UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Diciembre 2017

Estudio de la incidencia actitudinal de los estudiantes en pruebas de matemáticas tipo ICFES: Una aproximación semiparamétrica

Study of the students' attitudinal incidence in icfes type math tests: A semi-parametric approach

Diego Armando Corredor Rivera.[a] Wilmer Pineda Ríos.[b]

Lida Fonseca Gómez.[c]

Resumen

En temas académicos, los efectos marginales de la educación varían en función de distintos niveles de capacidades de los estudiantes. De esta manera, la omisión del comportamiento de la educación junto con el efecto de interacción de la misma con las capacidades de los estudiantes, conlleva a tener resultados erróneos si se utilizaran métodos lineales para la estimación de los parámetros de interés. Los modelos de regresión semiparamétrica tienen como principal objetivo analizar la relación entre la variable respuesta y las explicativas, esta relación se puede establecer por medio de una función de distribución conocida, que facilita la estimación de los parámetros y la interpretación de los resultados del modelo. Los errores pueden tener una estructura desconocida la cual puede tener efectos negativos en la estimación de los parámetros, ya que estos pueden llegar a tener gran parte de la información que no se ajusta adecuadamente al modelo, es por esta razón que se estudiaran los diferentes tipos de errores que se pueden obtener al ajustar un modelo semiparamétrico, obteniendo así la mejor estimación insesgada del modelo y la menor perdida de información en los errores.

***Palabras clave*:** modelo semiparamétrico bayesiano, muestreador de Gibbs, familia de los errores .

Abstract

In academic issues, the marginal effects of education vary according to different levels of student abilities, thus the omission of the non-linear form of education along with the effect of its interaction with the students' abilities, It leads to erroneous results if linear methods were used to estimate the parameters of interest. Semi-parametric regression models have as main objective to relate and analyze information that has a linear and non-linear structure, this relationship can be established by means of a known distribution function, which facilitates the estimation of the parameters and the interpretation of the results of the model. The residuals can have an unknown structure which can have negative effects on the estimation of the parameters, since these can have a large part of the information that does not fit properly to the model, that is why the different types will be studied of residuals that can be obtained by adjusting a semiparametric model, thus obtaining the best unbiased estimate of the model and the least loss of information in the residuals.

***Keywords*:** bayesian semiparametrics, Gibbs sampler, errors family .

[a]Estudiante de Estadística Universidad Santo Tomás
[b]Director de opción de grado, docente Universidad Santo Tomás
[c]Co-director de opción de grado, docente Universidad Santo Tomás

1. Introducción

En la actualidad la educación es un tema de gran importancia en cualquier país, ya que por medio del conocimiento y el avance en cuanto a los resultados en las pruebas tipo ICFES, aplicadas a los estudiantes se obtiene aceptación en organizaciones internacionales, que se encargan de la calidad educativa y medición intelectual de cada país. Los gobiernos se encargan de calificar y poner a prueba a los colegios, obteniendo los resultados de la calidad educativa en cada institución establecida por el Ministerio de Educación Nacional. Es por este motivo que los colegios toman decisiones en cuanto a la mejora de las calificaciones y el desempeño de los estudiantes. El presente estudio trata de ajustar un modelo semiparamétrico para un colegio de Bogotá interesado en relacionar los indicadores de los aspectos: afectivo, cognitivo y expresivo, ya que se considera que tienen gran impacto en las calificaciones de los estudiantes de matemáticas tipo ICFES. Se pretende sustentar estadísticamente la:

Teoría académica: la cual considera que la dimensión cognitiva se deja afectar por las dimensiones afectivo y expresivo, teniendo así un efecto en las calificaciones de matemáticas de los estudiantes de grado undécimo. Para el caso puntual de este estudio se estudian datos continuo, que poseen una estructura lineal y una no lineal (funcional), haciendo uso de un modelo semiparamétrico, analizado con diferentes tipos de errores, se tendrá la mejor estimación de los parámetros del modelo propuesto.

Los aportes de **Fonseca (2012)** a los datos ,los cuales se utilizaron en el trabajo de grado titulado: análisis estadístico a un indicador de logro estudiantil (Desempeño académico)", en el momento en que se trabajó con los datos el objetivo principal era **determinar factores escolares y sociales que afectan el indicador de logro estudiantil** a partir de análisis multivariado, con el fin de que el colegio pueda evaluar y proponer estratégias que permitan mejorar el indicador de logro (desempeño académico) de los estudiantes.

Hablando de temas educativos, siempre se ha querido saber los factores que aportan a un buen desempeño del estudiante, bien sea en todas las áreas del conocimiento o en solo algunas, entidades como las nombradas anteriormente, que se dedican a la calificación de las pruebas de desempeño académico aplicadas en los colegios, dichos resultados son dados a conocer tanto en cada colegio y a nivel nacional, teniendo la necesidad cada centro educativo de mejorar el desempeño de los estudiantes en cuanto a las pruebas se refiere. Entendiendo las dimensiones que se han trabajado anteriormente a los datos en estudio, se observa la necesidad de explicar la incidencia de la actitud, el conocimiento y la aplicación matemática en los puntajes obtenidos en el grado once para el colegio de investigación, solo se estudiará la relación de la asignatura de matemáticas con respecto a los aspectos: afectivo,cognitivo y expresivo.

Los modelos de regresión semiparamétrica tienen como principal objetivo relacionar y analizar información que tenga una estructura lineal y no lineal, esta relación se puede establecer por medio de una función de distribución conocida, la cual facilita la estimación de los parámetros y la interpretación de los resultados del modelo, puesto que a mayor grado del polinomio el ajuste de los datos mejora, no obstante tener un polinomio de grado elevado puede aumentar la complejidad del modelo, disminuir su precisión y dificultad su interpretación.

Hay dos grandes enfoques en el tema de modelos de suavizado con splines:

Splines de suavizado (smoothing splines): los splines de suavizado utilizan tantos parámetros como observaciones, lo que hace que su implementación no sea eficiente cuando el número de datos es muy elevado.

Splines de regresión (regresión splines): pueden ser ajustados mediante mínimos cuadrados una vez que se han seleccionado el número de nodos, pero la selección de los nodos se hace mediante algoritmos complejos.

La organización de este trabajo será la siguiente. En la sección 3 se exponen las generalidades de la regresión semiparamétrica y se describen los diferentes tipos de errores que se pueden ajustar en los modelos semiparamétricos. En la sección 4 se encuentra las especificaciones del modelo semiparamétrico propuesto, en la sección 5 se presenta la comparación del modelo con diferentes tipos de errores aplicado

a datos reales y sus correspondientes cadenas de convergencia y se presentan las conclusiones y finalmente en el apéndice se encuentra los diferentes resultados del software estadístico R utilizado en el desarrollo del trabajo e implementación del modelo.

2. Pedagogía conceptual

La aplicación en estudio tiene como objetivo establecer un modelo semiparamétrico que permita explicar el puntaje de matemáticas por medio de las dimensiones psicológicas que inciden en el proceso cognitivo. Los datos son obtenidos por medio de un colegio de Bogotá, el cual tiene interés en conocer si la dimensión cognitiva de los estudiantes es afectada por las dimensiones afectiva o expresiva.

Las dimensiones en el desarrollo psicológico a considerar son las siguientes:

Dimensión Afectiva: El modo como la gente reacciona emocionalmente, su habilidad para sentir el dolor o la alegría de otro ser viviente.Los objetivos afectivos apuntan típicamente a la conciencia y crecimiento en actitud, emoción y sentimientos. Comprende los niveles: recepción, respuesta, valoración, organización y caracterización.

Dimensión expresiva: Los objetivos expresivos generalmente apuntan en el cambio desarrollado en la conducta o habilidades.Comprende los niveles: percepción, disposición, mecanismo, respuesta compleja, adaptación, creación.

Dimensión Cognitiva: Es la habilidad para pensar las cosas. Los objetivos cognitivos giran en torno del conocimiento y la comprensión de cualquier tema dado. Se subdividen en seis niveles: conocimiento, comprensión, aplicación, análisis, síntesis y evaluación.

Según la teoría académica de la institución de educación a la cual se le esta realizando el estudio, la dimensión cognitiva en las matemáticas es afectada por la dimensiones afectiva y expresiva.

Hablando de temas educativos se deben tener en cuenta las principales entidades u organizaciones que son responsables de las calificaciones y el desarrollo de la calidad de la educación, los siguientes son aportes que han hecho a las nociones de competencia:

Secretaría de Educación de Bogotá: La competencia se entiende como un saber hacer?, frente a una tarea específica cuando el sujeto entra en contacto con ella. Supone conocimientos, saberes y habilidades que emergen en la interacción que se establece entre el individuo y la tarea.

Sistema Nacional de Evaluación, ICFES: El conocimiento no solo es concebido como la suma de principios y contenidos que deben ser aprehendidos para su transmisión sino como aquellas reglas de acción que nos garantizan su manejo? Aquí, la competencia es entendida como un saber hacer? o conocimiento implícito en un campo del actuar humano, Una acción situada que se define en relación con determinados instrumentos mediadores?

Programa de Naciones Unidas para el desarrollo de la educación: La universidad, en particular tiene la obligación de formar en la inteligencia social?, esto es la capacidad para adaptarse a un mundo que cambia rápidamente, lo cual supone adquirir y procesar información sumamente compleja, para toma de decisiones óptimas. Además de los conocimientos y habilidades propias de cada caso, hoy se subraya la importancia de que el alumno comprenda lo que hace. Así la comprensión es el elemento que diferencia al trabajador competente de hoy al trabajador calificado del pasado: comprender su trabajo y el medio donde trabaja es la clave que aporte a la solución de problemas,para tener iniciativa en la resolución de situaciones inesperadas y contar con la capacidad de aprender constantemente.

3. Regresión lineal paramétrica

Es una de las técnicas estadística más antiguas y está dada por la relación de datos $(x_i, Y_i), i = 1, .., n$. El siguiente modelo de regresión lineal simple describe la relación de las variables.

$$y = \beta_0 + \beta_1 x + \epsilon_i \tag{1}$$

Donde:

- B_0 y B_1 son los parámetros del modelo y
- e es una variable aleatoria

e es llamada error, la cual explica la variabilidad de y que no puede ser explicada con la relación lineal entre x y y. Los errores e, se consideran variables aleatorias independientes distribuidas normalmente con media cero y desviación estándar σ. Esto implica que el valor medio o valor esperado de y, es denotado por:

$$E(y/x) = \beta_0 + \beta_1 x \tag{2}$$

El objetivo principal de la regresión es proporcionar un resumen o una reducción de los datos observados con el fin de explorar y dar a conocer la relación existente entre la variable independiente x y la variable de respuesta y. Al momento de graficar los datos es lógico y natural que se evidencie una línea recta que ratifica la tendencia y relación. Principalmente la regresión utiliza el modelo (1) para la predicción dado un punto x, y obtener así una estimación del valor esperado de y. Esta estimación está dada por la siguiente expresión:

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x \tag{3}$$

Este modelo puede ser fácilmente interpretado y estimado con una buena precisión si los supuestos subyacentes son correctos. Si no se cumple con los supuestos, las estimaciones pueden ser inconsistentes y posiblemente tener un mal ajuste de la regresión.

3.1. Regresión no paramétrica

Por lo general, el análisis de Regresión lineal implica relacionar una variable de respuesta como una función de al menos una variable explicativa determinista, a través de alguna relación paramétrica. En tales circunstancias, la relación es específicamente predefinida por el modelo, por lo cual estos modelos son opciones fáciles en la práctica por la simplicidad y la interpretabilidad. Sin embargo, es evidente que pueden existir relaciones más complejas entre la respuesta y las variables explicativas, y en tal caso no es viable la utilizacion de estos modelos.

El uso de polinomios de orden superior puede, pero no siempre, dar solución a este problema. Existen en la literatura técnicas más flexibles para hacer frente a tales situaciones. La lista de tales técnicas incluye, pero no se limita a, modelos de regresión no paramétricos, tales como la estimación a través de Kernel (Azzalini Bowman 1993), Regresión polinomial local(Cleveland Loader 1996),y suavizado de spline (Silverman 1985).

La técnica de suavización conocida como Kernel, hace uso de una función que asigna pesos a las observaciones lejanas o cercanas al punto de interés. La función que asigna pesos en esta técnica de suavización está determinada por una función Kernel y por un parámetro de suavización, Härdle (1992).

Los splines con penalizaciones combinan lo mejor de ambos enfoques: utilizan menos parámetros que los splines de suavizado, pero la selección de los nodos no es tan determinante como en los splines de regresión. Hay tres razones fundamentales para el uso de este tipo de splines:

Splines de rango bajo, es decir, el tamaño de la base utilizada es mucho menor que la dimensión de los datos, al contrario de lo que ocurre en el caso de los splines de suavizado donde hay tantos nodos como datos, lo que hace que sea necesario trabajar con matrices de alta dimensión. El número de nodos, en el caso de los P-splines, no supera los 40, lo que hace que sean computacionalmente eficientes, sobre todo cuando se trabaja con gran cantidad de datos.

La introducción de penalizaciones relaja la importancia de la elección del número y la localización de los nodos, cuestión que es de gran importancia en los splines de rango bajo sin penalizaciones (ver por ejemplo Rice and Wu (2001)).

Varios autores, Nadaraya Watson (1964), Cleveland (1979), Gasser Müller (1979), Priestley Chao (1972), han propuesto diversos estimadores o suavizadores no paramétricos que permiten describir el comportamiento característico de un conjunto de datos, sin suponer a priori una función de regresión conocida. Supóngase para un conjunto de observaciones (x_i, y_i); $i = 1, 2, ., n$, la relación de regresión expresada por la siguiente forma:

$$y_i = f(x_i) + e_i \tag{4}$$

con $i = 1, ..., n$ donde y_i representa la variable dependiente, x_i los valores de la variable independiente o de diseño, $f(x_i)$ la función de regresión f desconocida evaluada en los puntos de diseño, y e_i una variable aleatoria denominada error. Un objetivo del análisis de regresión paramétrica es obtener una aproximación razonable de la función desconocida f, reduciendo el error de las observaciones y centrándose en la respuesta media de y, condicionada a x, lo que se expresa como:

$$f(x_i) = E(y|X = x_i) \tag{5}$$

con $E(e_i) = 0$ y var $Var(e_i) = \sigma^2$ para todo $i = 1, 2, .., n$.

Los estimadores de regresión no paramétrica tienen como fin estimar la forma de la función f a través de técnicas de suavización (Kernel, KVecinos más cercanos, Spline, etc.), sin necesidad de especificar a priori su estructura. La filosofía de estos métodos se basa en la idea de dejar que los datos hablen por sí mismos, Olaya (2005).

Eubank (1999), especifica que en regresión no paramétrica, se debe asumir que f es suave; es decir, que si se desea estimar la función f en un punto x, se espera que las observaciones y_i asociadas a los x_i cercanos a x, posean información de f en x, lo cual indica que es posible promediar de alguna forma las respuestas y_i más cercanas al punto donde se estime $f(x)$. Los estimadores más comunes suelen tener una estructura lineal en la variable dependiente, es decir, se expresa como una función lineal de la variable respuesta.

3.2. Distribución normal multivariada con mixtura en la escala

Según la inferencia estadística basada en el enfoque Bayesiano, para modelos de regresión bajo el supuesto de que los errores aditivos independientes siguen distribuciones normal, Student-t, diagonal, normal contaminada, Laplace o hiperbólica simétrica, es decir, los errores aditivos siguen una escala de mezclas de distribuciones normales, segun el tipo de error se tienen las siguientes estructuras:

De acuerdo con Andrews y Mallows (1974), un vector continuo $Y = (Y_1, ..., Y_r)^T$ sigue una mezcla de distribuciones normal multivariada con mixtura en la escala, denotada por SMN_r se puede escribir por:

$$Y = \mu + K^{1/2}(U)Z \tag{6}$$

Donde:

- $\mu \epsilon \mathbb{R}^r$, es el parámetro local
- $Z \sim N_r(0, \sigma^2 I)$
- K es una función positiva
- U es una variable aleatoria positiva independiente de Z, tiene función de distribución acumulada $H(u, \eta)$
- η es un vector de parámetros indexados de la distribución U.

Cuando los errores de la normal multivariada toma el valor del parámetro = 1, se tiene lo siguiente:

$$P\left[U = 1\right] = 1 \tag{7}$$

3.3. Distribución t-student multivariada

En este caso $U \sim Gamma(\eta/2, \eta/2)$ y $\eta > 0$, asi la densidad del vector aleatorio esta dado por:

$$f(y|\mu, \Sigma, \eta) = \frac{\Gamma(\frac{\eta+r}{2})}{\Gamma(\frac{\eta}{2})|\Sigma|^{\frac{1}{2}}(\pi\eta)^{\frac{r}{2}}} \left[1 + \frac{\delta^2}{\eta}\right]^{-(\frac{\eta+r}{2})} \tag{8}$$

Donde:

- $y \sim tr(\mu, \Sigma, \eta)$
- $\delta^2 = (y - \mu)^T \Sigma^{-1} (y - \mu)$

3.4. Distribución Slash multivariada

En este caso $U \sim Beta(\eta, 1)$, $\eta > 0$. Entonces la densidad del vector Y es:

$$f(y|\mu, \Sigma, \eta) = \frac{\eta}{(2\pi)^{\frac{r}{2}}|\Sigma|^{\frac{1}{2}}} \int_1^0 u^{\eta+\frac{r}{2}-1} exp\left[-u\delta^2\right] du \tag{9}$$

Donde: $Y \sim Sl_r(\mu, \Sigma, \eta)$

3.5. Distribución Normal contaminada multivariada

En este caso U es una variable aleatoria discreta, toma valores η_2 con probabilidad η_1 y valor 1 con probabilidad $(1 - \eta_1)$, asi dado U por:

$$h(u|\eta = (\eta_1, \eta_2)^T) = \eta_1 \mathbb{I}(u = \eta_2) + (1 - \eta_1)\mathbb{I}(u = 1) \tag{10}$$

sigue una distribución

$$f(y|\mu, \Sigma, \eta) = \eta_1 \phi_r(y|\mu, \eta_2^{-1}\Sigma) + (1 - \eta_1)\phi_r(y|\mu, \Sigma) \tag{11}$$

Donde: $0 < \eta_1 < 1$ y $0 < \eta_2 < 1$, entonces $CN_r \sim Sl_r(\mu, \Sigma, \eta)$

3.6. Distribución Hiperbólica simétrica multivariada

En este caso, $U \sim GIG(1, 1, \eta^2)$, La distribución del vector Y se reduce a:

$$f(y|\mu, \Sigma, \eta) = \frac{K_a(\eta\sqrt{\sigma^2+1})\eta^{r/2}(\sigma^2+1)^{-\frac{r}{4}+\frac{1}{2}}}{2^{r/2}\pi^{r/2}\left|\Sigma\right|^{1/2}K_1(\eta)} \tag{12}$$

Donde:

- $a = -\frac{r}{2} + 1$
- $k_a(\eta) = \frac{1}{2}\int_0^\infty x^{a-1}exp(-\frac{1}{2}\eta(x + x^{-1}))\partial x$

Por lo tanto $Y \sim SH_r(\mu, \Sigma, \eta), \eta > 0$

4. Regresión semiparamétrica

En términos generales, los métodos semiparamétricos son aproximaciones de medición que mantienen la estructura en un método empírico que es útil para la interpretación de los resultados, pero que no se apoya en supuestos específicos sobre características que resultan de interés secundario(Stoker,1991). En particular, los métodos semiparamétricos permiten la estimación de parámetros y funciones simultáneamente, sin supuestos específicos respecto de las formas de las funciones desconocidas. Como tal, los estimadores semiparamétricos son menos sensibles a supuestos que los estimadores paramétricos, y son capaces de describir la estructura de los datos de forma más clara que estos Caguari (2009).

La modelación semiparamétrica permite a un investigador tener lo mejor de ambos enfoques para obtener un modelo de regresión que describa mejor el comportamiento de los datos, es decir, aquellas características de los datos que son adecuadas para la modelación paramétrica son modeladas de esta forma y las componentes no paramétricas son empleadas solo donde sea necesario, Ruppert et al. (2009).

Dos características importantes en gran parte de la regresión semiparamétrica, Ruppert et al. (2009), son:

- Emplear la representación del modelo mixto de los splines penalizados Estas brindan varios beneficios: los efectos longitudinales y espaciales pueden ser fácilmente incorporados en el modelo, el ajuste y la inferencia pueden ser desarrollados dentro de los marcos establecidos de máxima verosimilitud y mejor predicción.
- Facilitar la parte de regresión no paramétrica utilizando splines penalizados de bajo rango.

$$y_i = m(x_i) + e_i \tag{13}$$

donde los $e_i \sim N(0, \sigma)$ son independientes de x_i y $m(.)$ es una función suavizada, que se define como una función suave de los datos porque puede ser modelada fácilmente utilizando splines como sugiere Crainiceanu et al. (2005). Es evidente que un modelo de este tipo es una generalización de un modelo de regresión que, indudablemente, tendrá un costo computacional, pero que permitirá estimar la función de una forma más precisa.

4.1. Modelo con términos paramétricos y no paramétricos

Un modelo que mezcla términos paramétricos y no paramétricos toma la siguiente expesión:

$$Y_i = \alpha + \Sigma K = 1^J f_k(Y_{ik}) + \Sigma K = j + 1^K X_{ik} B_{k-j} + \varepsilon_i \tag{14}$$

Donde:

$\varepsilon_i \sim SMN(0, \sigma^2 I, H, K)$

La primeras covariables j tienen un efecto lineal sobre Y_i y estan equipadas con suavizadores no paramétricos.

El resto de las variables entran en el modelo paramétrico.

4.2. Modelo Semiparamétrico Elíptico Generalizado

Se supone que se tiene un conjunto de n observaciones $Y_1, ..., Y_n$ y se genera de la siguiente manera:

$$Y_i = \mu_i + \sigma\epsilon_i, i = 1, ..n \tag{15}$$

Donde: $\epsilon_i \sim SMN(0, \sigma^2 I, H, K)$, los parametros de localización y dispersión pueden escribirse como:

$$\mu_i = x_i^T \beta + \sum_{j=1}^{S_1} f_j(v_{ij}) \tag{16}$$

Donde:

- $(x_i^T, v_{i1}, ..., v_{is_1})^T$ es un vector de las variables explicativas asociadas a parámetros de localización y dispersión del individuo i.
- $f_j(.)(j = 1, ..., s_1)$ Son funciones arbitarias de las variables explicativas continuas v_j y w_r.
- $\beta = (\beta_1, ..., \beta_p)^T$ y $\gamma = (\gamma_1, ..., \gamma_q)^T$ Son vectores de parámetros desconocidos.

4.3. Funciones no paramétricas

Para las funcion no paramétrica $f_j(.)$ se hace una aproximación usando B-splines, los valores $f_j(v_{ij})$ y puede ser aproximado por $b_{ij}^T \alpha_j$ y $d_{ir}^T \lambda_r$

Donde:

- $b_{ij} = (b_{1j}(v_{ij}), ..., b_{K_{\alpha_j} j}(v_{ij}))^T$ y $d_{ir} = (d_{1r}(w_{ir}), ..., d_{K_{\lambda_r} r}(w_{ir}))^T$ Son vectores de función base.
- $\alpha_j \in R^{K_{\lambda_j}}$ Son los vectores de coeficientes B-Spline.
- $K_{\alpha j} = M_{\alpha j} + k_{\alpha j}$ y $K_{\lambda r} = M_{\lambda r} + k_{\lambda r}$.

Puede ser escrito finalmente como:

$$\mu = X\beta + \sum_{j=1}^{S_1} B_j^T \alpha_j \tag{17}$$

Donde:
$X = (x_1, ..., x_n)^T$, $B_j = (b_{1j}, ..., b_{nj})^T$ Son las matrices asociadas a los efectos no paramétricos representados por $f_j(.)$.

La función de verosimilitud del vector de parámetros de interés $\theta = (\beta^T, \alpha_i^T)^T$, es decir, la función de verosimilitud es obtenida asumiendo realizaciones aleatorias de las variables $U_1, ..., Un$, denotado por $u_1, ..., u_n$, se puede expresar como:

$$L(\theta|y, X, v, u) \propto (\sigma^2)^{-n/2} exp{-1/2\sigma^2}(y - XB - \bar{B})_{(}u)^T(y - XB - \bar{B}) \tag{18}$$

Donde:

- $y = (y_1, ..., y_n)^T$
- $\bar{B} = (B_1, B_2, ..., B_{s1}$
- $\bar{\alpha} = (\alpha_1^T, ..., \alpha_{s1}^T)^T)$
- $L_{(u)} = diagK_{U1}, ..., K_{U_n}$
- $u = (u_1, ..., u_n)^T$
- $v = (v_1, ..., v_n)^T$

4.4. Enfoque bayesiano del modelo semiparamétrico

En estudios estadísticos, se puede tener información previa sobre los parámetros de interés. Esta información puede ser considerada formal para el análisis, resumiéndose en una función de densidad de dichos parámetros. Esta función de densidad, que es denotada como $P(\beta)$ y denominada función de densidad prior, depende de un conjunto conocido de parámetros β_p llamados como hiperparámetros.

Después de los valores de la variables de interés Y son observados, hay dos fuentes de información para los parámetros, una está dada por la función de densidad prior $P(\beta)$ y la otra fuente está dada por la función de verosimilitud $L(\beta|Y)$. Así, en el análisis bayesiano, la inferencia está basado en la funcián de densidad posterior de los parámetros, denotada por $\pi(\beta)$, la cual es obtenida a través de la aplicación del Teorema de Bayes, expresado como (Hoff 2009):

$$\pi\beta \propto L(\beta)P(\beta) \tag{19}$$

Basados en la anterior ecuación las funciones de densidad prior $P(\beta)$ son la componente esencial para la obtención de la distribución posterior , de allí que dado su naturaleza son independientes y pueden o no aportar información al modelo, por tanto según (Hoff 2009) son denominadas: informativas y no informativas. Para el caso particular del plantemiento de la estimación de los parámetros del modelo se han uso de funciones prior no informativas que de manera general se denotan como:

$$\beta \sim \mathbb{N}(0, \sigma^2) \tag{20}$$

donde el hiperparámetro $\sigma^2 > 0$.

Por otro lado, dado que el manejo variables independientes permite el cálculo de la función de verosimilitud a través de una productoria, nuestra propuesta de modelo pone de manifiesto la complejidad que

se genera en la obtención de dicha función, ya que se obtienen funciones conjuntas de probabilidad.

Bajo la anterior premisa y haciendo uso de la ecuación (19), la distribución de probabilidad posterior se expresa como:

$$P(\beta, \alpha|Y_1, ..., Y_k) \propto (Y_1, ..., Y_k|\beta, \alpha)P(\beta), P(\alpha) \tag{21}$$

donde $P(\beta)$ y $P(\alpha)$ representan las distribuciones prior.

La estimación de dichos parámetros requiere métodos iterativos enmarcados en MCMC tal como lo plantea (Gamerman Lopes 2006), para lo que hace uso de algoritmos como Metropolis Hasting y el muestrador de Gibbs, con el único fin de extraer muestras de distribuciones posterior, y observar la convergencia de las cadenas y así la estimación de los parámetros del modelo.

Antes de mostrar el comportamiento del dichos algoritmos, es necesario recordar que MCMC es una técnica que simula una cadena de Markov cuyos estados siguen una función de probabilidad dado un estado de grandes dimensiones tal como lo señala (Hoff 2009). De igual forma, una cadena de Markov es un modelo matemático ligado a sistemas estocásticos, donde los estados dependen de las probabilidades de transición, es decir, que el estado actual solo depende de su estado anterior.

Según lo expresado por (Koop Tobias 2007) el método de Monte Carlo está categorizado como un método no determinístico utilizado para aproximar expresiones matemáticas complejas de evaluar con exactitud. Donde este método posee un error absoluto de la estimación, el cual decrece como $\frac{1}{\sqrt{x}}$ de acuerdo con el Teorema de Límite Central, de allí que, a partir de varias repeticiones se busca reconocer el comportamiento del sistema, destacando que la base de estas simulaciones es ser generadas a partir de números aleatorios.

Para la presente investigación se hace uso del muestrador de Gibbs, siguiendo la propuesta de (Besag Mollie. 1991) bajo el argumento que es un caso particular del algorítmo de Metropolis Hasting, que permite aprovechar una de las ventajas del enfoque bayesiano, ya que no sólo se refiere a estimaciones puntuales, sino al hecho que dicho algoritmo es utilizado para la estimación de parámetros del modelo semiparamétrico, por medio del paquete **BayesGesm** que principalmente se basa en la capacidad de los B-splines para expresarse linealmente y obtener la distribución de los errores del modelo como una combinación de escala de distribuciones normales.
A priori se supone que los cuatro vectores de parámetros son independientes y normalmente distribuidos.

Inicialmente, la distribución previa para el vector de parámetro de interés debe ser especificado por, $\beta, \alpha_1, ..., \alpha_{s1}, \gamma, \lambda_1, ..., \lambda_{s2}$,supone que es independiente y se distribuye a priori de acuerdo con:

$$\beta \sim N_p(B_0, S_\beta), \alpha_j \sim N_{K_{aj}}(\alpha, 0, \tau_{aj}^2, I_{K_{aj}}) \tag{22}$$

Donde: los hiperparámetros β_0, α_{j0}, se asumen conocidos con I_k como la matriz identidad de orden K. Además, $\tau_{\alpha j}^2 \sim IG(a_{\tau_{aj}})$ proveniente de una distribución inversa gamma.

4.5. Muestreador de Gibbs

El muestreador de Gibbs es un caso particular del algoritmo de Metropolis Hasting, que consiste en tener bloques las distribuciones propuestas que coincidan con la distribuciones posterior condicionadas,

de modo que la probabilidad de aceptación es 1. Los pasos que realiza este algoritmo se enumeran a continuación:

- Se fija un valor inicial $\beta_1^0, ..., \beta_k^0$
- Se simula $\beta_1^{(t+1)} \sim \pi(\beta_1|\beta_2^t, ..., \beta_k^t)$
- Se simula $\beta_2^{(t+1)} \sim \pi(\beta_2|\beta_1^{(t+1)}, \beta_3^t, ..., \beta_k^t)$
- Se simula $\beta_3^{(t+1)} \sim \pi(\beta_3|\beta_1^{(t+1)}, \beta_2^{(t+1)}, \beta_4^t, ..., \beta_k^t)$
- Así sucesivamente
- Se simula $(\beta_k^{(t+1)} \sim \pi(\beta_k|\beta_1^{(t+1)}, ..., \beta_{(k-1)}^{(t+1)})$

4.6. Diagnósticos de convergencia

Posterior a la estimación bayesiana de los parámetros del modelo, se busca validar las estimaciones realizadas bajo la convergencia de las cadenas, siendo los diagnósticos de convergencia la herramienta que cumple con este objetivo. Por tanto, a continuación se presentan tres de los criterios más destacados dentro de la literatura:

- **Heidelberger and Welch:** es un diagnóstico de control de longitud de ejecución basado en un criterio de precisión relativa para la estimación de la media. El ajuste predeterminado corresponde a una precisión relativa de dos dígitos significativos. También elimina hasta la mitad de la cadena para asegurar que los medios se estimen a partir de una cadena que ha convergido. (P PD 1981)
- **Raftery and Lewis:** es un diagnóstico de control de longitud de ejecución basado en un criterio de exactitud de estimación del cuantil q. Está pensado para su uso en una prueba piloto corta de una cadena de Markov. También calcula el número de iteraciones de "quemado"que se descartarán al principio de la cadena.(Raftery Lewis 1995)
- **Geweke:**Diagnóstico de convergencia para cadenas de Markov basado en una prueba de igualdad de los medias de la primera y última parte de una cadena de Markov (por defecto el primer 10 % y el último 50 %). Si las muestras se extraen de la distribución estacionaria de la cadena, los dos medios son iguales y la estadística de Geweke tiene una distribución normal asintóticamente estándar. (Geweke 1992)

5. Resultados

5.1. Variables de estudio

Para el presente estudio, las variables de estudio se dividen de la siguiente manera:

Evaluación ICFES: Variable respuesta cuyas cifras fueron obtenidas de los resultados en el examen de estado ICFES, en el cual se mide el conocimiento de cada estudiante para las diferentes áreas y asignaturas vistas en cada institución de educación.

Dimensión Afectiva: Variable explicativa cuantitativa, obtenida de medir capacidades del estudiante

en los aspectos de razonamiento en los niveles: elemental básico, expresivo, avanzado.

Dimensión Cognitiva: Variable explicativa cuantitativa, obtenida de medir capacidades del estudiante en los niveles: elemental básico, expresivo, avanzado.

Dimensión Expresiva: Variable explicativa cuantitativa, obtenida de medir capacidades del estudiante en los niveles: elemental básico, expresivo, avanzado.

De la base de datos se tienen 111 calificaciones correspondientes a los estudiantes del colegio en estudio, las cuales se miden en las escalas: básica, elemental y avanzada.

5.2. Análisis factorial a un solo factor

Inicialmente, se realiza un analisis factorial a un factor para la reduccion de variables y dejar consolidadas las escalas de evaluación en terminos de las dimensiones analizadas, se utiliza la rotación Varimax, ya que es el método que más información permite recoger en el análisis factorial, se obtienen las siguientes graficas:

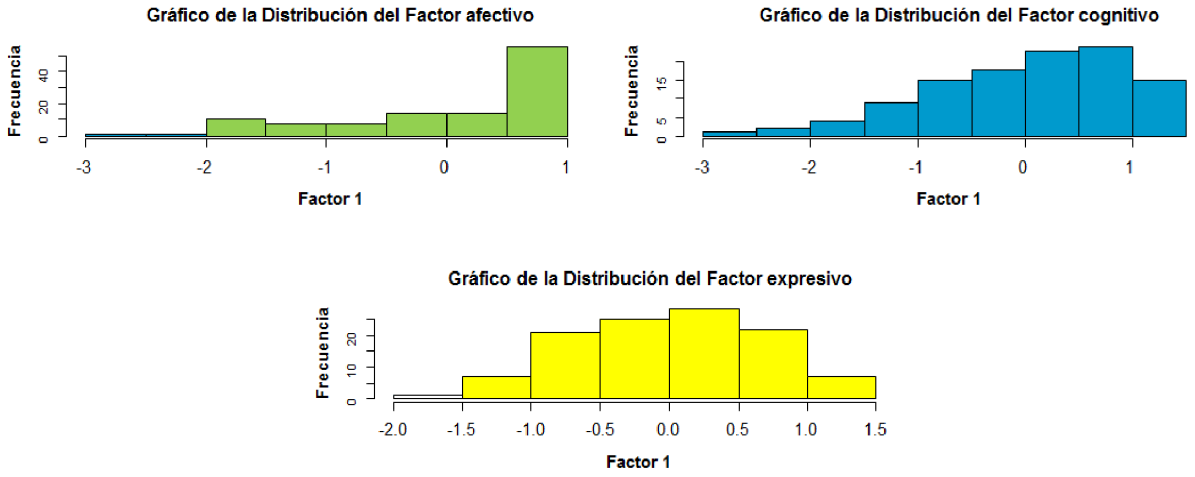


Figura 1: Pesos factoriales

Se observa que los nuevos pesos factoriales tienen una rango de escala normalizado entre -3 y 1, deduciendo que los estudiantes que tienen un buen desempeño en las matemáticas estarán hacia la derecha en cualquier gráfica que explica las dimensiones analizadas, se observaque para la gráfica del factor afectiva se tuvo una simetría marcada hacia la izquierda, lo que indica que los estudiantes que tengan un buen desempeño en esta dimensión, seguramente tendrán un score alto en matemáticas. Sobre los pesos obtenidos se va a realizar el ajuste del modelo y análisis semiparamétrico propuesto.

Se obtienen las gráficas correspondientes de las explicadas anteriormente según los pesos del análisis factorial:

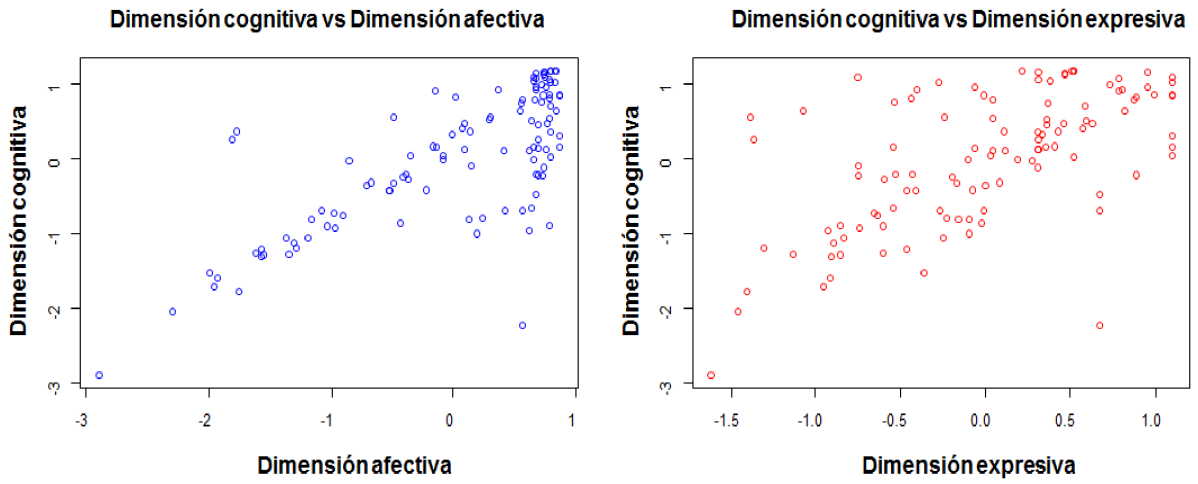


Figura 2: Dimensión cognitiva vs dimensión afectiva y expresiva

En la *figura 2* **Dimension cognitiva vs dimensión afectiva** se observa que los estudiantes que tienen un alto aprecio por las matemáticas en cuanto a la dimension afectiva, tendrán una puntuación mas alta en cuanto a la dimensión cognitiva, siendo esta el resultado de la calificación en las pruebas de matemáticas, se observa un comportamiento lineal en la gráfica, ya que la mayor parte de los puntos tienen una tendencia creciente.

De igual manera se observa que algunos estudiantes que tienen un alto aprecio por las matemáticas en cuanto a la dimension expresiva, tendrán una puntuación mas alta en cuanto a la dimensión cognitiva, siendo este el resultado de la calificación en las pruebas de matemáticas, a diferencia de la gráfica de la dimensión afectiva no se observa un claro comportamiento lineal de los datos ni una tendencia marcada.

5.3. Modelo de regresión normal

Se realizará el ajuste de un modelo lineal simple a los pesos resultantes del analisis factorial, con el fin de poder tener una estimación por el método clásico de los modelos, sin embargo de los resultados en la figura 2, se conoce que de las dos variables independientes una tiene un comportamiento lineal y la otra no.

Coefficients	*Estimate*	*Std. Error*	*t value*	$Pr(>\|t\|)$
Intercepto	-1.118e-18	5.505e-02	0.000	1.00000
scores afectivo	5.646e-01	7.641e-02	7.389	3.3e-11 ***
scores expresivo	3.381e-01	1.028e-01	3.289	0.00136 **
Error residual estandar: 0.58 con 108 grados de libertad				
R cuadrado múltiple: 0.5878, R cuadrado ajustado: 0.5802				
Estadística F: 77.01				
con 2 y 108 grados de libertad, p-value: $< 2.2e-16$				

Tabla 1: Resumen modelo lineal

Según los resultados del modelo normal, se tiene que las dimensiones afectivo y expresivo son significativas para explicar la dimensión congnitiva, este modelo tiene un ajuste del 58.78 %, al tener una sola variable que tiene un comportamiento lineal, se estaría disminuyendo el ajuste del modelo lineal, la dimensión expresiva es la que no posee una tendencia en sus valores, en este caso un modelo clásico no permite explicar en su mayoria la variable dependiente (dimensión cognitiva), ya que su principal aplicación para relacionar información es a partir de la tendencia de variables independientes.

Verificación de supuestos Se hace la verificación de supuestos del modelo normal, en donde se observa que los residuales del modelo sólo cumplen el supuesto de varianza constante, los supuestos de normalidad e independencia en los residuales no se cumplen.

- **Prueba de Normalidad**

  H_0 : Los residuales siguen normalidad
  H_1 : Los residuales no siguen normalidad

Tabla 2: Prueba de normalidad

Shapiro-Wilk normality test
W = 0.93584, p-value = 4.531e-05

- **Prueba de varianza constante**

  H_0 : Los residuales tienen varianza constante
  H_1 : Los residuales no tienen varianza constante

Tabla 3: Prueba de varianza constante

Non-constant Variance Score Test
Chisquare = 0.7650847 , Df = 1
p-value = 0.381742

- **Prueba de independencia**

  H_0 : Los residuales son independientes
  H_1 : Los residuales no son independientes

Tabla 4: Prueba de independencia

Autocorrelation D-W
D-W = 1.479794, Statistic = 1.479794
p-value = 0.01

5.4. Resultados del modelo semiparamétrico para los diferentes tipos de errores

Según las gráficas explicadas en el análisis factorial, se observo que la , dimension afectiva tenia una comportamiento lineal, mientras que la dimensión expresiva no presentaba ningun comportamiento lineal ni una tendencia, por esta razón se hará la aplicación del modelo semiparamétrico, ya que al tener una parte lineal y una no lineal se podrá ajustar el modelo planteado con los diferentes tipos de errores explicados en la sección tres, así se podrá dar explicación a la dimensión cognitiva con una parte lineal y una funcional.
Para la estimación se utilizará la función **gesm** del paquete **BayesGESM** que funciona en el software estadístico **R**, la función tiene la siguiente descripción:

Gesm se utiliza para obtener la inferencia estadística basada en el enfoque Bayesiano para modelos de regresión bajo el supuesto de que los errores aditivos independientes siguen una escala de mezclas de distribución normal (es decir, normal, Student-t, slash, normal contaminado, Laplace y distribución hiperbólica simétrica) , donde tanto la ubicación como los parámetros de dispersión de la distribución de la variable de respuesta incluyen componentes aditivos no paramétricos descritos por B-splines. Los parámetros se estiman implementando un algoritmo MCMC eficiente combinando el muestreador de Gibbs y el algoritmo de Metropolis-Hastings, este proceso de estimación lo hace la función automáticamente, obteniendo así el mejor parámetro con que se llegará a tener el mejor ajuste del modelo y el menor DIC.

Se obtienen los siguientes resultados:

DISTRIBUCIÓN	*DIC*
NORMAL CONTAMINADA	207.1435
NORMAL	204.9202
T-STUDENT	193.8102
HIPERBÓLICA	191.9921
SLASH	190.2907

Tabla 5: DIC para los modelos semiparamétricos

Según la tabla 5 el modelo semiparamétrico ajustado con el tipo de error Slash, obtuvo la mejor estimación con un menor DIC.
A continuación solo se presentará las gráficas del ajuste del modelo semiparamétrico ajustado con tipo de error Slash, sus pruebas y cadenas de convergencia bayesiana, para consultar las gráficas de los diferentes tipos de errores con sus pruebas de convergencia ver el apéndice.

5.5. Modelo semiparamétrico con tipo de error Slash multivariado

Se tienen la siguiente prueba de convergencia del modelo semiparamétrico con tipo de error Slash multivariado: En el que se utilizaron 100.000 iteraciones, se quemaron 1000 de calentamiento y se tomaron cada 100.

Diagnóstico de convergencia de Heidelberger y Welch's

```
     Stationarity start     p-value
     test         iteration
var1 passed       1         0.388

     Halfwidth Mean   Halfwidth
     test
var1 passed    -0.221 0.00715
```

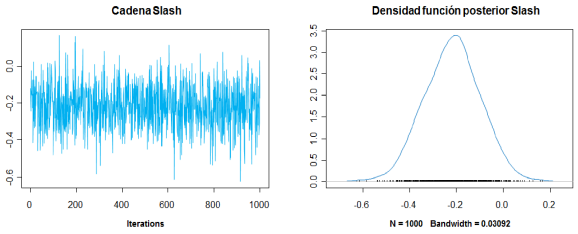


Figura 3: Prueba de convergencia

Según la prueba de convergencia la cadena es estacionaria y es constante en su media.

En la *figura 4* se observa el ajuste del modelo semiparamétrico con tipo de errores Slash multivariados, que obtuvo el mejor DIC, el ajuste de la regresión semiparamétrica esta dentro de las bandas de confianza, según el aprecio que tenga el estudiante por las matemáticas en cuanto a su dimensión afectiva y expresiva tendrá un puntaje alto en la dimensión cognitiva, correspondiente a la calificación final, se presenta un crecimiento no lineal ya que las dos covariables que se utilizaron, no tenian en su totalidad una tendencia marcada, lo que se ve reflejado en el ajuste del modelo resultante.

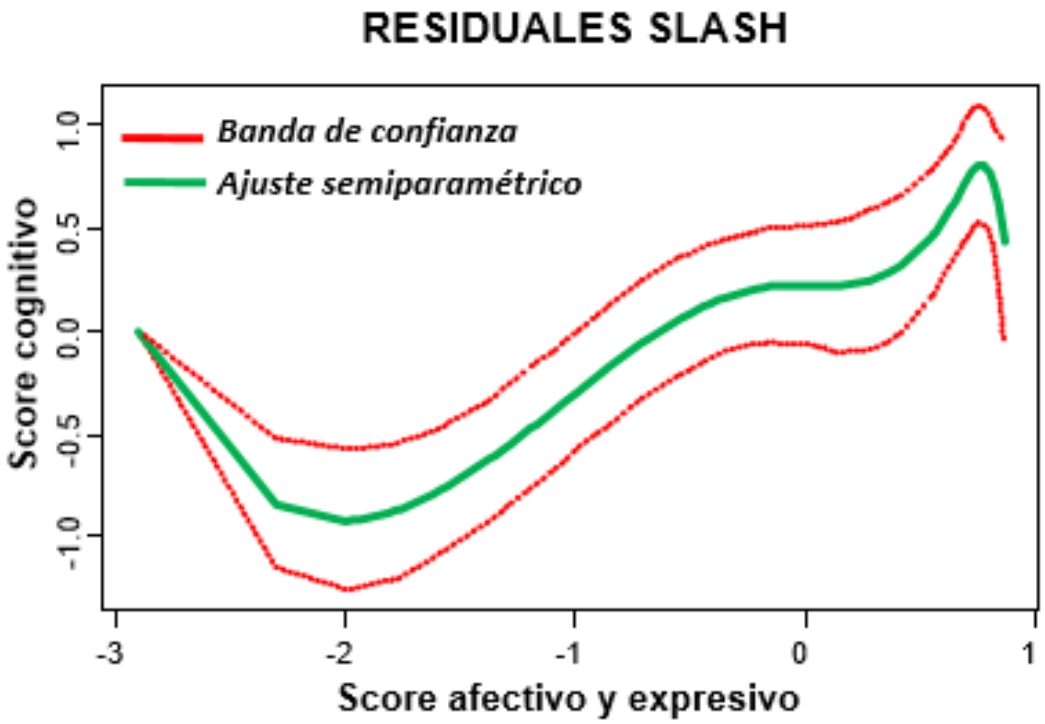


Figura 4: Modelo semiparamétrico con errores Slash multivariados

Discusión

El enfoque y desarrollo de este trabajo consiste en demostrar o refutar por medio de un método estadístico, que la dimensión cognitiva puede ser explicada por las dimensiones afectiva y expresiva para las pruebas de matemáticas tipo ICFES para un colegio de Bogotá, la aplicación estadística fue, ajustar un modelo semiparamétrico con distintos tipos de errores, ya que se pretende conocer el modelo con tipo de error que mejor se ajuste.
Durante mucho tiempo se pensó que se debía tener una competencia por cada dimensión del conocimiento (expresiva, afectiva y cognitiva), pero con el tiempo se ha propuesto que los conocimientos se desligan en una sola dimensión (Cognitiva), la mente humana cuando aprehende lo hace porque conjuga y estructura de forma coherente y sistemática las tres dimensiones de la mente, sin duda este ha sido uno de los temas más importantes en cuanto a la metodología de aprendizaje y el desarrollo académico.

Con lo obtenido en los resultados se demuestra estadísticamente que la dimensión cognitiva se ve afectada por los cambios o habilidades de la dimensión expresiva y afectiva para el contexto explicado anteriormente, refutando de esta manera la teoría académica sobre el desarrollo de las competencias por medio de las tres dimensiones de conocimiento.

Para tener un buen desempeño en la dimensión cognitiva en importante el interés que se pueda desarrollar en los estudiantes, en cuanto a la dimensiones expresiva y afectiva encargas del afecto que se pueda tener por el área de conocimiento en estudio, mejorando el score en las dimensiones nombradas se podrá llegar a tener mejores resultados en cuanto a las calificaciones de matemáticas tipo ICFES de los estudiantes de grado once.

Futuros estudios

Como futuros estudios se recomienda la aplicación de esta metodología estadística a diferentes áreas del conocimiento, y, con aportes de información de entes como el ICFES o administrativos encargados de la calificación de pruebas académicas, aplicar el modelo semiparamétrico y con base en los resultados obtenidos, poder implementar una política pública que permita mejorar la calidad de la educación, basándose en el desarrollo de las tres dimensiones analizadas en este trabajo.

Agradecimientos

Agradezco a mis directores de grado por su apoyo incondicional que me brindaron durante este proceso de mi carrera, especialmente al profesor Wilmer Pineda por su disposición, compromiso y dedicación, sus grandes aportes de conocimiento y experiencia fueron importantes al momento del desarrollo de esta investigación. También agradezco a mi familia, amigos y conocidos, principalmente a mi novia y fiel compañera Linnethy Julieth Lambraño Pérez, quien me acompaño y apoyo en este logro tan importante de mi vida.

Referencias

1. Andrews, D. F. Mallows, C. L. (1974), 'Scale mixtures of normal distributions',Journal of the Royal Statistical Society. Series B (Methodological)pp. 99-102.
2. Fonseca Gómez, Lida Rubiela. Aproximación de indicador de logro estudiantil.2012.
3. Garcia, Diana (2017), Modelo espacial con acercamiento bayesiano para el estudio de victimas en Colombia en 2017 :http://repository.usta.edu.co/handle/11634/178/recent-submissions?offset=0
4. Keele, Luke John. Semiparametric regression for the social sciences. John Wiley Sons, 2008.
5. Olaya, J. (n.d.). Modelación no paramétrica de la contaminación promedio octohoraria del aire debida al Monóxido de Carbono y al Ozono Troposférico. Scielo.
6. Herrera Cabezas, Andrea, Restrepo Álvarez, María Fernanda, Uribe Rodríguez, Ana Fernanda, López Lesmes, Claudia Natalia, Competencias académicas y profesionales del psicólogo. Diversitas: Perspectivas en Psicología [en linea] 2009, 5 (Junio-Diciembre).
7. Rondón Bolfarine (2015), Semi-parametric regression models using R
8. Ruppert, D.; Wand, M. P.; Carroll, R. J. Cambridge series in statistical and probabilistic mathematics: Semiparametric Regression. 2003.
9. Toquica, Sayuri (2017), Aproximación bayesiana de un modelo semiparamétrico: http://repository.usta.edu.co/handle/11634/178/recent-submissions?offset=0
10. Vanegas Hernando Luis, L. M. (2016). Semi-parametric regression models using R. Departamento de Estadística: Universidad Nacional de Colombia.

6. Apéndice

6.1. Modelo semiparamétrico con tipo de error normal multivariado

se tienen el siguiente ajuste y resultados del modelo semiparamétrico con tipo de error normal multivariado:

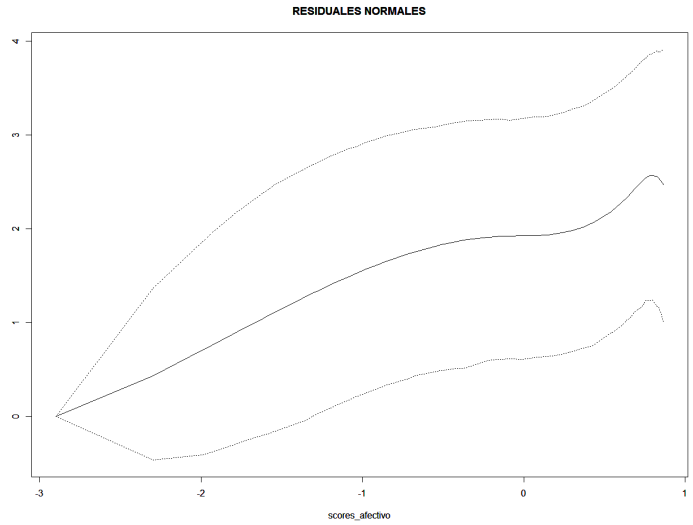
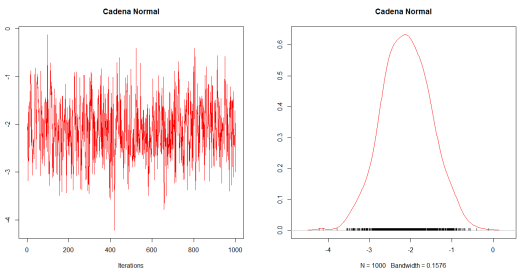


Figura 5: Modelo semiparamétrico con errores normales multivariados

Cadena de convergencia errores normal multivariado

```
     Stationarity start     p-value
     test         iteration
var1 passed       1         0.224

     Halfwidth Mean Halfwidth
     test
var1 passed    -2.1 0.0578
```




```
DIC1

[1] 204.9202
```

6.2. Modelo semiparamétrico con tipo de error T-student multivariado

se tienen el siguiente ajuste y resultados del modelo semiparamétrico con tipo de error T-student multivariado:

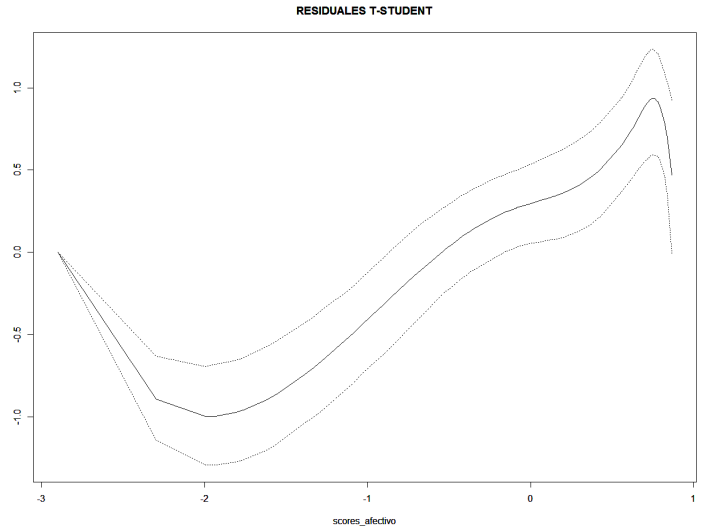
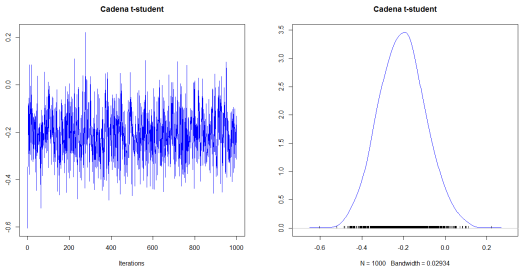


Figura 6: Modelo semiparamétrico con errores T-student multivariados

Cadena de convergencia errores T-student multivariado

```
     Stationarity start     p-value
     test         iteration
var1 passed       1         0.781

     Halfwidth Mean   Halfwidth
     test
var1 passed    -0.206 0.00692
```




```
DIC2
```

```
[1] 193.8102
```

6.3. Modelo semiparamétrico con tipo de error normal contaminada multivariada

se tienen el siguiente ajuste y resultados del modelo semiparamétrico con tipo de error normal contaminada multivariada:

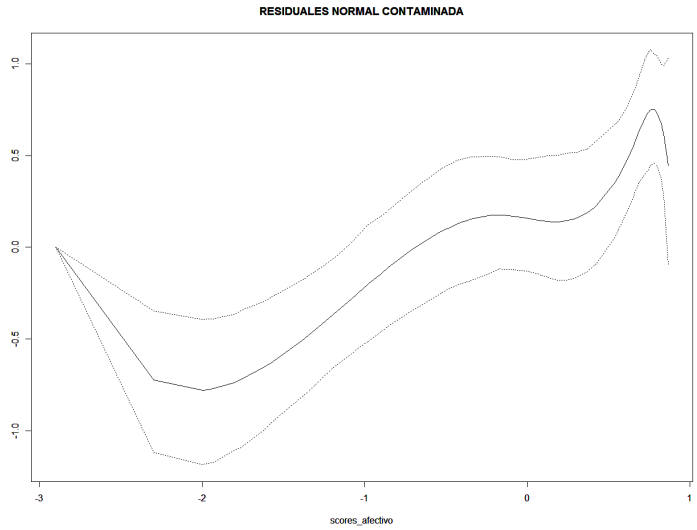
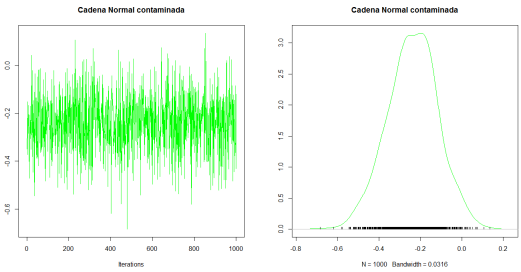


Figura 7: Modelo semiparamétrico con errores normales contaminados multivariados

Cadena de convergencia errores normal contaminado multivariado

```
     Stationarity start     p-value
     test         iteration
var1 passed       1         0.904

     Halfwidth Mean   Halfwidth
     test
var1 passed    -0.238 0.0075
```




```
DIC4
```

```
[1] 207.1435
```

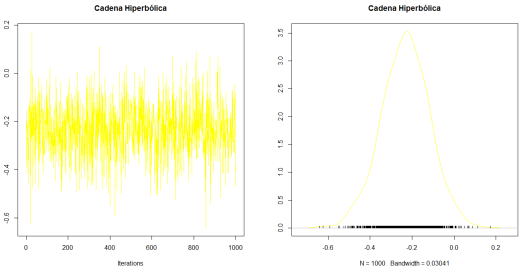
6.4. Modelo semiparamétrico con tipo de error hiperbólicos simétricos multivariados

se tienen el siguiente ajuste y resultados del modelo semiparamétrico con tipo de error hiperbólico simétrico multivariado:

Cadena de convergencia errores hiperbólicos simétricos multivariados

```
     Stationarity start     p-value
     test         iteration
var1 passed       1         0.881

     Halfwidth Mean   Halfwidth
     test
var1 passed    -0.232 0.00708
```



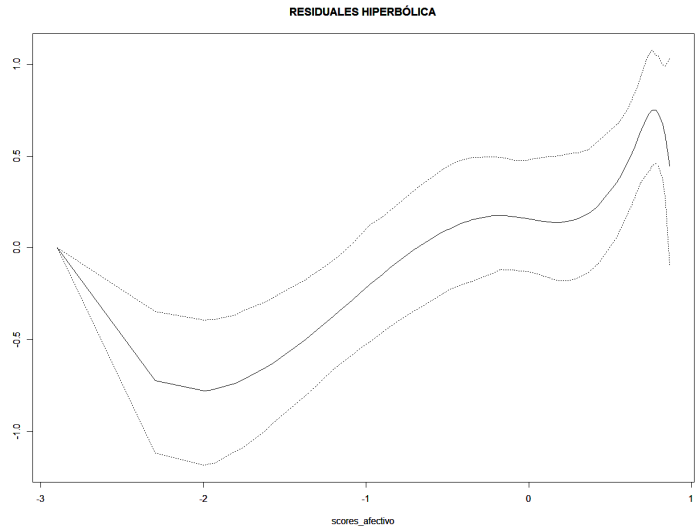


Figura 8: Modelo semiparamétrico con errores hiperbólicos simétricos multivariados

```
DIC5
[1] 191.9921
```