidadosamente y solo después de una revisión rigurosa de los datos, ya que la eliminación de datos puede tener un impacto en la precisión de los intervalos y lo que es más importante, en la utilidad de estos intervalos.
- Transformación de datos: otra forma de corregir los efectos de los *outliers* es transformar los datos antes de calcular los IRs. Por ejemplo, se puede aplicar la transformación logarítmica a los datos para reducir el impacto de los valores extremadamente altos en su distribución. Esto puede ser útil cuando los *outliers* son el resultado de distribuciones de datos no normales.
- Utilización de métodos robustos: son menos sensibles a los *outliers* que los métodos convencionales. Estos métodos pueden incluir la mediana y la desviación intercuartílica en lugar de la media y la desviación estándar.

Es importante que cualquier corrección se realice cuidadosamente y de forma adecuada para garantizar que los IRs sean precisos y representativos de la población a la que se aplican.

2.3.4. Estratificación por grupos

Para muchos análisis es necesario dividir a la población en grupos a la hora de realizar el muestreo [2]. En la práctica es habitual que los datos se particionan en grupos sin antes analizar si estos son comparables. Por ello es primordial determinar el posible efecto sobre la variable por la que queremos hacer la separación en la variable de interés (ej: hombres y mujeres). Una metodología para este procedimiento puede verse en [2, 10]

2.4. Métodos alternativos para la construcción de los IRs

Como ya se ha comentado, la forma más simple de establecer los IRs es determinar una pareja de valores entre los que se espera se encuentren una parte importante de los valores de la población de referencia. Para estimar esos valores se pueden estimar los percentiles de la distribución de probabilidad de la variable de interés. Los métodos más utilizados se basan en suponer una distribución Normal para los datos.

El método estándar para la construcción de IRs se basa en la estimación de IC, pero este método sólo es válido si los datos provienen de una población con distribución gausiana. En el caso de que los datos provengan de una distribución distinta de la Normal es necesario realizar una transfomación para conseguir que los datos sigan la distribución anterior,

después se calculan los intervalos y se vuelve a realizar otra transformación a las unidades originales.

Sin embargo, en la práctica, no suele ser adecuado utilizar directamente el modelo de probabilidad Normal, ya que la mayoría de parámetros biológicos suelen alejarse de ese modelo presentando asimetría o apuntamiento. Algunas veces una transformación de la variable de interés, como por ejemplo la transformación logarítmica, que podría solucionar estos problemas, no siempre es posible.

Por otra parte, es habitual encontrarse con que existe una dependencia de la variable de interés respecto a otras variables o factores como la edad y el sexo. En estos casos, no será adecuado proporcionar unos límites de referencia globales, sino que estos deberían ser determinados en función de estas variables explicativas. Utilizar por ejemplo modelos de regresión para estimar estas funciones podría ser una solución.

Desde esta perspectiva, la solución pasaría por utilizar modelos en los que la distribución de la variable respuesta no esté restringida a una distribución Normal y, además, no sólo permitan modelizar el parámetro de localización de dicha distribución, sino que la escala y la forma de la distribución puedan depender de una o varias variables explicativas de interés como ocurre en los modelos GAMLSS

Procedimiento estadístico

En la primera parte de esta sección se presenta el modelo de regresión lineal (LM, del inglés *Linear Model*), el modelo lineal generalizado (GLM, del inglés *Generalized Linear Model*), el modelo aditivo generalizado (GAM, del inglés *Generalized Additive Model*) y el modelo GAMLSS haciendo hincapié en las limitaciones de cada uno de ellos.

En la segunda parte, se describen algunas propiedades de interés de los modelos GAMLSS, así como su utilidad para la construcción de límites de referencia.

3.1. Modelos lineales

A continuación, se presenta una breve descripción de algunos modelos lineales para este trabajo junto con los principales métodos de estimación de sus parámetros.

3.1.0.1 Modelo de regresión lineal

El LM es un modelo simple que se utiliza para analizar la relación lineal entre una variable independiente y una o más variables dependientes. Este modelo asume que dado un conjunto de observaciones, la media μ de la variable dependiente o respuesta Y se relaciona de forma lineal con las variables regresoras:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

donde $\beta_0, \beta_1, ..., \beta_p$ son los coeficientes de regresión que representan el efecto de cada variable independiente sobre la variable dependiente, siendo p el número de regresores.

También se puede expresar de forma matricial:

$$\mu = \mathbf{X}\beta$$

donde μ es un vector de dimensiones $n \times 1$, $\mathbf{X}$ es la matriz de diseño de dimensiones $n \times p$ y β es un vector de dimensiones $p \times 1$, siendo n el número de observaciones.

Las ventajas del modelo de regresión lineal incluyen su simplicidad y facilidad de interpretación. Sin embargo, el LM tiene ciertas limitaciones como su capacidad para modelizar relaciones no lineales entre las variables independientes y la variable dependiente. Además, asume que los errores son independientes, tienen una distribución Normal centrada en 0 y varianza constante ($\mathcal{N}(0, \sigma)$). Estas suposiciones pueden no ser válidas en todos los casos, limitando la utilidad de este modelo.

3.1.0.2 Modelo lineal generalizado

Para superar las limitaciones de los LM, John Nelder y Robert Wedderburn [11] desarrollan los modelos GLM.

Previamente a la caracterización de estos modelos, es necesario introducir la familia exponencial de distribuciones.

Familia Exponencial de Distribuciones

Bajo la caracterización general de la familia exponencial de distribuciones se engloban distribuciones muy conocidas como la Normal, la binomial o la Poisson. Esta familia además cuenta con muy buenas propiedades teóricas, véase [12] para más información.

Sea Y una variable aleatoria cuya función de densidad $f_Y(y;\theta)$ depende del parámetro θ. Se dice que su distribución pertenece a la familia exponencial de distribuciones si:

$$f_Y(y;\theta) = s(y)t(\theta)\exp\{a(y)b(\theta)\},$$

donde $a(\cdot)$, $b(\cdot)$, $s(\cdot)$ y $t(\cdot)$ son funciones conocidas. La ecuación anterior puede reescribirse de la siguiente manera donde se evidencia el papel simétrico que juegan θ e y.

$$f_Y(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

con $s(y) = \exp\{d(y)\}$ y $t(\theta) = \exp\{c(\theta)\}$. Al término $b(\theta)$ se le denomina parámetro natural de la distribución. En el caso particular que $a(y) = y$ se dice que la distribución está expresada en su forma canónica.

De las propiedades de la familia exponencial de distribuciones se deduce que:

$$E_\theta(a(Y)) = \frac{-c^{'}(\theta)}{b^{'}(\theta)} \tag{1}$$

$$Var_\theta(a(Y)) = \frac{b^{''}(\theta)c^{'}(\theta) - c^{''}(\theta)b^{'}(\theta)}{[b^{'}(\theta)]^3} = \frac{-b^{''}(\theta)E_\theta(a(Y)) - c^{''}(\theta)}{[b^{'}(\theta)]^2} \tag{2}$$

• • •

En los modelos GLM la variable respuesta sigue una distribución que pertenece a la familia exponencial de distribuciones expresada en su forma canónica. Además, la dependencia lineal de la media de la variable respuesta (μ) con los regresores se modeliza a través de una función $g(\cdot)$, denominada función de enlace de la forma siguiente:

$$\eta = g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

En forma matricial se expresa:

$$\eta = g(\mu) = \mathbf{X}\beta$$

Nótese que un caso particular de GLM cuando se considera como función de enlace la identidad, es el LM.

Las ventajas del GLM incluyen su flexibilidad para modelizar diferentes tipos de variables dependientes y su capacidad para modelizar relaciones no lineales a través de la función de enlace. Sin embargo, el GLM también tiene limitaciones, la principal tiene que ver con su capacidad para modelizar relaciones complejas que incluyen dependencias no lineales. Además, en algunos casos la elección de la función de enlace y la distribución de probabilidad adecuadas puede no resultar sencillo.

3.1.0.3 Modelo aditivo generalizado

A diferencia de los modelos GLM, que requieren una relación lineal y constante entre los regresores y la media de la variable respuesta mediante la función de enlace, los modelos GAM permiten una relación más flexible y no lineal. Fueron introducidos por Hastie y Tibshirani [13] en 1990, ampliando el enfoque de los modelos GLM y proporcionando una herramienta más dinámica para el análisis de datos.

En los modelos GAM la dependencia lineal de los regresores con $g(\mu)$ viene dada por:

$$\eta = g(\mu) = \beta_0 + l_1(x_1) + l_2(x_2) + ... + l_p(x_p)$$

donde $l_j(\cdot)$, $j = 1, \ldots, p$, son funciones desconocidas cualesquiera. En la práctica, las funciones más empleadas son funciones de suavizado no lineales como por ejemplo los *splines* no paramétricos [13].

Las ventajas del modelo GAM incluyen la capacidad para modelizar relaciones no lineales complejas entre las variables independientes y la variable dependiente y su flexibilidad para incluir diferentes tipos de términos suaves no paramétricos en el modelo. Sin embargo, el modelo GAM también tiene algunas limitaciones, como una mayor exigencia computacional en el ajuste y el riesgo de sobreajuste si se incluyen demasiados términos no paramétricos.

3.1.0.4 Modelos aditivos generalizados para posición, escala y forma

Los modelos GLM y GAM, se limitan a situaciones donde la variable respuesta Y sigue una distribución de la familia exponencial. Además, en estos modelos, se establece una relación directa de la media μ de la variable respuesta como función de los regresores. Sin embargo, no ocurre lo mismo para la varianza, la asimetría o la curtosis. En los modelos GLM y GAM estos parámetros se modelizan de forma indirecta a través de su relación con la media como se puede deducir de la ecuación 2, en lugar de hacerlo directamente como función de los regresores.

Los modelos GAMLSS, introducidos por Rigby y Stasinopoulos en 2005 [14], son una extensión de los modelos GLM y GAM que superan las limitaciones anteriores. En estos modelos la distribución de la variable respuesta no pertenece necesariamente a la familia exponencial de distribuciones, modelizando explícitamente cada uno de los parámetros en función de las variables regresoras, a partir de funciones lineales y no lineales. Esto amplía la capacidad de los modelos para capturar patrones complejos y ajustarse de manera más precisa a los datos.

En los GAMLSS la dependencia de los parámetros de localización (μ), escala (σ) y forma (ν, τ, simetría y curtosis, respectivamente) se expresa de la siguiente forma:

$$\eta_1 = g_1(\mu) = \mathbf{X}\beta + l_1(x_1) + l_2(x_2) + \cdots + l_p(x_p)$$

$$\eta_2 = g_2(\sigma) = \mathbf{X}\beta + l_1(x_1) + l_2(x_2) + \cdots + l_p(x_p)$$

$$\eta_3 = g_3(\nu) = \mathbf{X}\beta + l_1(x_1) + l_2(x_2) + \cdots + l_p(x_p)$$
$$\eta_4 = g_4(\tau) = \mathbf{X}\beta + l_1(x_1) + l_2(x_2) + \cdots + l_p(x_p)$$

donde $\mathbf{X}\beta$ contiene los términos lineales del modelo y $l_j(\cdot)$, $j = 1, \ldots, p$ generalmente son funciones de suavizado no lineales y $g_h(\cdot)$, $h = 1, 2, 3, 4$, son las funciones enlace para cada parámetro.

Las ventajas del modelo GAMLSS incluyen la capacidad para modelizar una amplia variedad relaciones que incorporan formas lineales y no lineales para las características de la distribución de la variable dependiente, la flexibilidad para modelizar diferentes distribuciones de probabilidad y la capacidad para modelizar heterogeneidad en la varianza. Además, también proporciona información útil sobre la forma de la distribución de la variable dependiente. Todo esto amplía el rango de aplicabilidad de estos modelos a las distintas situaciones.

3.1.1. Estimación y ajuste de los modelos lineales

A continuación, se describen los principales métodos de estimación empleados para los modelos anteriores. En general, el método más utilizado es la estimación por máxima verosimilitud (MV).

3.1.1.1 Modelo de regresión lineal

Ajustar este modelo consiste en estimar los valores de los coeficientes de regresión $\widehat{\beta}$ y la varianza $\widehat{\sigma}^2$ que maximizan la verosimilitud de los datos. El método empleado con más frecuencia es el ajuste por mínimos cuadrados (OLS, del inglés *Ordinary Least Squares*) que coincide en este caso con la estimación por MV. El ajuste OLS busca minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo:

$$\hat{\beta} = \underset{\beta_0,\beta}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}\beta)^2$$

A partir de estos estimadores se obtiene:

$$\widehat{\mu} = \mathbf{X}\widehat{\beta}$$
$$\widehat{\sigma}^2 = \frac{(\mathbf{Y} - \widehat{\mu})^T(\mathbf{Y} - \widehat{\mu})}{n - p}$$

donde n es el número de observaciones, p el número de regresores e Y el vector de $n \times 1$ de datos observados .

3.1.1.2 Modelo lineal generalizado

La estimación en los GLM se lleva a cabo por MV. Sin embargo, las ecuaciones de verosimilitud no siempre pueden resolverse directamente. En estos casos, su solución puede aproximarse por procedimientos iterativos como el *iteratively reweighted least squares*

(IRLSM) o método de Newton-Rapson a partir de la esperanza de las segundas derivadas de la función de verosimilitud. Una descripción completa de estos métodos puede verse en [15].

3.1.1.3 Modelo aditivo generalizado

El ajuste de los modelos GAM se lleva a cabo habitualmente mediante un algoritmo de tipo *backfitting*. El empleo de este tipo de algoritmos es característico en el ajuste de modelos aditivos, como los modelos GAM [9]. El algoritmo de *backfitting* parte una solución inicial y resuelve iterativamente un problema de mínimos cuadrados sobre los residuos parciales, estimando en cada paso las funciones de suavizado. El proceso se repite hasta que el algoritmo converge, bien porque la solución no mejore lo suficiente, o bien porque se haya alcanzado un número N máximo de iteraciones. A continuación se proporciona una esquematización del algoritmo *backfitting*.

Sea $S_j(\cdot)$ la función de suavizado empleada (e.g. *spline*) para cada término $l_j(\cdot)$ y sea $\boldsymbol{l}_j = l_j(\mathbf{X}_{\cdot j})$, donde $\mathbf{X}_{\cdot j}$ denota la j-ésima columna en la matriz de diseño $\mathbf{X}$. Los pasos básicos del algoritmo *backfitting* para el ajuste de modelos aditivos son:

1. Inicialización: $\hat{\beta}_0 = \overline{\mathbf{Y}}$, $\hat{\boldsymbol{l}}_j^{(1)} = \mathbf{0}, \forall j$, $I = 0$
2. $I = I + 1$
3. Para $j = 1, \ldots, p$ calcular:

$$\hat{\boldsymbol{l}}_j^{\star} = S_j(\mathbf{Y} - \hat{\beta}_0 \cdot \mathbf{1} - \sum_{k<j} \hat{\boldsymbol{l}}_k^{(I+1)} - \sum_{k>j} \hat{\boldsymbol{l}}_k^{(I)})$$

$$\hat{\boldsymbol{l}}_j^{(I+1)} = \hat{\boldsymbol{l}}_j^{\star} - \overline{\hat{\boldsymbol{l}}_j^{\star}} \cdot \mathbf{1}$$

4. Repetir 2 y 3 hasta que $\text{máx}_j \, \|\hat{\boldsymbol{l}}_j^{(I+1)} - \hat{\boldsymbol{l}}_j^{(I)}\| < \delta$ o $I > N$, donde δ es una cantidad suficientemente pequeña.

En particular, para el ajuste de los modelos GAM se emplea una versión con pesos del algoritmo anterior, véase [13] para más detalle.

3.1.1.4 Modelos aditivos generalizados para posición, escala y forma

Los modelos GAMLSS se engloban también dentro de los modelos aditivos. Por ello, un algoritmo de tipo *backfitting* resulta, de nuevo, útil para ajustar estos modelos. En este caso, en 1990 Hastie and Tibshirani proponen una versión modificada del algoritmo que resuelve el problema de minimización para los residuos parciales considerando por separado la parte lineal y la no lineal, véase [13] para más detalle.

3.2. Los modelos GAMLSS

En esta sección se profundiza en las propiedadess de los modelos GAMLSS y su utilización para la construcción de IRs.

3.2.1. Distribuciones y características

Las distribuciones que se adaptan a las necesidades de este trabajo para el ajuste de los modelos GAMLSS son las que se presentan en la Tabla 1, las familias de distribuciones continuas definidas en $(0, \infty)$. La información completa se encuentra en [16], Apendice A. Para su implementación se ha utilizado la librería gamlss [17], una librería que contiene diferentes familias de distribuciones para la variable respuesta, tanto continuas como discretas o mixtas. Para más información se puede consultar el anexo C.

Acrónimo	**Distribución**	f_Y	**Nº parámetros**	**Características de forma**		**Función de enlace**			
				Asimetría	Curtosis	μ	σ	ν	τ
BCCG	Box-Cox cole and Green	$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right], \text{ si } \nu > 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), \text{ si } \nu = 0 \end{cases}$ $f_Y(y;\mu,\sigma,\nu) = \frac{y^{\nu-1}\exp(-\frac{1}{2}z^2)}{\mu^{\nu}\sigma\sqrt{2\pi}\Phi(\frac{1}{\sigma\lvert\nu\rvert})}$	3	ambos	-	identidad	logit	identidad	-
BCPE	Box-Cox Power Exponential	$Z = \begin{cases} \frac{1}{\sigma\tau}\left[\left(\frac{Y}{\mu}\right)^{\tau} - 1\right], \text{ si } \tau > 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), \text{ si } \tau = 0 \end{cases}$ $f_Y(y;\mu,\sigma,\nu,\tau) = \frac{y^{\nu-1}f_T(z)}{\mu^{\nu}\sigma F_T(\frac{1}{\sigma\lvert\nu\rvert})}$ $T \sim \mathbf{PE}(0,1,\tau)$	4	ambos	ambos	identidad	logit	identidad	log
BCT	Box-Cox-t	$Z = \begin{cases} \frac{1}{\sigma\tau}\left[\left(\frac{Y}{\mu}\right)^{\tau} - 1\right], \text{ si } \tau > 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), \text{ si } \tau = 0 \end{cases}$ $f_Y(y;\mu,\sigma,\nu,\tau) = \frac{y^{\nu-1}f_T(z)}{\mu^{\nu}\sigma F_T(\frac{1}{\sigma\lvert\nu\rvert})}$ $T \sim \mathbf{TF}(0,1,\tau)$	4	ambos	lepto	identidad	log	identidad	log
GA	Gamma	$f_Y(y;\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}} \frac{y^{\frac{1}{\sigma^2}-1}\exp\{\frac{-y}{(\sigma^2\mu)}\}}{\Gamma(\frac{1}{\sigma^2})}$	2	positiva	-	log	log	-	-
GB2	Generalized Beta Type 2	$f_Y(y;\mu,\sigma,\nu,\tau) = \frac{\Gamma(\nu+\tau)}{\Gamma(\nu)\Gamma(\tau)} \frac{\sigma(\frac{y}{\mu})^{\sigma\nu}}{y[1+(\frac{y}{\mu})^{\sigma}]^{\nu+\tau}}$	4	ambos	ambos	log	identidad	log	log
GG	Generalized Gamma	$z = (\frac{y}{\mu})^{\nu}$, $\theta = \frac{1}{\sigma^2\nu^2}$ $f_Y(y;\mu,\sigma,\nu) = \frac{\lvert\nu\rvert\theta^{\theta}z^{\theta}\exp\{-\theta z\}}{\Gamma(\theta)y}$	3	positiva	-	log	log	identidad	-
GIG	Generalized Inverse Gaussian	$c = [K_{\nu+1}(\frac{1}{\sigma^2})][K_{\nu+1}(\frac{1}{\sigma^2})]^{-1}$ $K_{\lambda}(t) = \frac{1}{2}\int_0^{\infty} x^{\lambda-1}\exp -\frac{1}{2}t(x+x^{-1})\,\mathrm{d}x$ $f_Y(y;\mu,\sigma,\nu) = (\frac{c}{\mu})^{\nu}[\frac{y^{\nu-1}}{2K_{\tau}(\frac{1}{\sigma^2})}]\exp[-\frac{1}{2\sigma^2}(\frac{cy}{\mu} - \frac{\mu}{cy})]$	3	positiva	-	log	log	identidad	-
IG	Inverse Gaussian	$f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2y^3}}\exp[-\frac{1}{2\mu^2\sigma^2y}(y-\mu)^2]$	2	positiva	-	log	log	-	-
LOGNO	Log Normal	$f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{y}\exp[-\frac{(\log(y)-\mu)^2}{2\sigma^2}]$	2	positiva	-	log	log	-	-
WEI	Weibull	$f_Y(y;\mu,\sigma) = \frac{\sigma y^{\sigma-1}}{\mu^{\sigma}}\exp[-(\frac{y}{\mu})^{\sigma}]$	2	positiva	-	log	log	-	-

Tabla 1: Familias GAMLSS con distribución continua definidas en $(0, \infty)$.

En la columna 'Características de forma' de la Tabla 1 se indica como es la asimetría y la curtosis en las distintas distribuciones. Para la asimetría, 'ambos' significa que puede ser positiva o negativa, mientras que para curtosis significa que puede ser tanto leptocúrtica como platicúrtica.

3.2.2. Selección de mejor modelo GAMLSS

En la búsqueda de un modelo GAMLSS apropiado para un conjunto de datos, se deben tomar dos decisiones: especificar la distribución de la variable respuesta y especificar qué términos aparecen en la modelización de sus parámetros.

3.2.2.1 Selección de la distribución

La selección de la distribución que mejor se adapta a los datos puede llevarse a cabo en la etapa de ajuste del modelo GAMLSS. Para ello se comparan diferentes modelos ajustados utilizando medidas de la calidad del ajuste. Concretamente, en este trabajo se ha utilizado el criterio de información de Akaike generalizado (GAIC) [18]. Esta medida de bondad de ajuste tiene la expresión

$$GAIC(k) = -2\sum_{i=1}^{n} \log[f_Y(y_i;\hat{\theta}_i)] + k \cdot gl_{eff}$$

donde $\sum_{i=1}^{n} \log[f_Y(y_i;\hat{\theta}_i)]$ con $\hat{\theta}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$ es la función de log-verosimilitud ajustada, k es un factor de penalización, por defecto, GAIC emplea una penalización $k = 2{,}5$ y gl_{eff} representa el número efectivo de grados de libertad del modelo elegido,

$$gl_{eff} = n - \sum_{\theta \in \{\mu,\sigma,\nu,\tau\}} gl_\theta$$

con gl_θ los grados de libertad utilizados en la modelización del parámetro correspondiente. Atendiendo a este criterio, se selecciona el modelo con menor GAIC.

3.2.2.2 Selección de los términos aditivos del modelo

Los modelos GAMLSS permiten modelizar todos los parámetros de la distribución elegida para la variable respuesta como constantes, como funciones lineales o no-lineales de las variables explicativas. La selección de los términos que aparecen en el modelo debe hacerse para todos los parámetros de la distribución elegida.

Para la selección de los términos aditivos del modelo se emplea el siguiente algoritmo basado en una combinación de los métodos *forward* y *backward* de selección de variables (ver Figura 1).

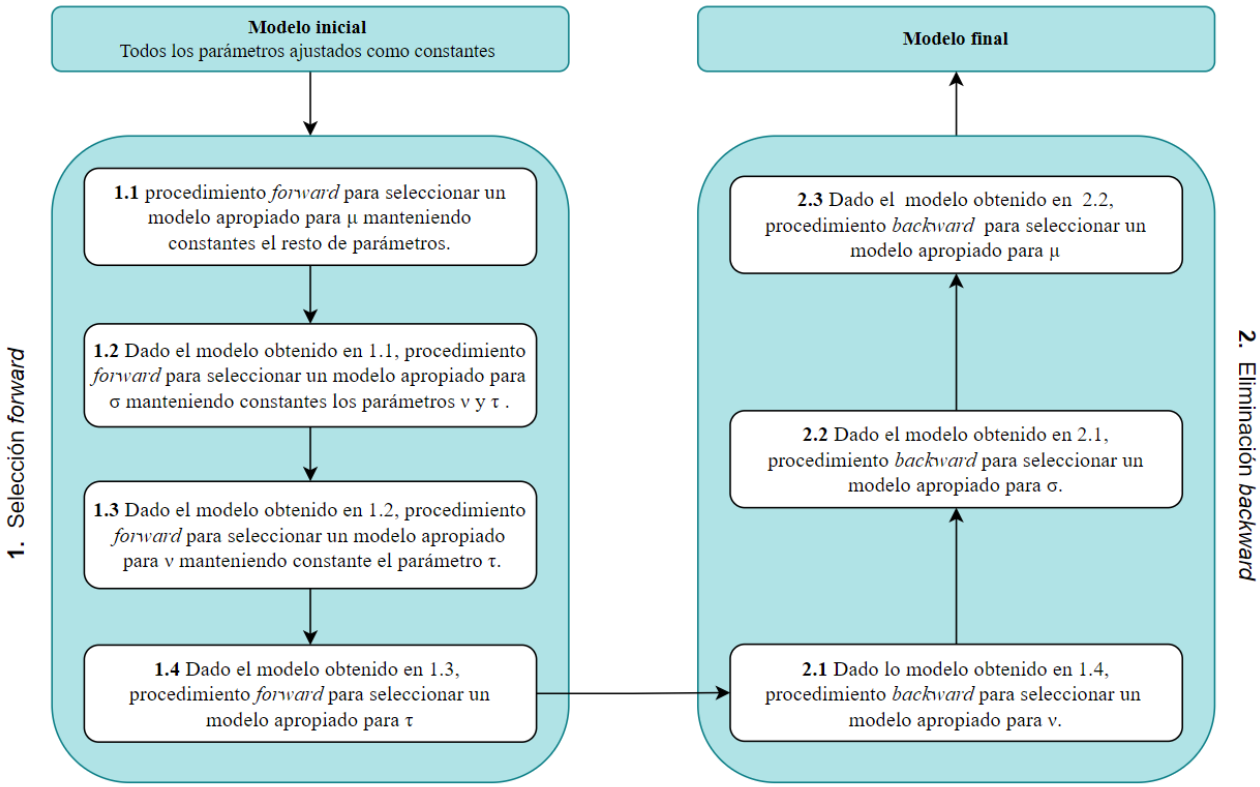

Figura 1: Algoritmo *stepwise* para la selección de las variables explicativas reelevantes para una distribución fija.

La elección entre modelos anidados se lleva a cabo utilizando el test de razón de verosimilitudes, mientras que para modelos no anidados se utiliza el GAIC.

3.2.3. Diagnóstico del modelo

El diagnóstico de los modelos de regresión se lleva a cabo utilizando los residuos crudos $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$ y sus valores estandarizados. Sin embargo, este tipo de residuos es difícil de utilizar en distribuciones diferentes a la distribución Normal. Por eso es habitual que en aproximaciones como los GLM, se utilicen otros tipos de residuos como pueden ser los residuos de desviación, los residuos por exclusión y/o los residuos de Pearson [15]. Sin embargo, estos tipos de residuos tampoco son directamente aplicables para los modelos GAMLSS, bien porque no están bien definidos para múltiples parámetros de la distribución de la variable respuesta, o porque no son apropiados para datos muy asimétricos y/o apuntados. En su lugar, en el contexto de los GAMLSS, se utilizan los residuos cuantil normalizados [19]. Este tipo de residuos se caracterizan por tener una distribución Normal estándar bajo la asunción de que el modelo es correcto, sea cual sea la distribución de la variable respuesta. De esta forma, el diagnóstico del modelo GAMLSS ajustado se reduce a verificar la normalidad de los residuos cuantil normalizados.

Los residuos cuantil normalizados se computan, para $i = 1, \ldots, n$, como

$$r_i = \Phi^{-1}(\hat{u}_i)$$

donde $\Phi^{-1}(\cdot)$ es la inversa de la función de distribución de una Normal estándar y $\hat{u}_i = F_Y(y_i; \hat{\theta}_i)$ donde $F_Y(y_i; \theta_i)$ es la función de distribución para la observación i-ésima de una variable respuesta continua Y. Si el modelo es correcto, u tiene una distribución uniforme entre 0 y 1 y $r = \Phi^{-1}(u)$ tendrá una distribución Normal estándar.

Para evaluar la normalidad de los r_i y poder inferir que el modelo es válido, en este trabajo se utilizan, como herramientas de diagnóstico básicas el histograma, el *QQ-plot* de los residuos, y el gráfico de residuales. En los dos primeros gráficos se debe observar que la distribución es aproximadamente Normal, mientras que los *plots* de residuales deberían ser *plots* nulos, sin observar patrones sistemáticos. Además, se presentan algunas medidas descriptivas como la media de los residuos, que debería ser aproximadamente 0; la varianza, aproximadamente 1; el coeficiente de simetría, cercano a 0; y el coeficiente de curtosis, cercano a 3.

Otra herramienta gráfica que se utiliza en este tipo de modelos es el *worm plot* [20], que basicamente son *QQ-plots* sin tendencia. La interpretación de los distintos elementos de este gráfico es la siguiente

- Los puntos muestran lo alejado que están los residuos de sus valores esperados representados por una línea horizontal.
- Las curvas elípticas representan regiones de confianza del 95 % para la *deviance*. Si el modelo es correcto aproximadamente el 95 % de los puntos deberían estar entre las dos curvas en la parte central del gráfico.
- El modelo cúbico ajustado a los datos. La forma de esta curva nos da información de diferentes desajustes del modelo según los criterios resumidos en la Tabla 2.

Forma de la curva	**Residuos**	**Desajuste**
Nivel: sobre el origen	Media muy grande	Parámetro de localización muy bajo
Nivel: debajo del origen	Media muy pequeña	Parámetro de localización muy alto
Pendiente positiva	Varianza muy grande	Parámetro de escala muy bajo
Pendiente negativa	Varianza muy pequeña	Parámetro de escala muy alto
Patrón de U	Simetría positiva	Parámetro de asimetría muy bajo
Patrón de U invertida	Asimetría negativa	Parámetro de asimetría muy alto
Patrón de S, cola izquierda por debajo	Leptocúrticos	Parámetro de curtosis muy bajo
Patrón de S, cola izquierda por arriba	Platicúricos	Parámetro de curtosis muy alto

Tabla 2: Criterios de desajuste y posible solución de los modelos GAMLSS ajustados observables en el *worm plot* [20].

3.2.4. Estimación de los percentiles

Los IRs se construyen a partir de los modelos GAMLSS ajustados mediante la estimación de los percentiles.

Denotamos como y_α al percentil 100α de una variable Y, es decir, que verifica que $P(Y \leq y_\alpha) = \alpha$. Se tiene entonces que $y_\alpha(x) = F^{-1}_{Y|x}(\alpha)$, donde $F^{-1}_{Y|x}(\cdot)$ es la inversa de la función de distribución de Y condicionada a los valores de los regresores. En este trabajo se han obtenido los percentiles $100\alpha = (2.5, 5, 95, 97.5)$ considerando como variables explicativas edad, sexo y raza. Puesto que sexo y raza son variables cualitativas, esto se reduce a obtener los percentiles a partir de la distribución de Y condicionada por la edad para cada grupo de sexo y raza.

Resultados

Los datos que se van a analizar en este trabajo se han obtenido de la encuesta *The National Health and Nutrition Examination Survey* (NHANES). Esta encuesta se realiza anualmente, desde 1971, por el US *National Center for Health Statistics* (NCHS), y su objetivo es evaluar el estado nutricional y de salud general de población americana, tanto adulta como infantil.

Contiene información procedente, no sólo de diversos cuestionarios, sino también de exámenes físicos entre los que se encuentran análiticas de sangre y de orina.

La información obtenida se utiliza por el *Centre for Disease Control and Prevention* (CDC) [21] para determinar la prevalencia de enfermedades importantes, así como de sus factores de riesgo, siendo una herramienta muy útil para la promoción de la salud y la prevención de enfermedades.

Desde 1999 los datos son públicos y están disponibles en la dirección web:

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes.

Concretamente, en este trabajo los datos se corresponden con la encuesta llevada a cabo en el periodo 2017-2018, los más recientes disponibles gratuitamente. Se procede a la aplicación de los modelos GAMLSS para la construcción de IRs para los valores de creatinina medidos en orina, utilizando como posibles variables regresoras la edad, el sexo y la raza.

La creatinina es un producto de desecho que se genera por los músculos y cuyos niveles, cuando son anormalmente altos en sangre o anormalmente bajos en orina, se utilizan como marcadores de enfermedad renal.

La medición de este producto suele formar parte de chequeos médicos rutinarios, por lo que disponer de IRs es muy útil en la práctica clínica. Además, es especialmente interesante en este caso, puesto que se conoce que sus niveles están influenciados por una serie de factores, entre los que se encuentran la edad, el sexo y la raza [22].

El código R utilizado en este trabajo se encuentra en el anexo D.

4.1. Análisis descriptivo

La muestra que se ha utilizado en este trabajo tiene un tamaño total de 7626 individuos, estratificados en dos grupos: 2499 individuos en edad pediátrica (entre 3 y 19 años) y 5127 individuos en edad adulta (>19 años). La muestra total se divide aleatoriamente en dos submuestras:

- La muestra de referencia, formada por el 90 % de los individuos, que se utiliza para ajustar los modelos GAMLSS y estimar los IRs.
- La muestra de validación, formada por el 10 % de los individuos y que, puede utilizarse para validar los IRs.

En la Tabla 3 se comparan las muestras de referencia y validación para todas las variables consideradas en este trabajo. Podemos observar que no hay diferencias significativas entre las dos muestras en ninguno de los estratos, por tanto, podemos asumir que ambas proceden de la misma población por lo que la muestra de validación es apropiada para validar los IRs .

Estrato	**Variable**	**Nivel**	**Muestra de referencia**		**Muestra de validación**		**P-valor**
			n	Media±DT %(IC95 %)	n	Media±DT %(IC95 %)	
Edad Pediátrica (3-19 años)	Creatinina		2249	119.29±82.00	250	119.53±81.97	>0.05
	Edad (meses)		2249	136.32±56.28	250	138.84±59.4	>0.05
	Sexo	Hombres	1130	50.34(48.15,52.33)	116	46.4(40.12,52.78)	>0.05
		Mujeres	1119	49.66(47.75,51.84)	134	53.6(47.21,59.87)	>0.05
	Raza	Afroamericano	512	22.76(21.05,24.56)	56	22.4(17.49,28.17)	>0.05
		Asiático	249	11.07(9.81,12.45)	26	10.4(7.03,15.03)	>0.05
		Caucásico	692	30.76(28.87,32.73)	72	28.8(23.35,34.9)	>0.05
		Hispano	383	17.02(15.51,18.66)	57	22.8(17.85,28.6)	>0.05
		Multirracial	231	10.27(9.06,11.61)	20	8(5.07,12.26)	>0.05
		Nativo americano	182	8.09(7.01,9.31)	19	7.6(4.76,11.79)	>0.05
Edad Adulta (>19 años)	Creatinina		4614	127.57±85.72	513	134.02±87.62	>0.05
	Edad (años)		4614	51.39±17.60	513	50.12±17.49	>0.05
	Sexo	Hombres	2208	47.85(46.4,49.3)	267	52.05(47.62,56.43)	>0.05
		Mujeres	2406	52.14(50.69,53.59)	246	47.95(43.56,52.37)	>0.05
	Raza	Afroamericano	1067	23.12(21.92,24.37)	133	25.92(22.23,29.98)	>0.05
		Asiático	664	14.39(13.39,15.44)	72	14.03(11.2,17.41)	>0.05
		Caucásico	1587	34.39(33.02,35.78)	171	33.33(29.29,37.62)	>0.05
		Hispano	616	13.35(12.38,14.37)	74	14.42(11.56,17.83)	>0.05
		Multirracial	238	5.15(4.54,5.84)	19	3.7(2.3,5.82)	>0.05
		Nativo americano	442	9.57(8.75,10.47)	44	8.57(6.36,11.42)	>0.05

Tabla 3: Medidas descriptivas para las variables de interés según el estrato y la submuestra: referencia y validación.

DT: Desviación típica; IC: Intervalo de confianza

En relación a los niveles de creatinina para la muestra de referencia en la Tabla 3, se puede observar que los valores de creatinina generalmente son mayores en el estrato de edad adulta. La Tabla 4 muestra los niveles de creatinina según los factores considerados en la muestra de referencia. Sin tener en cuenta la edad, los hombres tienen valores más altos de esta sustancia, al igual que las personas afroamericanas seguidas de las clasificadas en el grupo multirracial, mientras que las personas asiáticas son las que presentan valores más bajos de creatinina en orina.

Factor	Nivel	Edad pediátrica (0-19 años)		Edad adulta (>19 años)	
		n	Media±DT	n	Media±DT
Sexo	Hombres	1130	125.30±82.27	2208	145.15±87.39
	Mujeres	1119	113.21±81.33	2406	111.42±80.87
Raza	Afroamericano	512	150.08±90.25	1067	163.87±100.18
	Asiático	249	97.14±71.78	664	96.16±68.16
	Caucásico	692	110.82±81.20	1587	118.49±77.27
	Hispano	383	110.23±72.12	616	121.78±75.76
	Multirracial	231	118.62±77.11	238	138.53±87.99
	Nativo americano	182	115.07±77.06	442	121.85±82.06

Tabla 4: Niveles de creatinina según el estrato y los factores de interés en la muestra de referencia.

DT: Desviación típica

Comparando ambos estratos de forma gráfica (Figuras 2 y 3), llama la atención cómo en el estrato de edad pediátrica aumentan los valores de creatinina con la edad. Este patrón no se observa en los adultos, donde los niveles tienden a ser menores a medida que aumenta la edad. Para más gráficos sobre el análisis descriptivo, consultar el anexo A.

Figura 2: Niveles de creatinina en la muestra de referencia para el estrato de edad pediátrica según la edad, el sexo y la raza.

Figura 3: Niveles de creatinina en la muestra de referencia para el estrato de edad adulta según la edad, el sexo y la raza.

4.2. Selección de los modelos GAMLSS

Se selecciona el mejor modelo para cada uno de los estratos según la metodología detallada en la Sección 3. Los modelos se ajustan utilizando la muestra de referencia considerando como variable respuesta los niveles de creatinina y como posibles regresores la edad, el sexo y la raza. Las distribuciones que se comparan son las que se detallan en la Tabla 1 (Sección 3.2.1), definidas en el soporte $(0, \infty)$.

En la Tabla 5 se resumen las características de los modelos GAMLSS ajustados para los dos estratos:

- Para la edad pediátrica, el modelo con menor GAIC se corresponde con la distribución Weibull (WEI) en el que se modelizan los parámetros de localización y escala en términos de los tres factores considerados. La edad se incluye en el modelo como un *spline* cúbico.
  Con valor de GAIC muy próximo se encuentra la distribución gamma generalizada (GG), que incluye un parámetro de asimetría que no depende de las variables explicativas.
- Para la edad adulta, el modelo con menor GAIC se corresponde con la distribución Box-Cox Power Exponential (BCPE) en la que se modelizan los parámetros de localización, escala, asimetría y curtosis. Los dos primeros en función de los tres factores considerados con la edad incluida como un *spline* cúbico; el tercero en función de la edad y el sexo, y el último independiente de las variables explicativas.
  Con un valor de GAIC muy próximo (su diferencia es inferior a la unidad), se encuentra la distribución gamma generalizada (GG), que excluye de la anterior el parámetro de curtosis.

Familia	Edad pediátrica		Edad adulta	
	Modelo	GAIC	Modelo	GAIC
BCCG	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + sexo$ $\nu \sim 1$	24734.8	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + sexo$ $\nu \sim edad + sexo$	51820.2
BCPE	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + sexo$ $\nu \sim 1$ $\tau \sim 1$	24720.7	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$ $\nu \sim edad + sexo$ $\tau \sim 1$	51798.3
BCT	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + sexo$ $\nu \sim 1$ $\tau \sim 1$	24765.1	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$ $\nu \sim edad + sexo$ $\tau \sim 1$	51816.2
GA	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$	24730.1	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim sexo + cs(edad) + raza$	51815
GB2	$\mu \sim cs(edad) + raza$ $\sigma \sim cs(edad) + raza + sexo$ $\nu \sim 1$ $\tau \sim 1$	24748.7	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$ $\nu \sim edad + sexo$ $\tau \sim 1$	51859.6
GG	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$ $\nu \sim 1$	24704.8	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$ $\nu \sim edad + sexo$	51799.6
GIG	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim 1$ $\nu \sim raza + sexo$	24735.7	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim sexo$ $\nu \sim cs(edad) + raza + sexo$	51806.6
IG	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$	25041.5	$\mu \sim edad + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$	52390.9
LOGNO	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$	25041.5	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + sexo$	52162.9
WEI	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$	24702.1	$\mu \sim cs(edad) + raza + sexo$ $\sigma \sim cs(edad) + raza + sexo$	51857.6

Tabla 5: Ajuste de los modelos GAMLSS en la muestra de referencia para las distintas distribuciones de la variable respuesta consideradas.

4.3. Diagnóstico de los modelos seleccionados

Se estudia la normalidad de los residuos cuantil normalizados para evaluar la validez de los dos mejores modelos.

4.3.1. Estrato 1: edad pediátrica

En la Tabla 6 se muestran los descriptivos de estos residuos.

Estos estadísticos soportan la asunción de que su distribución sea aproximadamente Normal estándar en ambos modelos: media próxima a 0, varianza a 1, y los coefiecientes de asimetría y curtosis cercanos a 0 y 3, respectivamente.

	WEI	GG
Media	-0.001041406	-0.001418982
Varianza	0.9887786	0.997616
Coef. asimetría	0.1010494	0.03231352
coef. curtosis	2.630195	2.59675

Tabla 6: Resumen de la distribución delos residuos cuantiles en el estrato de edad pediátrica en la muestra de referencia.

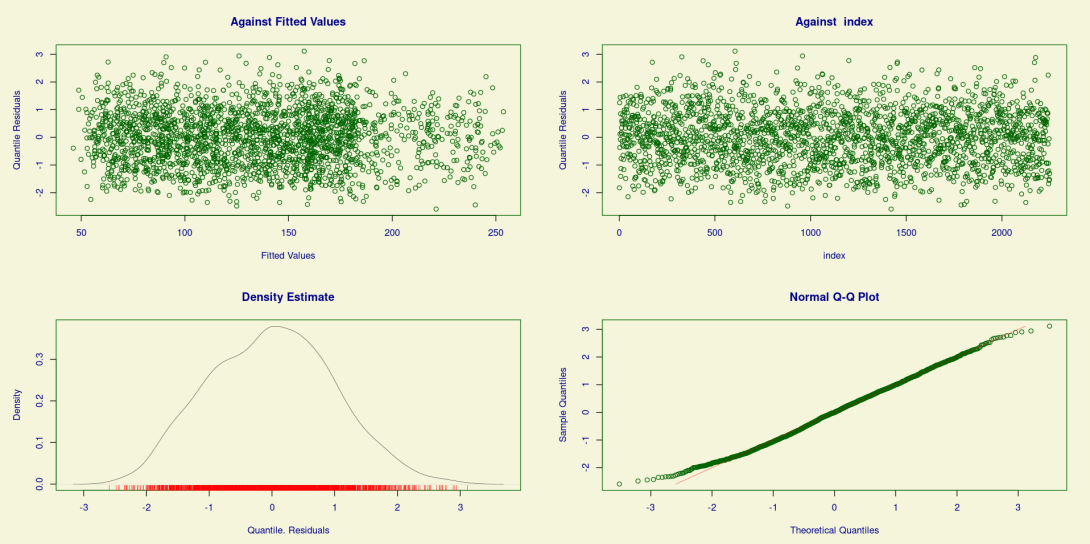

Figura 4: *Plots* de residuales del modelo basado en la distribución WEI para los niveles de creatinina en el estrato de edad pediátrica en la muestra de referencia.

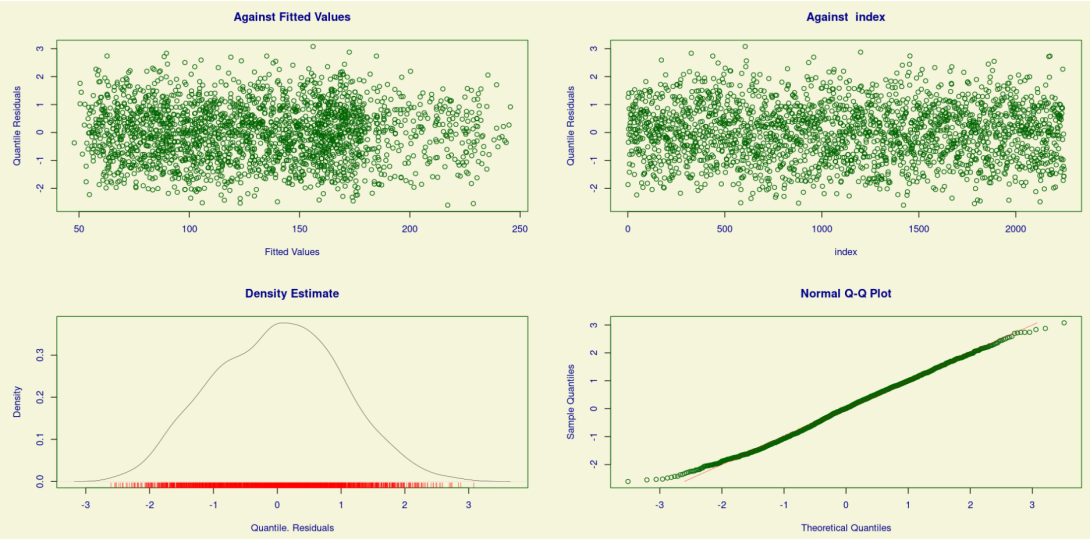

Figura 5: *Plots* de residuales del modelo basado en la distribución GG para los niveles de creatinina en el estrato de edad pediátrica en la muestra de referencia.

Como se puede observar en las Figuras 4 y 5, los residuos cuantil normalizados se comportan bien. El plot de densidad y el *QQ-plot* (colocados en la parte inferior de las figuras) no muestran desviaciones de la asunción de normalidad. Además, los *plots* de residuales

(colocados en la parte superior) muestran un patrón aleatorio alrededor de la recta $Y = 0$.

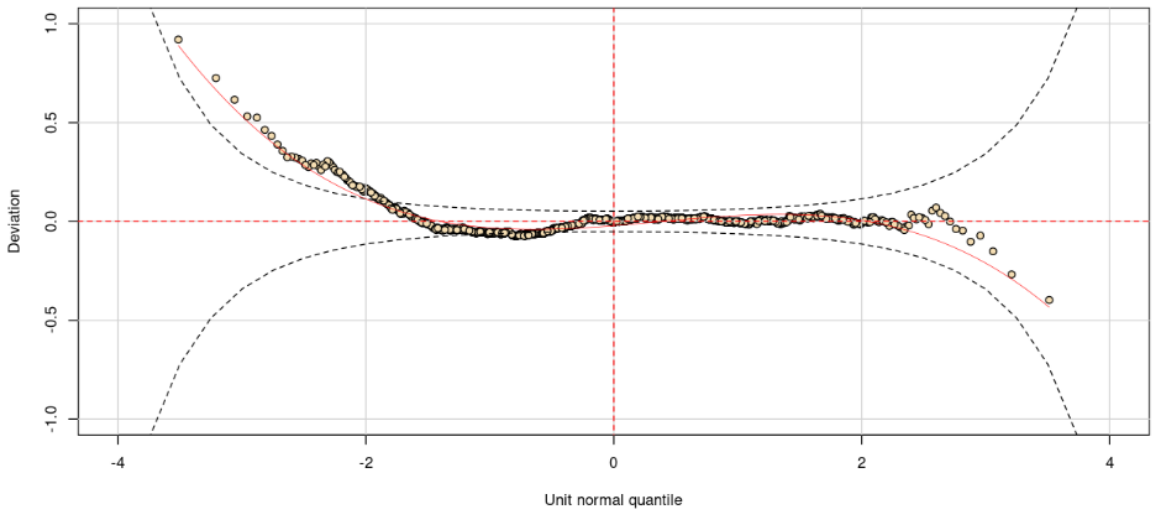

Figura 6: *Worm plot* del modelo basado en la distribución WEI para los niveles de creatinina en el estrato de edad pediátrica en la muestra de referencia.

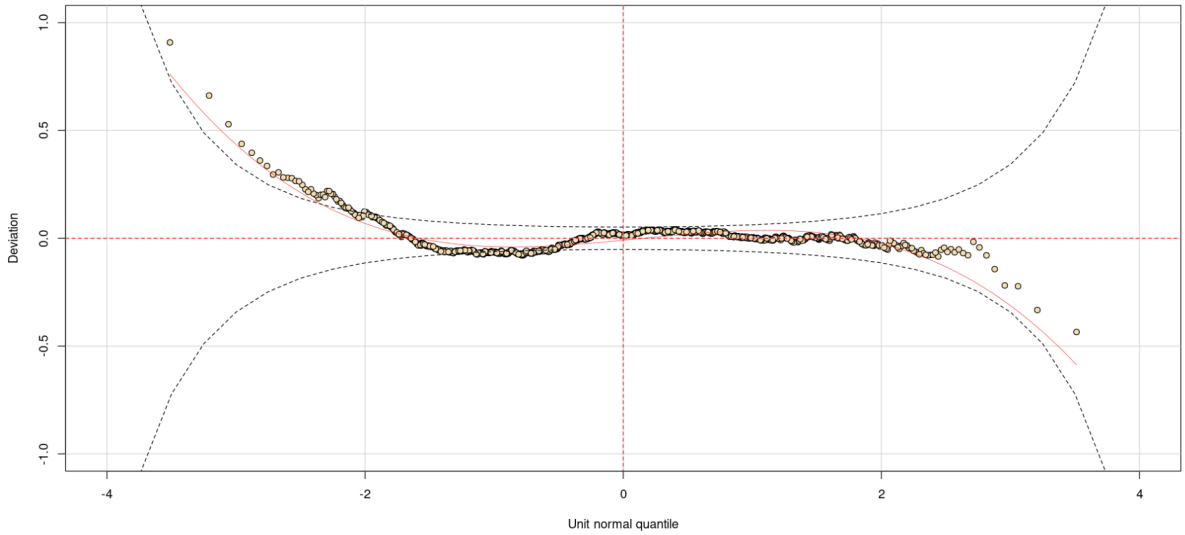

Figura 7: *Worm plot* del modelo basado en la distribución GG para los niveles de creatinina en el estrato de edad pediátrica en la muestra de referencia.

Como se puede observar en las Figuras 6 y 7, ambos modelos se ajustan adecuadamente ya que aproximadamente el 95 % de los puntos se encuentra dentro de las regiones de confianza en la parte central del gráfico.

Teniendo en cuenta que los dos modelos potenciales son válidos y prácticamente igual de buenos, se decide, por simplicidad, el modelo basado en la distribución WEI para la construcción de los IRs.

4.3.2. Estrato 2: edad adulta

Al igual que en el estrato de edad pediátrica se estudia la normalidad de los residuos cuantil normalizados para evaluar la validez de los dos mejores modelos elegidos en la sección anterior.

En la Tabla 7 se muestran los descriptivos de estos residuos. Estos estadísticos validan la asunción de que su distribución sea aproximadamente Normal en ambos modelos.

	BCPE	GG
Media	0.006628771	-0.0001535682
Varianza	00.9999853	1.000347
coef. asimetría	-0.01295024	0.0101127
coef. curtosis	3.008499	2.874897

Tabla 7: Resumen de la distribución de los residuos cuantiles en el estrato de edad adulta en la muestra de referencia.

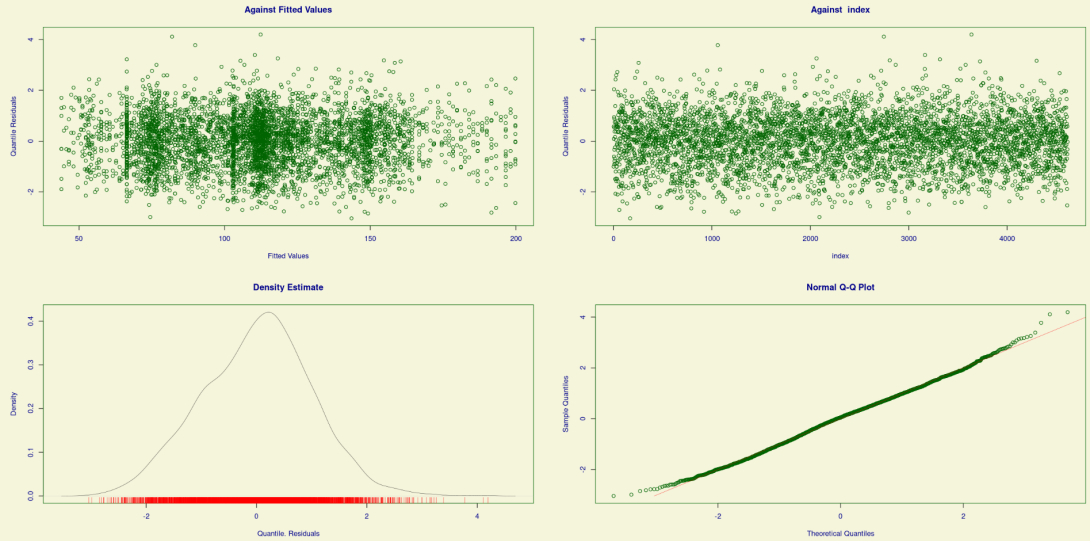

Figura 8: *Plots* de residuales del modelo basado en la distribución BCPE para los niveles de creatinina en el estrato de edad adulta en la muestra de referencia.

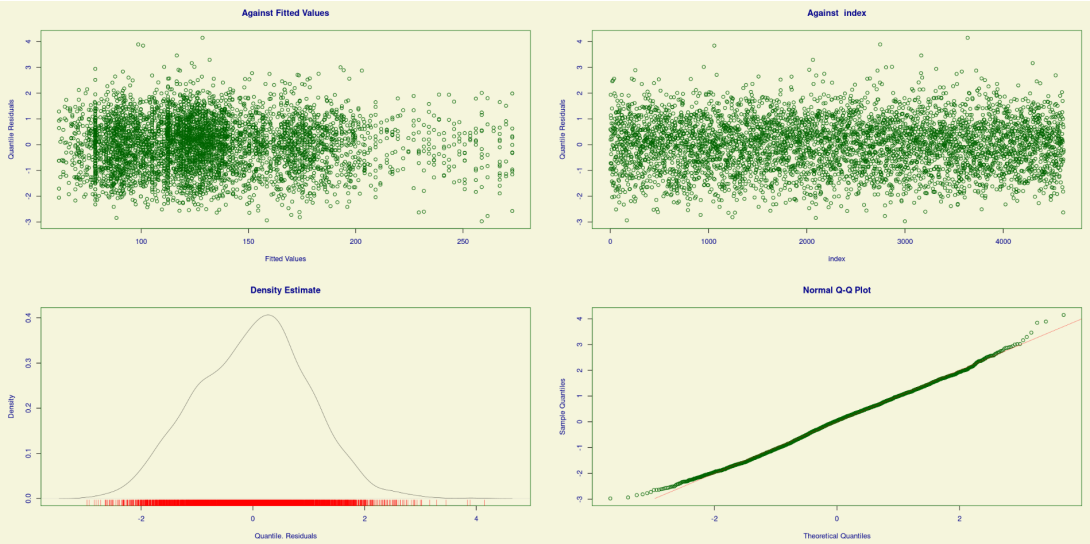

Figura 9: *Plots* de residuales del modelo basado en la distribución GG para los niveles de creatinina en el estrato de edad adulta en la muestra de referencia.

Como se puede observar en las Figuras 8 y 9, los residuos normalizados se comportan bien. Y al igual que ocurría en el estrato de dad pediátrica, ninguno de los gráficos, *plots* de residuales, el de densidad y *QQ-Plot*, muestran desviaciones importantes de la asunción de normalidad. Lo mismo ocurre con los *worm plot* de las Figuras 10 y 11 .

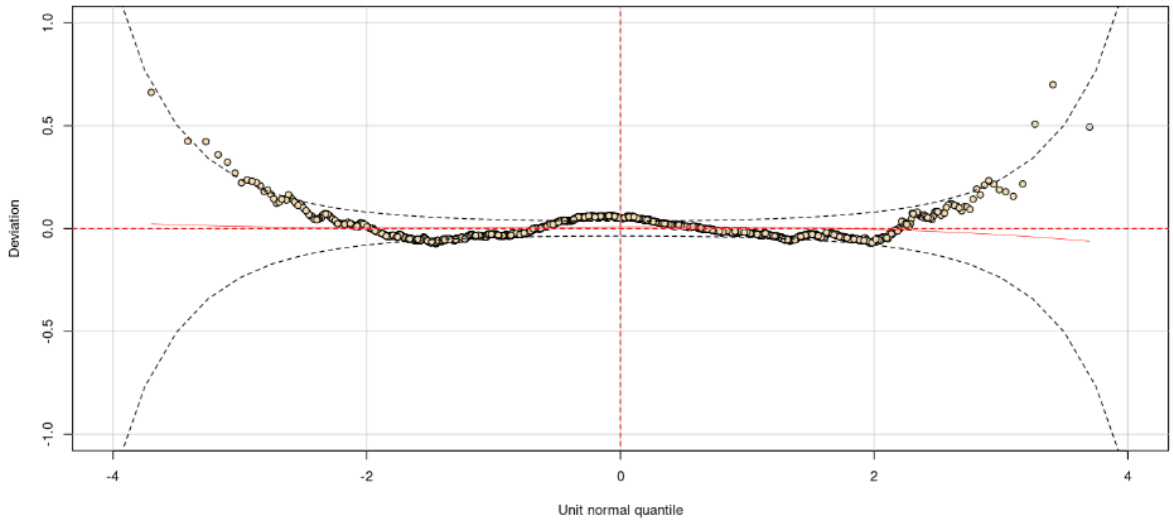

Figura 10: *Worm plot* del modelo basado en la distribución BCPE para los niveles de creatinina en el estrato de edad adulta en la muestra de referencia.

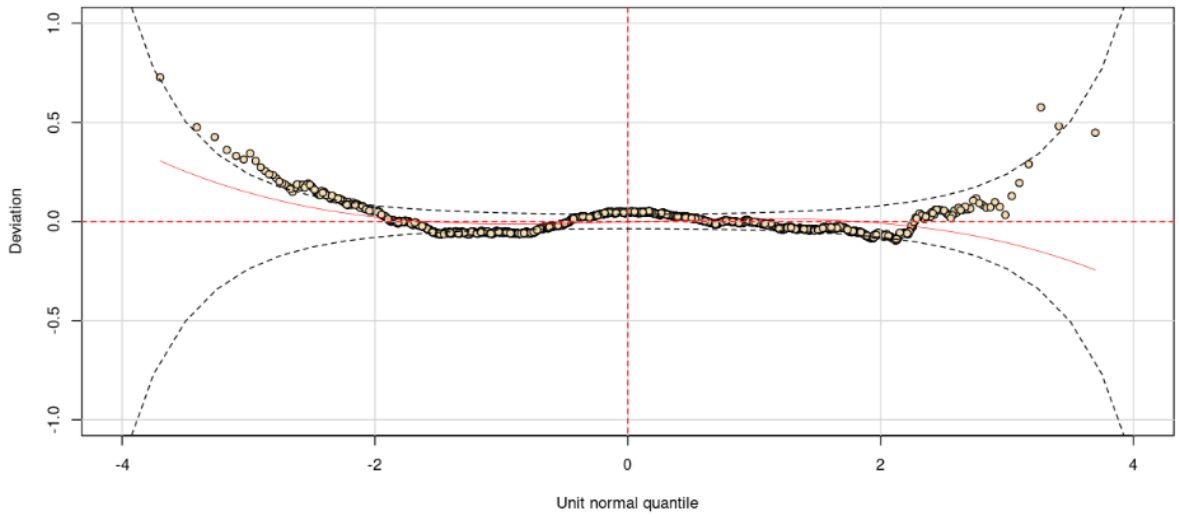

Figura 11: *Worm plot* del modelo basado en la distribución GG para los niveles de creatinina en el estrato de edad adulta en la muestra de referencia.

Utilizando el principio de simplicidad, en este caso se selecciona el modelo basado en la distribución GG.

4.4. Estimación de los parámetros de los modelos GAMLSS

En las Tablas 8 y 9 se muestra la estimación de los parámetros ajustados para el mejor modelo en cada estrato. Se ha escogido como categoría de referencia la que presenta niveles más altos de creatinina: hombre para el sexo y afroaméricano para la raza.

Parámetro	Variable	Estimación	Error Estándar	t Valor	P-valor
μ	Intercept	4.3252	0.0392	110.060	< 0.05
	cs(edad)	0.0058	0.0002	25.557	< 0.05
	Sexo: mujer	-0.0903	0.0241	-3.746	0.0001
	Raza: asiático	-0.4324	0.0466	-9.279	< 0.05
	Raza: caucásico	-0.3076	0.0317	-9.698	< 0.05
	Raza: hispano	-0.3202	0.0366	-8.746	< 0.05
	Raza: multirracial	-0.2055	0.0429	-4.791	< 0.05
	Raza: nativo americano	-0.2786	0.0483	-5.762	< 0.05
σ	Intercept	0.8157	0.0530	15.366	< 0.05
	cs(edad)	-0.0001	0.0002	-0.631	0.5278
	Sexo: mujer	-0.1412	0.0331	-4.261	< 0.05
	Raza: asiático	-0.2322	0.0609	-3.809	< 0.05
	Raza: caucásico	-0.1436	0.0459	-3.125	< 0.05
	Raza: hispano	-0.0971	0.0529	-1.836	0.0665
	Raza: multirracial	-0.0945	0.0624	-1.515	0.1298
	Raza: nativo americano	-0.1331	0.0668	-1.992	< 0.05

Tabla 8: Coeficientes estimados para los parámetros del modelo basado en la distribución WEI en el estrato de edad pediátrica en la muestra de referencia.

Parámetro	Variable	Estimación	Error Estándar	t Valor	P-valor
μ	(Intercept)	5.7988	0.0479	121.037	< 0.05
	cs(edad)	-0.0100	0.0008	-12.482	< 0.05
	Sexo: mujer	-0.3587	0.0295	-12.122	< 0.05
	Raza: asiático	-0.5334	0.0308	-17.299	< 0.05
	Raza: caucásico	-0.2907	0.0235	-12.353	< 0.05
	Raza: hispano	-0.3180	0.0295	-10.773	< 0.05
	Raza: multirracial	-0.2121	0.0420	-5.048	< 0.05
	Raza: nativo americano	-0.2832	0.0341	-8.299	< 0.05
σ	(Intercept)	-0.6194	0.0454	-13.622	< 0.05
	cs(edad)	-0.0003	0.0007	-0.553	0.5801
	Sexo: mujer	0.1802	0.0255	7.044	< 0.05
	Raza: asiático	0.1084	0.0323	3.349	< 0.05
	Raza: caucásico	0.0636	0.0261	2.436	< 0.05
	Raza: hispano	0.0108	0.0334	0.324	0.7455
	Raza: multirracial	0.0281	0.0471	0.597	0.5505
	Raza: nativo americano	0.0585	0.0371	1.575	0.1154
ν	(Intercept)	1.8185	0.2350	7.735	< 0.05
	Edad	-0.0072	0.0039	-1.830	0.0673
	Sexo: mujer	-0.5010	0.1421	-3.525	< 0.05

Tabla 9: Coeficientes estimados para los parámetros del modelo basado en la distribución GG en el estrato de edad adulta en la muestra de referencia.

En el estrato de edad pediátrica (Tabla 8), pediátrica podemos observar, en el parámetro de localización, que las mujeres tienen un valor diferencial de -0.0903 unidades, es decir, de 0.0903 mg/dL inferior en creatinina respecto de los hombres de su misma edad y raza. Fijándonos en los valores diferenciales de los distintos grupos étnicos, llama la atención que los afroamericanos tienen niveles significativamente mayores que el resto de grupos raciales considerados, especialmente ésta diferencia es muy significativa con el grupo de asiáticos.

En lo que se refiere al parámetro de variabilidad, las mujeres presentan una variabilidad significativamente menor que los hombres. En cuanto a la raza, asiáticos, caucásicos y nativos americanos presentan una variabilidad en el nivel de creatinina significativamente menor que el resto.

En la Tabla 9 se representan las estimaciones de los coeficientes del modelo ajustado para el estrato de edad adulta. En el parámetro μ resulta interesante ver que la diferencia entre los niveles de creatinina para personas de la misma edad y raza varía de manera más importante que en el estrato de edad pediátrica, con un valor diferencial del -0.3587 unidades.

En cuanto a σ, la variabilidad va disminuyendo a medida que aumenta la edad. También podemos observar dos grupos en relación con la raza, los hispanos, multirraciales y nativos americanos no presentan diferencias significativas en la variabilidad con los afroamericanos, mientras que las otras dos razas restantes si las presentan.

Por último, en relación con la asimetría podemos ver como ésta es menor en las mujeres que en los hombres de la misma edad. Por otra parte, dado que el signo de la estimación de ν para la edad es negativo, sabemos que la asimetría disminuye según va aumentando la edad.

4.5. Estimación de los IRs

En las Figuras 12 y 13 se representan los IRs para el estrato de edad pediátrica y adulta, respectivamente. Las observaciones representadas en estas figuras se corresponden con los individuos de la muestra de validación.

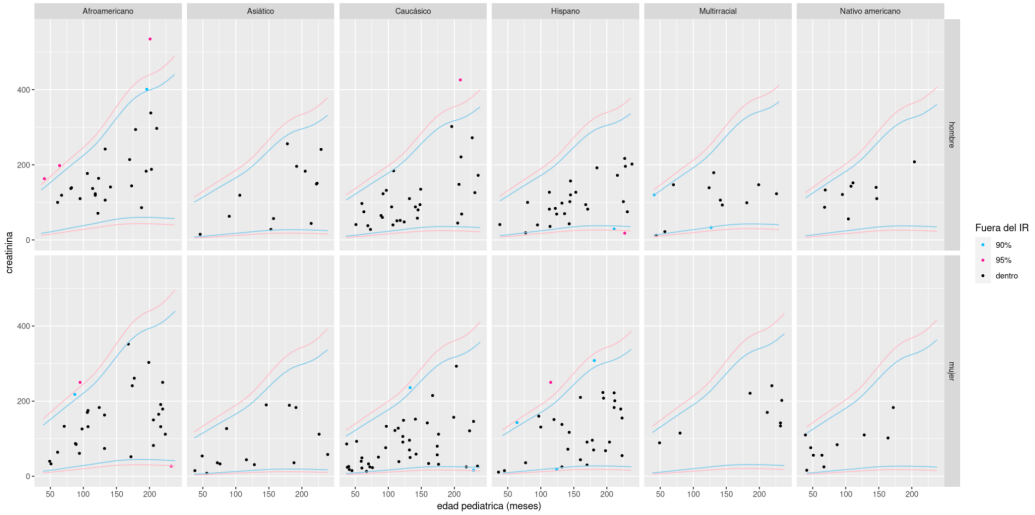

Figura 12: IRs para el modelo basado en la distribución WEI en el estrato de edad pediátrica estimados con la muestra de referencia.

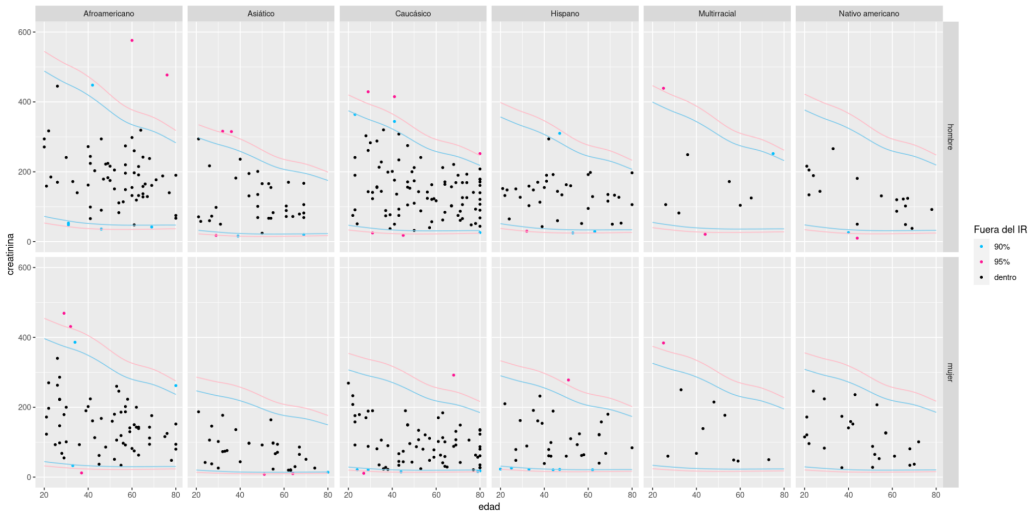

Figura 13: IRs para el modelo basado en la distribución GG en el estrato de edad adulta estimados con la muestra de referencia.

En ambos casos se observa una clara dependencia de la edad, aunque el tipo de dependencia es diferente en cada estrato.

En edad adulta los niveles de creatinina disminuyen con la edad y la variabilidad es menor en las mujeres que en los hombres.

En cuanto a la raza podemos ver cómo los valores más altos de creatinina se encuentran en los afroamericanos en ambos estratos.

Para ver los IRs construidos a partir de los segundos mejores modelos en cada estrato, consultar el anexo B.

4.6. Validación de los IRs

Para validar los IRs estimados en la sección anterior, se va a estudiar que porcentaje de individuos quedan dentro de los IRs del 90 % y al 95 % de confianza. Se espera que, como mucho, el 5 % y el 10 % de las observaciones de la muestra de validación quedan fuera del IR correspondiente. La Tabla 10 muestra el porcentaje de individuos de la muestra de validación clasificados como "normales" por los IRs junto con sus IC, según su estrato, sexo y raza.

			GAMLSS (dentro del IR)					
			IR 95 %			**IR 90 %**		
Estrato	**Factor**	**Nivel**	n	%	IC95 %	n	%	IC95 %
Edad pediátrica	Sexo	Hombre	111	95.68	(89.17,100)	107	92.24	(85.72,98.53)
		Mujer	131	97.76	(91.25,100)	125	93.28	(86.77,99.79)
	Raza	Afroamericano	51	91.07	(88.22,100)	49	87.50	(85.16,94.15)
		Asiático	26	100	(97.34,100)	26	100	(97.58,100)
		Caucásico	71	98.61	(93.76,100)	69	95.83	(93.02,100)
		Hispano	55	96.49	(91.71,100)	51	89.47	(85.94,93.97)
		Multirracial	20	100	(97.24,100)	18	90.00	(87.79,95.18)
		Nativo americano	19	100	(96.91,100)	19	100	(98.13,100)
		Total	242	96.80	(87.13,100)	232	92.85	(86.84,100)
Edad adulta	Sexo	Hombre	253	94.75	(92.11,97.12)	238	89.13	(85.06,94.37)
		Mujer	237	96.34	(93.87,99.88)	222	90.24	(86.62,96.93)
	Raza	Afroamericano	128	96.24	(94.33,99.12)	120	90.22	(89.18,96.39)
		Asiático	67	93.05	(90.82,97.10)	64	88.88	(86.95,92.49)
		Caucásico	164	95.90	(93.48,98.86)	156	91.22	(89.63,97.46)
		Hispano	72	97.29	(94.74,100)	63	85.13	(83.75,90.25)
		Multirracial	16	84.21	(82.94,86.36)	15	78.94	(77.90,91.44)
		Nativo americano	43	97.72	(95.47,100)	42	95.45	(93.73,99.23)
		Total	490	95.5	(90.40,100)	460	89.62	(86.84,100)

Tabla 10: Porcentaje de individuos de la muestra de validación clasificados como "normales" según su estrato.

IR: Intervalo de referencia; IC: Intervalo de confianza;

En la Tabla 10 podemos observar que la mayoría de los individuos son clasificados como "normales".

Podemos ver como hay algunos porcentajes que están ligeramente por debajo de lo esperado, pero no son estadísticamente distintos.

Es el caso de los afroamericanos en el estrato de edad pediátrica, y de los asiáticos y multirraciales en el estrato de edad adulta. Esto se debe a que la muestra es muy pequeña y como podemos ver en las Figuras 12 y 13, los individuos que se encuentran fuera de los IRs para estos grupos raciales se encuentran muy cerca del límite del intervalo.

Discusión y conclusiones

La creatinina es un producto de desecho producido por los músculos a partir de la creatina, una sustancia que se utiliza para generar energía en los músculos. La creatinina se filtra en los riñones y se elimina del cuerpo a través de la orina. Debido a que su producción depende de la masa muscular, los niveles de creatinina en el cuerpo pueden variar según la edad, el sexo y la raza [1].

En general, los hombres tienden a tener niveles más altos de creatinina que las mujeres debido a su mayor masa muscular. Además, los niveles de creatinina tienden a disminuir con la edad, ya que se produce una pérdida gradual de masa muscular.

Algunos estudios han encontrado diferencias significativas en los niveles de creatinina entre diferentes grupos étnicos. Por ejemplo, los afroamericanos pueden tener niveles más altos de creatinina que los caucásicos debido a una mayor masa muscular y una tasa de filtración glomerular más baja, véase [23].

La medición de los niveles de creatinina en sangre o en orina son un indicador importante de la función renal, si estos son elevados en sangre o bajos en el caso de la orina pueden ser un signo de nefropatía o alguna otra enfermedad que afecte al funcionamiento de los riñones como: enfermedados autoinmunitarias, infecciones bacterianas, bloqueo en las vias urinarias o complicaciones de la diabetes.

Los datos utilizados en este trabajo apoyan estas teorías. Además de la clara dependencia del sexo y la raza, se ha encontrado un efecto importante de la edad. Este efecto no es el mismo en cada uno de los dos estratos considerados, algo que parece lógico, dado que los individuos en el estrato de edad pediátrica estarán en una etapa de crecimiento, lo que indudablemente tendrá un efecto en su masa muscular.

En este trabajo, se han establecido IRs ajustados por edad, sexo y raza para el nivel de creatinina en orina utilizando los modelos GAMLSS. La importancia de manejar unos buenos IRs está fuera de toda duda, puesto que constituyen una herramienta muy útil para el diagnóstico de muchas enfermedades en estudios epidemiológicos, estudios clínicos y la práctica clínica. La definición de IRs más utilizada es el intervalo de valores que contienen el 90 % o 95 % de individuos de una población de referencia [24]. Una posible estimación de estos intervalos es calcular los percentiles correspondientes, directamente a partir de la distribución de nuestros datos. Sin embargo, este método va a producir estimaciones sesgadas, sobre todo si las muestras no son muy grandes, por lo que puede ser pertinente utilizar otros procedimientos. Algunos de los métodos más utilizados se basan en suponer una distribución normal para los datos. Sin embargo, en la práctica, no suele ser adecuado, ya que la mayoría de parámetros biológicos tienen distribuciones que suelen alejarse de ese modelo, presentando asimetría o apuntamiento. Aunque una simple transformación de los datos podría solventar este problema, en la aplicación que se ha planteado en este trabajo, no sería suficiente, puesto que la distribución de los niveles de la creatinina depende, de forma muy importante, de características individuales, como la edad, sexo y el grupo racial. En estos casos no será adecuado proporcionar unos IRs globales, sino que éstos deberían ser determinados en función de estas variables explicativas.

Además, la dependencia puede existir no sólo en los valores medios de nuestras variables de interés, sino también en los otros parámetros que caracterizan a la distribución: variabilidad, simetría y/o apuntamiento. Por todo ello, los modelos GAMLSS [25], como una extensión de los GLM y de los GAM, son una herramienta adecuada para llevar a cabo las estimaciones de las curvas centiles que permiten establecer unos IRs útiles.

Como se ha mostrado en este trabajo, los IRs construidos, tienen un buen comportamiento en una muestra externa procedente de la misma población. Sin embargo, para terminar la fase de validación, sería muy útil disponer de una muestra de individuos con alguna patología renal, para evaluar la capacidad de identificar sujetos con este tipo de desorden. Lamentablemente, no se ha podido completar esta parte por falta de datos.

Referencias

[1] Yutaka Tonomura, Mitsunobu Matsubara, and Itsuro Kazama. Biomarkers in urine and use of creatinine. In *General Methods in Biomarker Research and their Applications*, pages 165–186. Springer International Publishing, 2015.

[2] Paul S Horn and Amadeo J Pesce. Reference intervals: an update. *Clinica Chimica Acta*, 334(1-2):5–23, 2003.

[3] Itziar Fernández, Amalia Enríquez-de Salamanca, Alejandro Portero, Carmen García-Vázquez, Margarita Calonge, and José M Herreras. Age-and sex-adjusted reference intervals in tear cytokine levels in healthy subjects. *Applied Sciences*, 11(19):8958, 2021.

[4] Elaine Borghi, Mercedes de Onis, Cutberto Garza, Jan Van den Broeck, Edward A Frongillo, Laurence Grummer-Strawn, S Van Buuren, H Pan, L Molinari, Reynaldo Martorell, et al. Construction of the world health organization child growth standards: selection of methods for attained growth curves. *Statistics in medicine*, 25(2):247–265, 2006.

[5] Liliya Chamitava, Vanessa Garcia-Larsen, Lucia Cazzoletti, Paolo Degan, Andrea Pasini, Valeria Bellisario, Angelo G Corsico, Morena Nicolis, Mario Olivieri, Pietro Pirina, et al. Determination of adjusted reference intervals of urinary biomarkers of oxidative stress in healthy adults using gamlss models. *PloS one*, 13(10):e0206176, 2018.

[6] A Aellig, A Albert, G Blin, J Buret, E Daubrosse, M Drosdowsky, J Favre, B Gouget, L Guize, J Henny, et al. Société française de biologie clinique. section of physiopathology. commissionreference values". utilisation of reference values.(document j, stage 3, version 1). In *Annales de biologie clinique*, volume 40, pages 697–708, 1982.

[7] Luisa Martinez-Sanchez, Fernando Marques-Garcia, Yesim Ozarda, Albert Blanco, Nannette Brouwer, Francesca Canalias, Christa Cobbaert, Marc Thelen, and Wendy den Elzen. Big data and reference intervals: rationale, current practices, harmonization and standardization prerequisites and future perspectives of indirect determination of reference intervals using routine data. *Advances in Laboratory Medicine/Avances en Medicina de Laboratorio*, 2(1):9–16, 2020.

[8] Callum G Fraser. *Biological variation: from principles to practice*. Amer. Assoc. for Clinical Chemistry, 2001.

[9] Raúl León Barua and Roberto Berenson Seminario. Medicina teórica.: Definición de la salud. *Revista Médica Herediana*, 7(3):105–107, 1996.

[10] Ari Lahti, Per Hyltoft Petersen, James C Boyd, Callum G Fraser, and Nils Jørgensen. Objective criteria for partitioning gaussian-distributed reference values into subgroups. *Clinical chemistry*, 48(2):338–352, 2002.

[11] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[12] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. CRC press, 2018.

[13] Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

[14] D Mikis Stasinopoulos and Robert A Rigby. Generalized additive models for location

scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46, 2008.

[15] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[16] Robert A Rigby and DM Stasinopoulos. A flexible regression approach using gamlss in r. *London Metropolitan University, London*, page 47, 2009.

[17] Mikis Stasinopoulos, Bob Rigby, Vlasios Voudouris, Calliope Akantziliotou, Marco Enea, and Daniil Kiose. Package 'gamlss'. *Dist'2020Available online: http://www. gamlss. org (accessed on 16 July 2021)*, 2023.

[18] Hirotugu Akaike. Information measures and model selection. *Int Stat Inst*, 44:277–291, 1983.

[19] Peter K Dunn and Gordon K Smyth. Randomized quantile residuals. *Journal of Computational and graphical statistics*, 5(3):236–244, 1996.

[20] Stef van Buuren and Miranda Fredriks. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, 20(8):1259–1277, 2001.

[21] Centers for Disease Control, Prevention, et al. National health and nutrition examination survey https://www. cdc. gov/nchs/nhanes, 2017.

[22] Lesley A Stevens, Shani Shastri, and Andrew S Levey. Assessment of renal function. In *Comprehensive clinical nephrology*, pages 31–38. Elsevier Inc., 2010.

[23] Andrew S Levey, Josef Coresh, Kline Bolton, Bruce Culleton, Kathy Schiro Harvey, T Alp Ikizler, Cynda Ann Johnson, Annamaria Kausz, Paul L Kimmel, John Kusek, et al. K/doqi clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases*, 39(2 SUPPL. 1):i–ii+, 2002.

[24] PA Wayne. Clsi defining, establishing, and verifying reference intervals in 2005clinical laboratory-approved guideline. *CLSI Document EP28-A3C. Third edition. Available online: http://shop. clsi. org/site/Sample_pdf/EP28A3C_sample. pdf (accessed on 19 October 2010)*, 2008.

[25] Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

Índice de figuras

Índice de tablas

Índice de abreviaturas

- BCPE: Cox-Cox Power Exponential.
- CLSI: Clinical and Laboratory Standards Institute.
- EMV: Estimación por Máxima Verosimilitud)
- GAM: Generalized Additive Model.
- GAMLSS: Generalized Additive Models for Location, Scale, and Shape.
- GG: Generalized Gamma.
- GLM: Generalized Linear Model.
- IRLSM: Iteratively Reweighted Least Squares.
- IC: Confidence Interval.
- IFCC: International Federation of Clinical Chemistry and Laboratory Medicine.
- IR: Reference Interval.
- IRs: Reference Intervals.
- LM: Linear Model.
- NCCLS: National Committee for Clinical Laboratory Standards.
- OLS: Ordinary Least Squares.
- OMS: World Health Organization.
- WEI: Weibull.

Anexo A: Análisis descriptivo

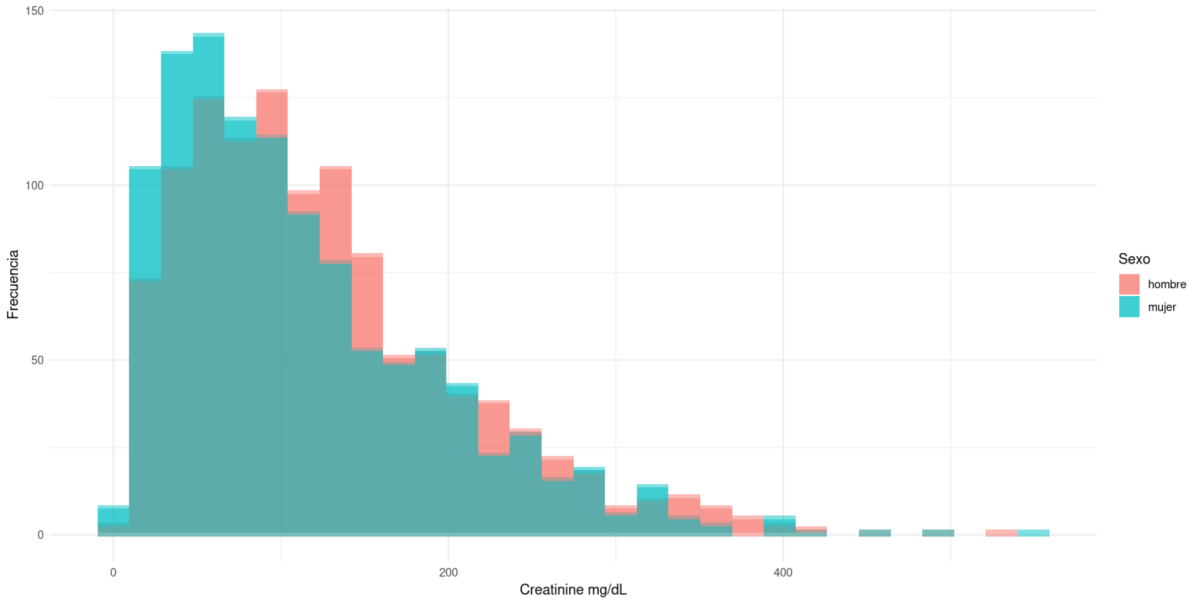

Figura 14: Frecuencia de creatinina en orina segun el sexo para el estrato de edad pediátrica en la muestra de referencia.

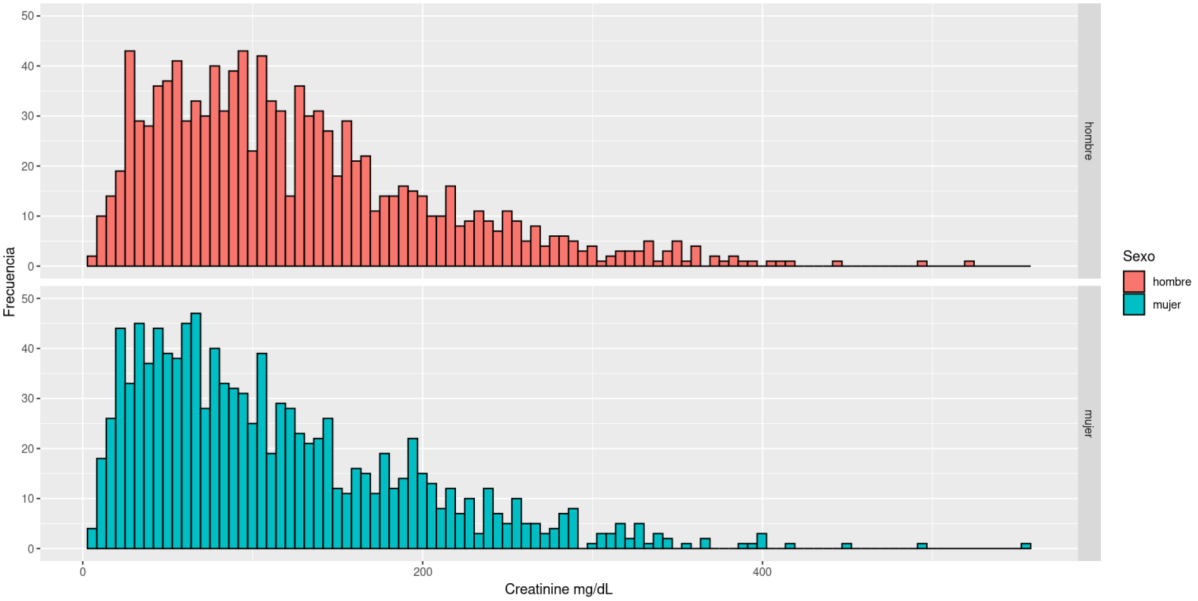

Figura 15: Frecuencia de los valores de creatinina en orina segun el sexo para el estrato de edad pediátrica en la muestra de referencia.

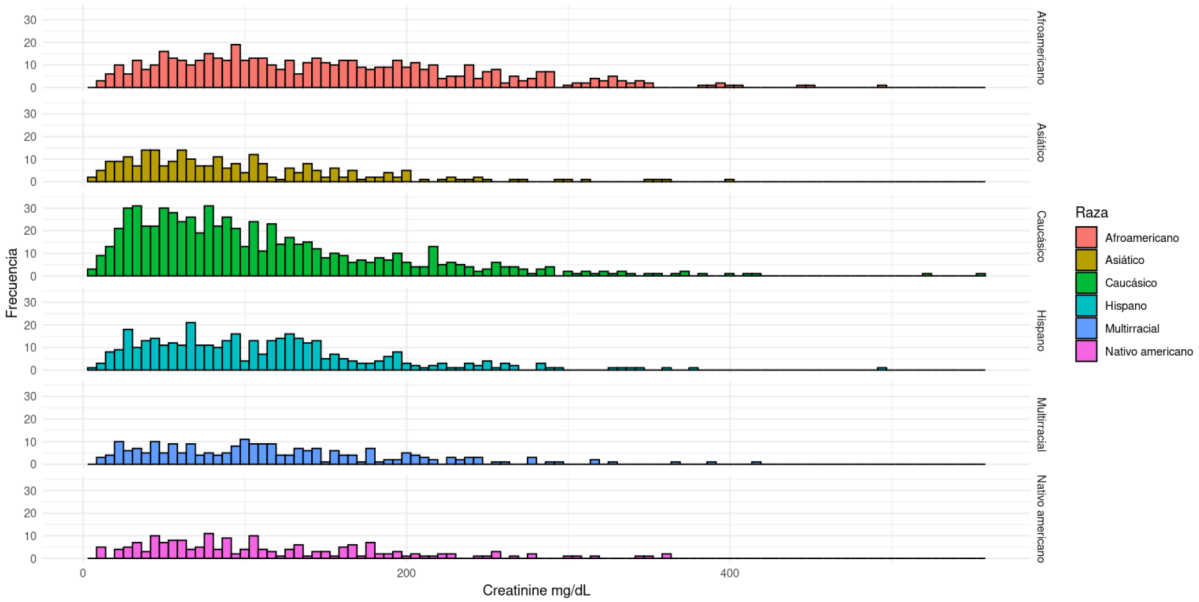

Figura 16: Frecuencia de los valores de creatinina en orina segun la raza para el estrato de edad pediátrica en la muestra de referencia.

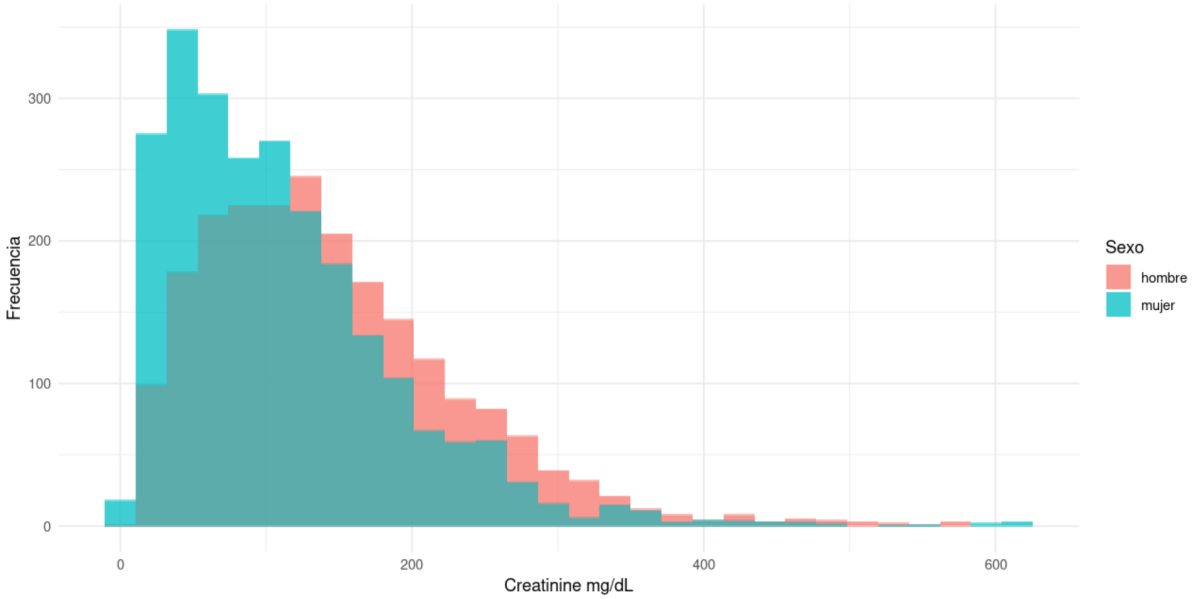

Figura 17: Frecuencia de los valores de creatinina en orina segun el sexo para el estrato de edad adulta en la muestra de referencia.

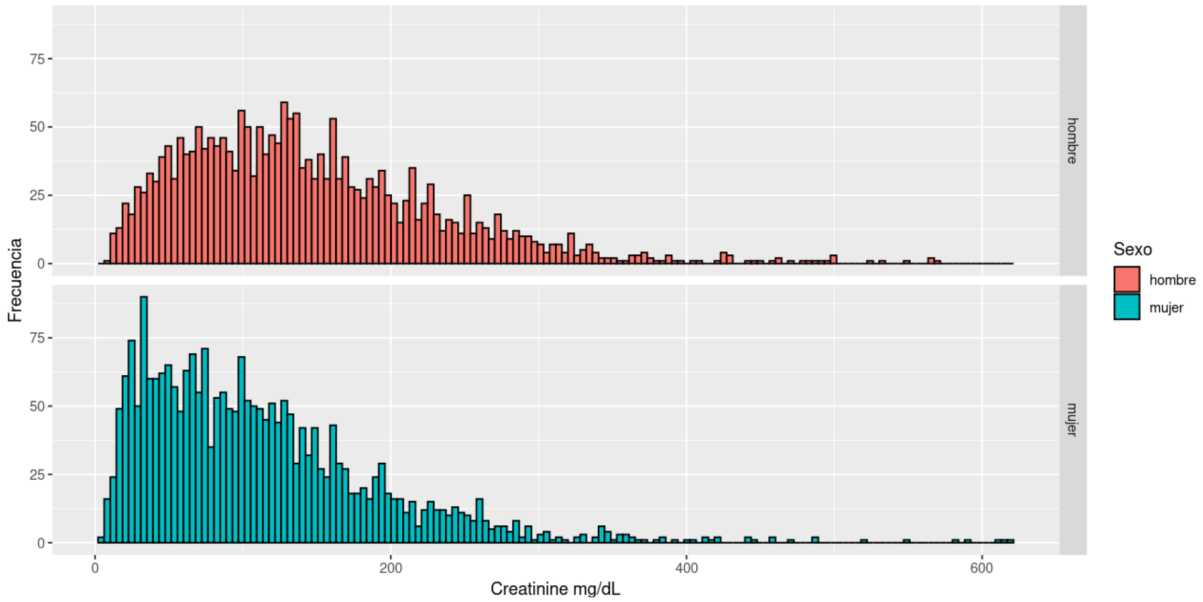

Figura 18: Frecuencia de los valores de creatinina en orina segun el sexo para el estrato de edad adulta en la muestra de referencia.

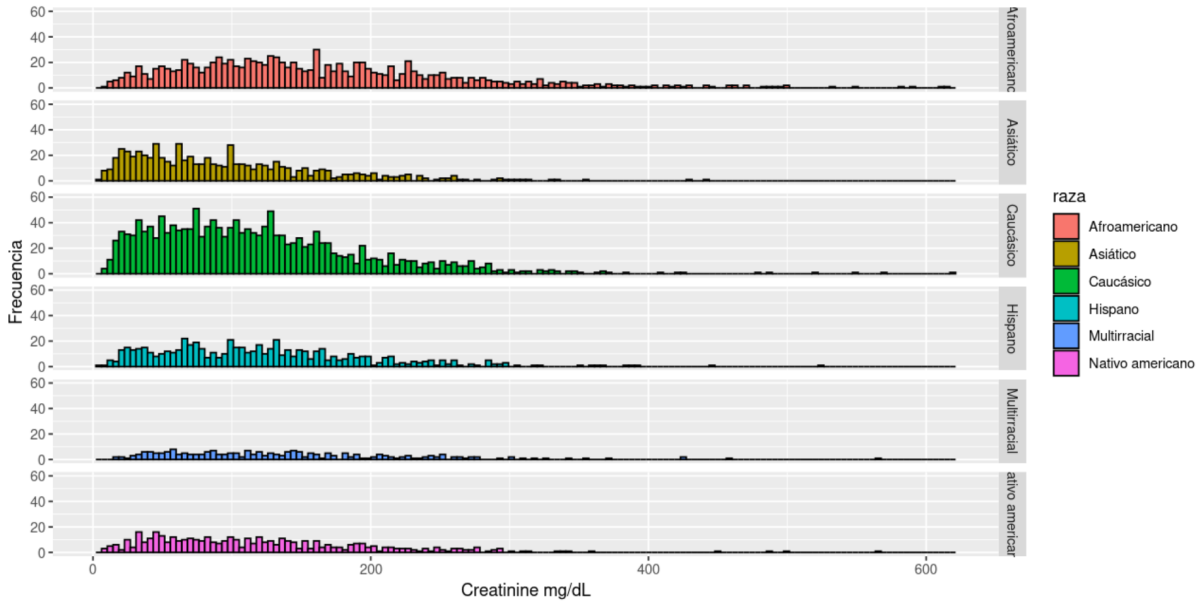

Figura 19: Frecuencia de los valores de creatinina en orina segun la raza para el estrato de edad adulta en la muestra de referencia.

Anexo B: IRs segundo mejor modelo

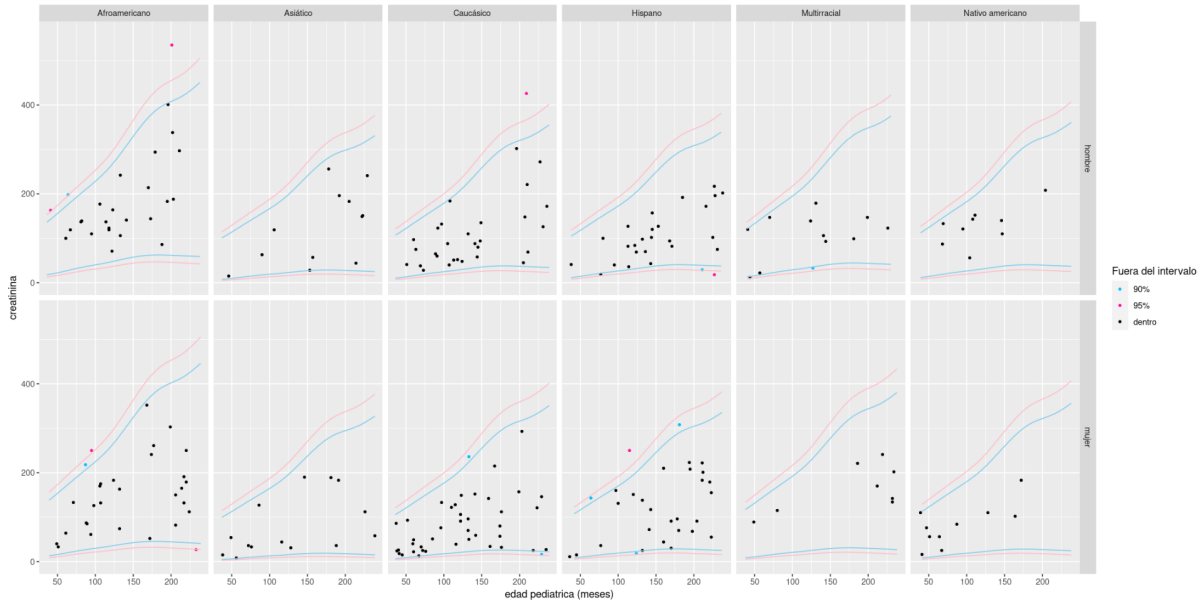

Figura 20: IRs para el modelo basado en la distribución GG en el estrato de edad pediátrica estimados con la muestra de referencia.

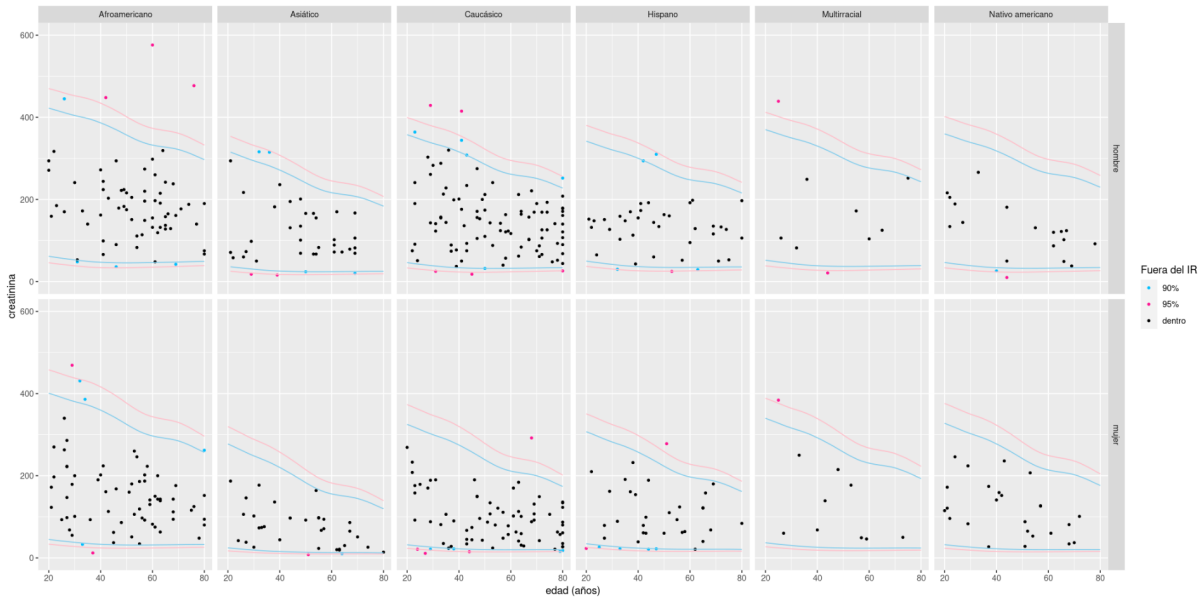

Figura 21: IRs para el modelo basado en la distribución BCPE en el estrato de edad adulta estimados con la muestra de referencia.

Anexo C: Librería gamlss de R

La librería gamlss en R es una herramienta para el ajuste de modelos GAMLSS, que permiten modelar la distribución completa de una variable de respuesta, incluyendo su ubicación (media), escala (varianza) y forma (parámetros de distribución) [17].

La librería gamlss proporciona funciones para ajustar y analizar modelos GAMLSS de manera flexible y eficiente. Algunas características y funcionalidades clave de gamlss son:

- Especificación de modelos: permite especificar modelos GAMLSS utilizando fórmulas similares a las utilizadas en lm() o glm(), lo que facilita el modelado de relaciones no lineales y no paramétricas.
- Selección de distribuciones: ofrece una amplia gama de distribuciones para modelar diferentes tipos de datos, incluyendo distribuciones gaussianas, exponenciales, Gamma, Weibull, lognormal, entre otras.
- Estimación de parámetros: utiliza métodos de estimación robustos y eficientes para estimar los parámetros del modelo, incluyendo el algoritmo de backfitting y el método de penalización de máxima verosimilitud.
- Diagnóstico de ajuste: proporciona herramientas para evaluar el ajuste del modelo, incluyendo gráficos de diagnóstico, pruebas de bondad de ajuste y comparación de modelos.
- Predicción y visualización: permite realizar predicciones a partir del modelo ajustado y generar gráficos para visualizar los resultados, como gráficos de ajuste o gráficos de residuos.

Para utilizar gamlss, se debe instalar previamente ejecutando **install.packages("gamlss")**. Después, se carga la librería en lu sesión de R con **library(gamlss)**

Función gamlss()

La función gamlss() devuelve un objeto de la clase "gamlss", que es un modelo GAMLSS, ésta es muy similar a la función gam(), pero puede ajustar más distribuciones y puede modelar todos los parámetros de la distribución como funciones de las variables explicativas.

Uso

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

Argumentos

formula: Un objeto, con la variable de respuesta a la izquierda del operador $\sim$, y los términos separados por el operador + a la derecha. Los términos de suavizado no paramétrico se indican con pb() para splines beta penalizados, cs para splines de suavizado, lo para términos de suavizado loess y random o ra para términos aleatorios. Las interacciones con términos de suavizado no paramétrico no están completamente soportadas, pero no generarán errores, simplemente producirán la interacción paramétrica habitual.

sigma.formula: Un objeto para ajustar un modelo al parámetro σ, como en la fórmula anterior.

nu.formula: Unpara ajustar un modelo al parámetro ν.

tau.formula: Un objeto para ajustar un modelo al parámetro τ.

family: Un objeto gamlss.family, que se utiliza para definir la distribución y las funciones de enlace de los diversos parámetros. Las familias de distribución soportadas por gamlss() se pueden encontrar en gamlss.family.

data: Un data frame que contiene las variables que aparecen en la fórmula.

weights: Un vector de pesos, Los pesos se pueden utilizar para ponderar las observaciones o para un análisis de verosimilitud ponderada donde la contribución de las observaciones a la verosimilitud difiere según los pesos. La longitud de weights debe ser la misma que el número de observaciones en los datos. Por defecto, el peso se establece en uno.

contrasts: Una lista de contrastes que se utilizarán para algunos o todos los factores que aparecen como variables en la fórmula del modelo. Los nombres de la lista deben ser los nombres de las variables correspondientes.

method: Los algoritmos actuales para GAMLSS son RS(), CG() y mixed(). method=RS() utilizará el algoritmo de Rigby y Stasinopoulos, method=CG() utilizará el algoritmo de Cole y Green, y mixed(2,10) utilizará el algoritmo RS dos veces antes de cambiar al algoritmo de Cole y Green para hasta 10 iteraciones adicionales.

start.from: Un modelo GAMLSS ajustado cuyos valores ajustados se utilizarán como valores iniciales para el modelo actual.

mu.start: Vector o escalar de valores iniciales para el parámetro de ubicación μ.

sigma.start: Vector o escalar de valores iniciales para el parámetro de escala σ.

nu.start: Vector o escalar de valores iniciales para el parámetro ν.

tau.start: Vector o escalar de valores iniciales para el parámetro τ.

mu.fix: Indica si el parámetro μ debe mantenerse fijo en los procesos de ajuste.

sigma.fix: Indica si el parámetro σ debe mantenerse fijo en los procesos de ajuste.

nu.fix: Indica si el parámetro ν debe mantenerse fijo en los procesos de ajuste.

tau.fix: Indica si el parámetro τ debe mantenerse fijo en los procesos de ajuste.

control: Establece los parámetros de control del algoritmo de iteraciones externas.

i.control: Esto establece los parámetros de control de las iteraciones internas del algoritmo RS

Salidas

Variable	Descripción
family	La familia de distribución del objeto gamlss
parameters	Los nombres de los parámetros ajustados
call	La llamada a la función gamlss
y	La variable de respuesta
control	La configuración de control del ajuste de gamlss
weights	El vector de pesos
G.deviance	La deviance global
N	El número de observaciones en el ajuste
rqres	Una función para calcular los residuos normalizados cuantiles
iter	El número de iteraciones externas en el proceso de ajuste
type	El tipo de distribución o variable de respuesta
method	El algoritmo utilizado para el ajuste, RS(), CG() o mixed()
converged	Si el ajuste del modelo ha convergido
residuals	Los residuos normalizados cuantiles del modelo
mu.fv	Los valores ajustados del modelo μ, también sigma.fv, nu.fv, tau.fv para los otros parámetros si están presentes
mu.lp	El predictor lineal del modelo μ, también sigma.lp, nu.lp, tau.lp para los otros parámetros si están presentes
mu.wv	La variable de trabajo del modelo μ, también sigma.wv, nu.wv, tau.wv para los otros parámetros si están presentes
mu.wt	Los pesos de trabajo del modelo μ, también sigma.wt, nu.wt, tau.wt para los otros parámetros si están presentes
mu.link	La función de enlace para el modelo μ, también sigma.link, nu.link, tau.link para los otros parámetros si están presentes
mu.terms	Los términos del modelo μ, también sigma.terms, nu.terms, tau.terms para los otros parámetros si están presentes
mu.x	La matriz de diseño para μ, también sigma.x, nu.x, tau.x para los otros parámetros si están presentes
mu.qr	La descomposición QR del modelo μ, también sigma.qr, nu.qr, tau.qr para los otros parámetros si están presentes
mu.coefficients	Los coeficientes lineales del modelo μ, también sigma.coefficients, nu.coefficients, tau.coefficients para los otros parámetros si están presentes
mu.formula	La fórmula para el modelo μ, también sigma.formula, nu.formula, tau.formula para los otros parámetros si están presentes
mu.df	Los grados de libertad de μ, también sigma.df, nu.df, tau.df para los otros parámetros si están presentes
mu.nl.df	Los grados de libertad no lineales, también sigma.nl.df, nu.nl.df, tau.nl.df para los otros parámetros si están presentes
df.fit	Los grados de libertad totales utilizados por el modelo
df.residual	Los grados de libertad residuales restantes después de ajustar el modelo
aic	El criterio de información de Akaike
sbc	El criterio de información bayesiano

Anexo D: Código R

Librerias necesarias

```
library(readr)
library(dplyr)
library(pastecs)
library(psych)
library(Hmisc)
library(skimr)
library(ggplot2)
library(ggpubr)
library(gamlss)
```

Importación de los datos

```
#Pediatricos
pediatricos <- read_csv("tfg/training.pediatricos.csv")
test_pediatricos <- read_csv("tfg/test.pediatricos.csv")

#adultos
adultos<- read_csv("tfg/training.adultos.csv
test_adultos <- read_csv("tfg/test.adultoss.csv"
```

Análisis descriptivo numérico individuos pediátricos

```r
#training
summary(pediatricos)
describe(pediatricos)
describeBy(pediatricos$creatine, group = pediatricos$sexo, mat = T, digits = 2)
describeBy(pediatricos$creatine, group = pediatricos$raza, mat = T, digits = 2)

#sexo
counts <- aggregate(pediatricos$sexo~ sexo, data = pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'hombre')))
counts <- aggregate(pediatricos$sexo ~ sexo, data =pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'mujer')))
counts$porcentaje <- counts$`pediatricos$sexo`[ ,2] /2249 *100
counts$ic_inf <- tapply(counts$`pediatricos$sexo`[ ,2], counts$sexo, function(x
    ) {
  prop.test(sum(x), 2249)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`pediatricos$sexo`[ ,2], counts$sexo, function(x
    ) {
  prop.test(sum(x), 2249)$conf.int[2] *100
})


#raza
counts <- aggregate(pediatricos$raza~ raza, data = pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'Asiatico')))
counts <- aggregate(pediatricos$raza ~ raza, data =pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'Caucasico')))
counts <- aggregate(pediatricos$raza ~ raza, data =pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'Hispano')))
counts <- aggregate(pediatricos$raza ~ raza, data =pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'Afroamericano')))
counts <- aggregate(pediatricos$raza ~ raza, data =pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'Multirracial')))
counts <- aggregate(pediatricos$raza ~ raza, data =pediatricos, FUN = function(
    x) c(n = length(x), sum = sum(x == 'Nativo americano')))
counts$porcentaje <- counts$`pediatricos$raza`[ ,2] /2249 *100
counts$ic_inf <- tapply(counts$`pediatricos$raza`[ ,2], counts$raza, function(x
    ) {
  prop.test(sum(x), 2249)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`pediatricos$raza`[ ,2], counts$raza, function(x
    ) {
  prop.test(sum(x), 2249)$conf.int[2] *100
})

#test
summary(test_pediatricos)
describe(test_pediatricos)
describeBy(test_pediatricos$creatine, group = test_pediatricos$sexo, mat = T,
    digits = 2)
```

```r
describeBy(test_pediatricos$creatine, group = test_pediatricos$raza, mat = T,
    digits = 2)



#sexo
counts <- aggregate(test_pediatricos$sexo~ sexo, data = test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'hombre')))
counts <- aggregate(test_pediatricos$sexo ~ sexo, data =test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'mujer')))
counts$porcentaje <- counts$`test_pediatricos$sexo`[ ,2] /250 *100
counts$ic_inf <- tapply(counts$`test_pediatricos$sexo`[ ,2], counts$sexo,
    function(x) {
  prop.test(sum(x), 250)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`test_pediatricos$sexo`[ ,2], counts$sexo,
    function(x) {
  prop.test(sum(x), 250)$conf.int[2] *100
})


#raza
counts <- aggregate(test_pediatricos$raza~ raza, data = test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'Asiatico')))
counts <- aggregate(test_pediatricos$raza ~ raza, data =test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'Caucasico')))
counts <- aggregate(test_pediatricos$raza ~ raza, data =test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'Hispano')))
counts <- aggregate(test_pediatricos$raza ~ raza, data =test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'Afroamericano')))
counts <- aggregate(test_pediatricos$raza ~ raza, data =test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'Multirracial')))
counts <- aggregate(test_pediatricos$raza ~ raza, data =test_pediatricos, FUN =
     function(x) c(n = length(x), sum = sum(x == 'Nativo americano')))
counts$porcentaje <- counts$`test_pediatricos$raza`[ ,2] /250 *100
counts$ic_inf <- tapply(counts$`test_pediatricos$raza`[ ,2], counts$raza,
    function(x) {
  prop.test(sum(x), 250)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`test_pediatricos$raza`[ ,2], counts$raza,
    function(x) {
  prop.test(sum(x), 250)$conf.int[2] *100
})
```

Análisis descriptivo numérico individuos adultos

```r
#training
summary(adultos)
describe(adultos)
describeBy(adultos$creatine, group = adultos$sexo, mat = T, digits = 2)
describeBy(adultos$creatine, group = adultos$raza, mat = T, digits = 2)



#sexo
counts <- aggregate(adultos$sexo~ sexo, data = adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'hombre')))
counts <- aggregate(adultos$sexo ~ sexo, data =adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'mujer')))
counts$porcentaje <- counts$`adultos$sexo`[ ,2] /4614 *100
counts$ic_inf <- tapply(counts$`adultos$sexo`[ ,2], counts$sexo, function(x) {
  prop.test(sum(x), 4614)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`adultos$sexo`[ ,2], counts$sexo, function(x) {
  prop.test(sum(x), 4614)$conf.int[2] *100
})


#raza
counts <- aggregate(adultos$raza~ raza, data = adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'Asiatico')))
counts <- aggregate(adultos$raza ~ raza, data =adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'Caucasico')))
counts <- aggregate(adultos$raza ~ raza, data =adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'Hispano')))
counts <- aggregate(adultos$raza ~ raza, data =adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'Afroamericano')))
counts <- aggregate(adultos$raza ~ raza, data =adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'Multirracial')))
counts <- aggregate(adultos$raza ~ raza, data =adultos, FUN = function(x) c(n =
    length(x), sum = sum(x == 'Nativo americano')))
counts$porcentaje <- counts$`adultos$raza`[ ,2] /4614 *100
counts$ic_inf <- tapply(counts$`adultos$raza`[ ,2], counts$raza, function(x) {
  prop.test(sum(x), 4614)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`adultos$raza`[ ,2], counts$raza, function(x) {
  prop.test(sum(x), 4614)$conf.int[2] *100
})



#sexo
counts <- aggregate(test_adultos$sexo~ sexo, data = test_adultos, FUN =
   function(x) c(n = length(x), sum = sum(x == 'hombre')))
counts <- aggregate(test_adultos$sexo ~ sexo, data =test_adultos, FUN =
   function(x) c(n = length(x), sum = sum(x == 'mujer')))
```

```r
counts$porcentaje <- counts$`test_adultos$sexo`[ ,2] /513 *100
counts$ic_inf <- tapply(counts$`test_adultos$sexo`[ ,2], counts$sexo, function(
    x) {
  prop.test(sum(x), 513)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`test_adultos$sexo`[ ,2], counts$sexo, function(
    x) {
  prop.test(sum(x), 513)$conf.int[2] *100
})


#raza
counts <- aggregate(test_adultos$raza~ raza, data = test_adultos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'Asiatico')))
counts <- aggregate(test_adultos$raza ~ raza, data =test_adultos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'Caucasico')))
counts <- aggregate(test_adultos$raza ~ raza, data =test_adultos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'Hispano')))
counts <- aggregate(test_adultos$raza ~ raza, data =test_adultos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'Afroamericano')))
counts <- aggregate(test_adultos$raza ~ raza, data =test_adultos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'Multirracial')))
counts <- aggregate(test_adultos$raza ~ raza, data =test_adultos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'Nativo americano')))
counts$porcentaje <- counts$`test_adultos$raza`[ ,2] /513 *100
counts$ic_inf <- tapply(counts$`test_adultos$raza`[ ,2], counts$raza, function(
    x) {
  prop.test(sum(x), 513)$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$`test_adultos$raza`[ ,2], counts$raza, function(
    x) {
  prop.test(sum(x), 513)$conf.int[2] *100
})
```

Análisis descriptivo gráfico individuos pediátricos

```
ggplot(pediatricos, aes(x = pediatricos$edad, y = pediatricos$creatine, color =
    pediatricos$raza)) +
  geom_point(aes(shape=sexo)) +scale_shape_manual(values=c(4,20))+
  facet_wrap(~ pediatricos$raza,)+
  labs(x="edad pediatrica (meses)",
       y="creatinina")+
  scale_colour_discrete(name="raza")+
  theme_minimal() +  scale_y_continuous(limits = c(0,650))
```

Análisis descriptivo gráfico individuos adultos

```
ggplot(adultos, aes(x = adultos$edad, y = adultos$creatine, color = adultos$
    raza)) +
  geom_point(aes(shape=sexo)) +scale_shape_manual(values=c(4,20))+
  facet_wrap(~ adultos$raza,)+
  labs(x="edad",y="creatinina")+
  scale_colour_discrete(name="raza")+
  theme_minimal()+scale_y_continuous(limits = c(0,650))
```

Ajuste de los modelos GAMLSS

```
# Funcion que ajusta a cada familia, eligiedo las variables explicativas
f.chooseX<-function(DATA,fam){

  mod0 <- gamlss(creatine~1,data=DATA,family=fam, trace=FALSE)  # modelo
    constante

  f.upper=as.formula(paste("~","1+edad+sexo+raza+cs(edad)",sep="")) # modelo
    maximo

  lpar<-length(mod0$parameters)
  if(lpar==2){
    #parametro mu: forward
    mod1<-stepGAIC(mod0, parameter="mu",scope=list(lower=~1,upper=f.upper),
    trace=FALSE,direction="forward")
    #parametro sigma: forward
    mod2 <- stepGAIC(mod1, parameter="sigma", scope=list(lower=~1,upper=f.upper
    ),direction="forward",trace=FALSE)
    # modelo final: backward de mod1 para modelizar mu
    model<-stepGAIC(mod2, parameter="mu",trace=FALSE,direction="backward")
  }


  if(lpar==3){
    #parametro mu: forward
    mod1<-stepGAIC(mod0, parameter="mu",scope=list(lower=~1,upper=f.upper),
    trace=FALSE,direction="forward")
    #parametro sigma: forward
    mod2 <- stepGAIC(mod1, parameter="sigma", scope=list(lower=~1,upper=f.upper
    ),direction="forward",trace=FALSE)
    #parametro nu: forward
    mod3 <- stepGAIC(mod2, parameter="nu", scope=list(lower=~1,upper=f.upper),
    direction="forward",trace=FALSE)
    # backward de mod3 para modelizar sigma
    mod4<-stepGAIC(mod3, parameter="sigma",trace=FALSE,direction="backward")
    # backward de mod4 para modelizar mu
    model<-stepGAIC(mod4, parameter="mu",trace=FALSE,direction="backward")
  }

  if(lpar==4){
    #parametro mu: forward
    mod1<-stepGAIC(mod0, parameter="mu",scope=list(lower=~1,upper=f.upper),
    trace=FALSE,direction="forward")
    #parametro sigma: forward
    mod2 <- stepGAIC(mod1, parameter="sigma", scope=list(lower=~1,upper=f.upper
    ),direction="forward",trace=FALSE)
    #parametro nu: forward
    mod3 <- stepGAIC(mod2, parameter="nu", scope=list(lower=~1,upper=f.upper),
    direction="forward",trace=FALSE,method=mixed())
    #parametro tau: forward
    mod4 <- stepGAIC(mod3, parameter="tau", scope=list(lower=~1,upper=f.upper),
```

```
    direction="forward",trace=FALSE)
    # backward de mod4 para modelizar nu
    mod5<-stepGAIC(mod4, parameter="nu",trace=FALSE,direction="backward")
    # backward de mod5 para modelizar sigma
    mod6<-stepGAIC(mod5, parameter="sigma",trace=FALSE,direction="backward")
    # backward de mod6 para modelizar mu
    model<-stepGAIC(mod6, parameter="mu",trace=FALSE,direction="backward")
  }
  LR.test(mod0,model)

  return(model)
}

#pediatricos
DATA<-pediatricos
fam<-WEI
modelWEI<-f.chooseX(DATA,fam)
fam<-GG
modelGG<-f.chooseX(DATA,fam)

#adultos
DATA<-adultos
fam<-BCPE
amodelBCPE<-f.chooseX(DATA,fam)
fam<-GG
amodelGG<-f.chooseX(DATA,fam)
```

Elección mejor modelo pediatricos

```
term.plot(modelWEI, parameter = "mu", pages = 1, ask = FALSE, rug = TRUE)
drop1(modelWEI, parameter = "mu", parallel = "multicore", ncpus = 4)
term.plot(modelWEI, parameter = "sigma", pages = 1, ask = FALSE, rug = TRUE)
drop1(modelWEI, parameter = "sigma", parallel = "multicore", ncpus = 4)
wp(modelWEI, ylim.all =1)
plot(modelWEI)


term.plot(modelGG, parameter = "mu", pages = 1, ask = FALSE, rug = TRUE)
drop1(modelGG, parameter = "mu", parallel = "multicore", ncpus = 4)
term.plot(modelGG, parameter = "sigma", pages = 1, ask = FALSE, rug = TRUE)
drop1(modelGG, parameter = "sigma", parallel = "multicore", ncpus = 4)
wp(modelGG, ylim.all = 1)
plot(modelGG)
```

Elección mejor modelo adultos

```
plot(fitted(amodelBCPE), residuals(amodelBCPE))
term.plot(amodelBCPE, parameter = "mu", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelBCPE, parameter = "mu", parallel = "multicore", ncpus = 4)
term.plot(amodelBCPE, parameter = "sigma", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelBCPE, parameter = "sigma", parallel = "multicore", ncpus = 4)
term.plot(amodelBCPE, parameter = "nu", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelBCPE, parameter = "nu", parallel = "multicore", ncpus = 4)
term.plot(amodelBCPE, parameter = "tau", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelBCPE, parameter = "tau", parallel = "multicore", ncpus = 4)
wp(amodelBCPE, ylim.all = 1)
plot(amodelBCPE)

term.plot(amodelGG, parameter = "mu", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelGG, parameter = "mu", parallel = "multicore", ncpus = 4)
term.plot(amodelGG, parameter = "sigma", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelGG, parameter = "sigma", parallel = "multicore", ncpus = 4)
term.plot(amodelGG, parameter = "nu", pages = 1, ask = FALSE, rug = TRUE)
drop1(amodelGG, parameter = "nu", parallel = "multicore", ncpus = 4)
wp(amodelGG, ylim.all = 1)
plot(amodelGG)
```

Cálculo de percentiles

```
# Funcion que calcula los percentiles
f.calcula.percentiles<-function(model.opt,new.DATA,cent=c
    (0.025,0.05,0.95,0.975)){
  ## predecimos los valores
  pred <- predictAll(model.opt, newdata = new.DATA, type = "response")

  ## familia del modelo
  fname <- model.opt$family[1]
  qfun <- paste("q",fname,sep="")

  lpar <- length(model.opt$parameters)

  mat<-NULL
  for(i in c(0.025,0.05,0.95,0.975)){
    if(lpar==1)
      newcall <- call(qfun, i, mu = pred$mu)

    if(lpar==2)
      newcall <- call(qfun, i, mu = pred$mu,sigma = pred$sigma)

    if(lpar==3)
      newcall <- call(qfun, i, mu = pred$mu,sigma = pred$sigma, nu = pred$nu)

    if(lpar==4)
      newcall <- call(qfun, i, mu = pred$mu,sigma = pred$sigma, nu = pred$nu,
    tau = pred$tau)

    ll <- eval(newcall)
    mat <- cbind(mat, ll)
  }

  mat <- as.data.frame(mat)
  nnn <- paste("C", as.character(cent*100), sep = "")
  names(mat) <- nnn

  res<-data.frame(new.DATA,mat)

  return(res)
}
```

IRs pediatricos

```r
#pediatricos
newdatamodel<- data.frame(
  edad = pediatricos$edad,
  raza = pediatricos$raza,
  sexo = pediatricos$sexo
)

wei<-data.frame(pediatricos$id,f.calcula.percentiles(modelWEI,newdatamodel))
a<- geom_line(data = wei   , aes(x = edad, y = C2.5), color = c("pink"))
b<- geom_line(data = wei   , aes(x = edad, y = C97.5), color = c("pink"))
c<- geom_line(data = wei   , aes(x = edad, y = C5), color = c("skyblue"))
d<- geom_line(data = wei   , aes(x = edad, y = C95), color = c("skyblue"))
test_pediatricos$fueraic95wei <- "dentro"
for(i in 1:dim(test_pediatricos)[1]){
  for(j in 1:dim(a$data)[1]){
    condicion1 <- abs(test_pediatricos$edad[i] - a$data$edad[j]) <= 1
    condicion2 <- test_pediatricos$raza[i] == a$data$raza[j]
    condicion3 <-test_pediatricos$creatine[i] > a$data$C97.5[j]
    condicion4 <-test_pediatricos$creatine[i] < a$data$C2.5[j]
    condicion5 <- test_pediatricos$sexo[i] == a$data$sexo[j]
    condicion6 <-test_pediatricos$creatine[i] > a$data$C95[j]
    condicion7 <-test_pediatricos$creatine[i] < a$data$C5[j]
    if(condicion1 && condicion2 && condicion5 && (condicion3||condicion4)) {
      test_pediatricos$fueraic95wei[i] <- "95%"
    } else if(condicion1 && condicion2 && condicion5 && (condicion6||condicion7
    )) {
      test_pediatricos$fueraic95wei[i] <- "90%"
    }
  }
}
ggplot(test_pediatricos, aes(x = edad, y = creatine, color = fueraic95wei)) +
  geom_point(shape=20) +
  facet_grid(sexo ~ raza) + a+b+c+d+
  labs(x = "edad pediatrica (meses)", y = "creatinina") +
  scale_color_manual(name = "Fuera del IR", values = c("dentro" = "black", "
    95%" = "deeppink", "90%" = "deepskyblue")) +
  scale_y_continuous(limits = c(0, 560)) + scale_x_continuous(limits = c(36,
    239))nd{Rcode}
```

IRs adultos

```
#Adultos
newdatamodela<- data.frame(
  edad = adultos$edad,
  raza = adultos$raza,
  sexo = adultos$sexo
)

ggA<-f.calcula.percentiles(amodelGG,newdatamodela)
a<- geom_line(data = ggA  , aes(x = edad, y = C2.5), color = c("pink"))
b<- geom_line(data = ggA  , aes(x = edad, y = C97.5), color = c("pink"))
c<- geom_line(data = ggA  , aes(x = edad, y = C5), color = c("skyblue"))
d<- geom_line(data = ggA  , aes(x = edad, y = C95), color = c("skyblue"))
test_adultos$fueraic95gg <- "dentro"
for(i in 1:dim(test_adultos)[1]){
  for(j in 1:dim(a$data)[1]){
    condicion1 <- abs(test_adultos$edad[i] - a$data$edad[j]) <= 1
    condicion2 <- test_adultos$raza[i] == a$data$raza[j]
    condicion3 <-test_adultos$creatine[i] > a$data$C97.5[j]
    condicion4 <-test_adultos$creatine[i] < a$data$C2.5[j]
    condicion5 <- test_adultos$sexo[i] == a$data$sexo[j]
    condicion6 <-test_adultos$creatine[i] > a$data$C95[j]
    condicion7 <-test_adultos$creatine[i] < a$data$C5[j]
    if(condicion1 && condicion2 && condicion5 && (condicion3||condicion4)) {
      test_adultos$fueraic95gg[i] <- "95%"
    } else if(condicion1 && condicion2 && condicion5 && (condicion6||condicion7
    )) {
      test_adultos$fueraic95gg[i] <- "90%"
    }
  }
}
ggplot(test_adultos, aes(x = edad, y = creatine, color = fueraic95gg)) +
  geom_point(shape=20) +
  facet_grid(sexo ~ raza) + a+b+c+d+
  labs(x = "edad", y = "creatinina") +
  scale_color_manual(name = "Fuera del IR", values = c("dentro" = "black", "
    95%" = "deeppink", "90%" = "deepskyblue")) +
  scale_y_continuous(limits = c(0, 600)) + scale_x_continuous(limits = c(19,
    80))
```

Validación de los IRs

```r
#validaion de porcentajes (pediatricos)

#sexo- Gamlss
counts <- aggregate(fueraic95wei ~ sexo, data = test_pediatricos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'dentro')))
counts <- aggregate(fueraic95wei ~ sexo, data = test_pediatricos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'dentro')+sum(x == '90%')))
counts$porcentaje <- counts$fueraic95wei[,2] /counts$fueraic95wei[,1]
counts$ic_inf <- tapply(counts$fueraic95wei[,2], counts$sexo, function(x) {
  prop.test(sum(x), sum(counts$fueraic95wei[,2]))$conf.int[1] *100
})

counts$ic_sup <- tapply(counts$fueraic95wei[,2], counts$sexo, function(x) {
  prop.test(sum(x), sum(counts$fueraic95wei[ ,2]))$conf.int[2] *100
})

#raza - Gamlss
counts <- aggregate(fueraic95wei ~ raza, data = test_pediatricos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'dentro')))
counts <- aggregate(fueraic95wei ~ raza, data = test_pediatricos, FUN =
    function(x) c(n = length(x), sum = sum(x == 'dentro')+sum(x == '90%')))
counts$porcentaje <- counts$fueraic95wei[,2] /counts$fueraic95wei[,1]
counts$ic_inf <- tapply(counts$fueraic95wei[,2], counts$raza, function(x) {
  prop.test(sum(x), sum(counts$fueraic95wei[ ,2]))$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$fueraic95wei[,2], counts$raza, function(x) {
  prop.test(sum(x), sum(counts$fueraic95wei[ ,2]))$conf.int[2] *100
})



#validaion de porcentajes (adultos)

#sexo- Gamlss
counts <- aggregate(fueraic95gg ~ sexo, data = test_adultos, FUN = function(x)
    c(n = length(x), sum = sum(x == 'dentro')))
counts <- aggregate(fueraic95gg ~ sexo, data = test_adultos, FUN = function(x)
    c(n = length(x), sum = sum(x == 'dentro')+sum(x == '90%')))
counts$porcentaje <- counts$fueraic95gg[,2] /counts$fueraic95gg[,1]
counts$ic_inf <- tapply(counts$fueraic95gg[,2], counts$sexo, function(x) {
  prop.test(sum(x), sum(counts$fueraic95gg[ ,2]))$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$fueraic95gg[,2], counts$sexo, function(x) {
  prop.test(sum(x), sum(counts$fueraic95gg[ ,2]))$conf.int[2] *100
})

#raza - Gamlss
counts <- aggregate(fueraic95gg ~ raza, data = test_adultos, FUN = function(x)
    c(n = length(x), sum = sum(x == 'dentro')))
counts <- aggregate(fueraic95gg ~ raza, data = test_adultos, FUN = function(x)
```

```
    c(n = length(x), sum = sum(x == 'dentro')+sum(x == '90%')))
counts$porcentaje <- counts$fueraic95gg[,2] /counts$fueraic95gg[,1]*100
counts$ic_inf <- tapply(counts$fueraic95gg[,2], counts$raza, function(x) {
  prop.test(sum(x), sum(counts$fueraic95gg[ ,2]))$conf.int[1] *100
})
counts$ic_sup <- tapply(counts$fueraic95gg[,2], counts$raza, function(x) {
  prop.test(sum(x), sum(counts$fueraic95gg[ ,2]))$conf.int[2] *100
})
```