ónΩón

UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Proyecto de Grado
Abril 2020

Modelo de curvas de crecimiento: Modelos lineales mixtos vs Datos funcionales

Nancy Patricia Balaguera Carrasquilla [a]
nancybalaguera@usantotomas.edu.co

Wilmer Pineda Rios[b]
wilmerpineda@usantotomas.edu.co

Resumen

Se realiza una comparación entre el uso de modelos lineales mixtos, explícitamente de curvas de crecimiento para el análisis de datos longitudinales contra una aplicaciones de datos funcionales por medio de la ANOVA funcional, se quiere evidenciar cambios en el comportamiento de los datos a través del tiempo así como que tanto afecta el nivel en que se encuentre el paciente en este punto. La base de datos que se usó para el desarrollo de este trabajo fue tomada de (Mirman, 2014). Este conjunto de datos contiene información del Proyecto de Psicolingüística de Afasia de Moss (Mirman et al., 2010), los resultados indican que no hay diferencia significativa en las respuestas de los pacientes a la prueba aplicada de acuerdo al nivel de diagnostico en el que se encuentra su enfermedad, independientemente de la metodología usada los resultados no varían, además de esto se evidencio que computacional mente es mas económico realizar este tipo de análisis por medio de los datos convertidos a funciones además los resultados son gráficamente mas fáciles de comprender.

Palabras clave Curva de crecimiento, modelos lineales mixtos, datos funcionales, ANOVA Funcional.

Abstract

A comparison is made between the use of mixed linear models, explicitly of growth curves for the analysis of longitudinal data against an application of functional data by means of functional ANOVA, we want to show changes in the behavior of the data over time so as it affects how much the patient is at this point. The database that was used for the development of this work was taken (Mirman, 2014). This data set contains information from the Moss Aphasia Psycholinguistics Project (Mirman et al., 2010), the results indicate that there is no significant difference in the patients' responses to the test applied according to the level of diagnosis at which Their disease is found, regardless of the methodology used the uncontrolled results, in addition to this it is evidenced that computationally it is more economical to perform this type of analysis by means of data converted to functions in addition to the results are graphically easier to understand.

Keywords Growth curve, mixed linear models, functional data, Functional ANOVA.

[a]Estudiante pregrado Estadística, U. Santo Tomás, sede Bogotá
[b]Docente Facultad de Estadística, U. Santo Tomás, sede Bogotá

1. Introducción

El estudio longitudinal, a diferencia del estudio transversal, permiten el seguimiento de los mismos individuos a través del tiempo y de sus generaciones precedentes y siguientes, este tipo de investigación está relacionada en su gran mayoría con estudio de tipo conductual, en la investigación que se desarrolló a lo largo de este trabajo se encontraron algunas aplicaciones que se han dado desde la metodología funcional y modelos mixtos de curvas de crecimiento de manera independiente.

Desde el punto de vista funcional se han desarrollado técnicas estadísticas para estos análisis, en la investigación se encontró el desarrollo de una tesis de maestría en la Universidad Nacional de Colombia donde, el objetivo básico de investigación en potenciales evocados era detectar diferencias en las respuestas bioeléctricas del cerebro que estén relacionadas con la producción de procesos comportamentales o psicológicos tales como percepción, motivación o aprendizaje y que representen campos de acción al detectar tempranamente cualquier atipicidad presentada, por medio de el desarrollo de una nueva técnica para el análisis espacial de la ANOVA funcional. Esta tesis fue guía útil para el desarrollo del trabajo de grado.

La idea general de este trabajo de grado es comparar un modelo de curva de crecimiento el cual no es más que un generalización de un modelo mixto, teniendo en cuenta un polinomio característico de acuerdo al comportamiento de los datos, que para el ejemplo a desarrollar será de forma lineal, contra una prueba ANOVA funcional que tienen como propósito detectar diferencias entre los diagnósticos de cada paciente. El objetivo principal es identificar las diferencias existentes entre las dos metodologías aplicadas e identificar cual es la mas acertada para este conjunto de datos.

2. Datos Longitudinales

Los datos longitudinales son observaciones medidas a una misma población en dos o más ocasiones a través del tiempo, este proceso observa y analiza la evolución de un fenómeno dado con el fin de estimar incidencias y anticipar riesgos. Este tipo de estudios traen consigo algunas limitaciones que hacen que su precisión a la hora de estimar pueda verse afectada y es que la variable de investigación bajo la que el fenómeno se repite puede no ser controlada por el investigador, por tanto, los resultados se verán altamente afectados.

El modelamiento de este tipo de datos hace uso de modelos lineales y su enfoque está basado en el análisis de la varianza, los modelos tradicionales tienen una desventaja y es que se requiere que los datos sean balanceados lo que, en contexto aplicado, es difícil de conseguir. Por esto, se han desarrollado modelos alternativos como el estudio de curvas de crecimiento, del que se han derivado gran cantidad de métodos. Todos estos métodos, además de modelar la varianza entre e intra individuos, no requieren datos balanceados. En la actualidad, se aplican los modelos lineales mixtos como una alternativa global de análisis.

Existen algunas dificultades al realizar análisis de datos longitudinales, por un lado, el análisis suele ser más complejo debido a la dependencia que suele darse entre las medidas repetidas de la misma variable observada. por otro lado, no es posible controlar todas las circunstancias bajo las que obtiene las medidas repetidas, y esto lleva a que en la mayoría de los casos los datos no sean balanceados o estén incompletos.

Para el análisis de datos longitudinales, existen diferentes metodologías. una de las formas para determinar cuál es el procedimiento correcto a realizar es identificando la distribución de la variable respuesta, si esta es normal o métrica, las técnicas de análisis multivariante, análisis de curvas de crecimiento, modelos de efectos mixtos y los modelos de ecuaciones estructurales son los más adecuados para la buena manipulación de los datos (Liang y Zeger, 1986). Por otro lado, cuando la variable no se distribuye de forma normal, es posible usar los modelos log-lineales y los modelos basados en las ecuaciones de estimación

generalizadas.

3. Modelo de Curva de Crecimiento

Una curva de crecimiento es una representación gráfica de cómo una cantidad particular aumenta con el tiempo. Las curvas de crecimiento se usan en estadísticas para determinar el tipo de patrón de crecimiento de la cantidad, ya sea lineal, exponencial o cúbico.

Los métodos de regresión modelan explícitamente el tiempo como una variable continua, que es la forma natural de cuantificar el cambio en el tiempo, una ventaja de estos modelos es que cuantifican simultáneamente los patrones a nivel grupal y a nivel individual, en un modelo multinivel de deben tener dos aspectos claros los cuales son los coeficientes de las variables y un modelo del que parte cada uno de ellos.

3.1. Estructura del modelo de curva de crecimiento.

Para presentar la estructura básica de un modelo de curva de crecimiento, comencemos con el caso lineal simple ilustrado en la ecuación 1.

$$Y = \beta_0 + \beta_1 * Time \tag{1}$$

Donde β_0 es el intercepto, es decir el valor que toma Y cuando $Time$=0 y β_1 es la pendiente la cual representa el cambio promedio en Y para cada unidad de $Time$. En el contexto de un modelo de regresión, esa relación general se elabora para convertirse en un modelo de las observaciones individuales Y_{ij} para el individuo i en el $Time$ j como se observa en la ecuación 2:

$$Y_{ij} = \beta_{0i} + \beta_{1i} * Time_j + \varepsilon_{ij} \tag{2}$$

Donde ε_{ij} es el error residual, es decir, la cantidad en que la observación real Y_{ij} difiere del valor estimado; generalmente, se supone que los errores residuales son (IID), lo que significa que todos los ε_{ij} provienen de la misma distribución y que cualquier valor de ε_{ij} en particular no proporciona información sobre otros valores de ε_{ij}.

Ahora realizamos un ejemplo ilustrado en la ecuación 3 asumiendo 2 condiciones: una condición de control que servirá como base y una condición experimental que llamaremos condición C, esto sera el modelo de nivel 2 para el coeficiente β_{0i} del modelo de nivel 1:

$$\beta_{0i} = \gamma_{00} + \gamma_{0C} * C + \zeta_{0i} \tag{3}$$

Donde γ_{00} es la linea base de β_{0i}, $0C$ es el efecto fijo (también llamado a veces efecto estructural) de la condición C en la intersección, y $0i$ es la desviación aleatoria (también llamada residual o estocástica) de esa línea base para el individuo i. Este modelo de nivel 2 hace dos cosas.

Primero, permite que la condición C tenga una interceptación única que es diferente de la interceptación de referencia. Conceptualmente, esto es más o menos como una prueba t que compara la condición C con la línea de base y, típicamente, estos son los efectos de interés primario.

En segundo lugar, define una estructura para la variación aleatoria en la intersección: dice que todas las observaciones dentro del conjunto indexado por i deberían tener la misma intersección, que puede

diferir para otros valores de i. En otras palabras, la estructura de efectos aleatorios captura la estructura anidada de los datos al especificar que todas las observaciones indexadas por un valor particular de i corresponden a una sola intersección, que es diferente de las intersecciones para otros valores de i. En un caso típico, se refiere a participantes individuales en un estudio y ζ_{0i} variaría aleatoriamente entre los participantes.

Para el análisis de los modelos de curva de crecimiento las estimaciones de parámetros proporcionadas por el resumen no tienen valores p, para este tipo de modelos, la mejor prueba estadística es la comparación de modelos anidados, que se realiza por medio de la función ANOVA.

3.2. Modelos lineales mixtos

Torres-Saavedra (2018) dan lugar a que el objetivo de un modelo mixto es el análisis de una variable aleatoria, su valor esperado y su variabilidad, esto, dicho en otras palabras, lo que intenta es modelar un efecto fijo a todos los sujetos de estudio y otro efecto aleatorio asociado a cada uno de los sujetos el cual esta representado por la ecuación 4:

$$Y_{ij} = (\beta_0 + \beta_1 t_{ij}) + (b_{i0} + b_{1i} t_{ij}) + \varepsilon_{ij}. \tag{4}$$

3.3. Efectos Fijos y Efectos Aleatorios

Michael Clark (2018) afirma que este tipo de efectos se evidencia generalmente en los modelos mixtos, los efectos fijos son los coeficientes de regresión que se tienen en los enfoques de modelado estándar y los efectos aleatorios se entiende como las variables que no puedo controlar y que le generar un efecto particular a cada individuo.

Adicional argumenta que los efectos aleatorios permiten que cada grupo tenga su propio efecto único además del efecto fijo general. Esto es simplemente una desviación aleatoria, casi siempre distribuida normalmente en la práctica, de la intersección general y las pendientes. Los modelos mixtos son un enfoque equilibrado entre ignorar estas contribuciones únicas y contextualizar en exceso ejecutando regresiones separadas para cada grupo.

4. Análisis de datos funcionales (ADF)

Cuando se habla de datos longitudinales se entienden como se vio en la sección anterior, medidas repetidas para un mismo individuo, que generalmente son representados a través de funciones, de aquí nace la necesidad de adaptar las medidas estadísticas a estas funciones para su análisis.

Ferraty and Vieu (2006) definen una variable aleatoria funcional , como una variable aleatoria que toma valores en un espacio de funciones (espacio funcional). Una observación x de la variable aleatoria se denomina dato funcional.

Un conjunto de datos funcionales $x1, x2, ..., xn$ es la observación de n variables funcionales distribuidas como x. Un dato funcional $xi(t), t \in T = [a, b] \subset R$, es representado usualmente como un conjunto finito de pares $(tj, xij) tj \in T, j = 1, 2, ..., M$, donde M representa la cantidad de puntos en los cuales es observada la variable de interés $yij = i(tj)$.

4.1. Suavizado de funciones mediante funciones B-splines

(Ramsay and Silverman, 2005). Desarrollaron una herramienta no paramétrica de mucha utilidad en el análisis de datos funcionales, que consiste en el suavizado de curvas a través de funciones, a pesar de que existen varias técnicas para realizar este tipo de procedimiento para el desarrollo de este trabajo se hizo por medio de la base de B-splines.

Un spline se define como una curva diferenciable definida en porciones mediante polinomios, son principalmente útiles en los procesos de interpolación requiriendo polinomios de bajo grado para su estimación evitando fuertes oscilaciones que se podrían causar por polinomios de alto grado, ello hace que un polinomio pueda coincidir con una función en muchos puntos.

Dado un conjunto de L puntos interiores del intervalo $T = [a, b]$, es decir, $a = 1 < 2 < ... < L = b$, un spline cubico es una función S definida sobre T, tal que S es un polinomio cubico en el intervalo $[\tau_l 1, \tau_l], l = 1, 2, 3, ..., L + 1$. (Ramsay and Silverman, 2005).

Luego de que los datos están suavizados y convertidos en funciones correctamente la técnica de análisis a usar es la regresión funcional. La extensión del modelo lineal al caso funcional se presenta cuando existe por lo menos una variable (respuesta o explicativa) funcional en dicho modelo.

4.2. ANOVA funcional

Asumiendo que se tienen G tratamientos cada uno con un numero n_g de sujetos y que $y(t)$ es una respuesta funcional, el modelo lineal para la i-esima función (curva respuesta) en el g-esimo grupo yig(t), como se muestra en la ecuación 5: (Ramsay and Silverman, 2005)

$$Y_{ig} = \mu(t) + \alpha_g(t) + \varepsilon_{ig}(t) \tag{5}$$

donde la función $\mu(t)$ es la media general, $\alpha_g(t)$ representa la función media para cada tratamiento y $\varepsilon_{ig}(t)$ es la función de error en cada caso.

En términos matriciales el modelo puede expresarse como:

$$Y_{ig} = \sum_{j=1}^{G+1} x_{ig} j \beta_j(t) + \varepsilon_{ig}(t) \tag{6}$$

$$Y(t) = \mathbf{X}(t)\beta(t) + \epsilon(t) \tag{7}$$

Puede observarse que la ecuación (6) es una extensión natural de la suma de cuadrados del error (SSE) del modelo lineal clásico.

Al igual que en el modelo lineal múltiple, la fuente primaria de información para investigar la importancia de los tratamientos, es la función de suma de cuadrados residual:

$$SSE(t) = \sum_{ig} (y_{ig}(t) - \mathbf{X}_{ig}\beta(t))^2 \tag{8}$$

Dicha función es comparada con la función de suma de cuadrados residual obtenida al utilizar solo la media general (t) en el modelo:

$$SSEY(t) = \sum_{ig} (y_{ig}(t) - \mu(t))^2 \tag{9}$$

Ası un camino para realizar la comparación, se establece utilizando la función del cuadrado de correlación múltiple:

$$R^2(t) = \frac{(SSY(t) - SSE(t))}{SSY(t)} \tag{10}$$

Por tanto la función de cuadrado medio de la regresión es la diferencia entre la suma de cuadrados total (SSY (t)) y la suma de cuadrados del error (SSE(t)), dividida por la diferencia entre los grados de libertad del error para los dos modelos (grados de libertad de la regresión):

$$CMR(t) = \frac{SSY(t) - SSE(t)}{gl} \tag{11}$$

Finalmente se puede construir la función F como:

$$F(t) = \frac{CMR(t)}{CME(t)} \tag{12}$$

5. Resultados y Discusión

La base de datos NamingRecovery se tomo de(Mirman et al., 2010). Este conjunto de datos contiene información del Proyecto de Psicolingüística de Afasia de Moss (Mirman et al., 2010). Se analiza el cambio en la proporción de respuestas de nombres de imágenes que fueron errores semánticos (como decir caballo a una imagen de una vaca) para un grupo de pacientes afásicos. La afasia es el trastorno del lenguaje que se produce como consecuencia de una patología cerebral, se trata de la pérdida de capacidad de producir o comprender el lenguaje, debido a lesiones en áreas cerebrales especializadas en estas funciones, en otras palabras es una pérdida adquirida en el lenguaje oral.

Cada paciente completó la prueba de denominación de imágenes cinco veces en el transcurso de varias semanas (véase también Schwartz y Brecher, 2000). Los pacientes se agrupan por subtipo de afasia: anómica ($N = 6$) esta aparece solo cuando la dificultad de encontrar palabras de uso común aparezca de forma relativamente aislada, conducción ($N = 9$) esta la comprensión es casi normal; pero la fluidez queda gravemente comprometida debido a problemas en la producción de palabras aislada; convirtiéndose así en una habla secuencial y de oraciones cortas y por ultimo la afasia de Wernicke ($N = 8$) se caracteriza por un habla fluida pero con un gran número de sustituciones y parafasias, junto con dificultades en la comprensión.

Para comenzar se hace una descripción general de los datos, en la figura 1. se muestra la relación de cada tipo de diagnostico en referencia al porcentaje de error semántico con cada administración de prueba.

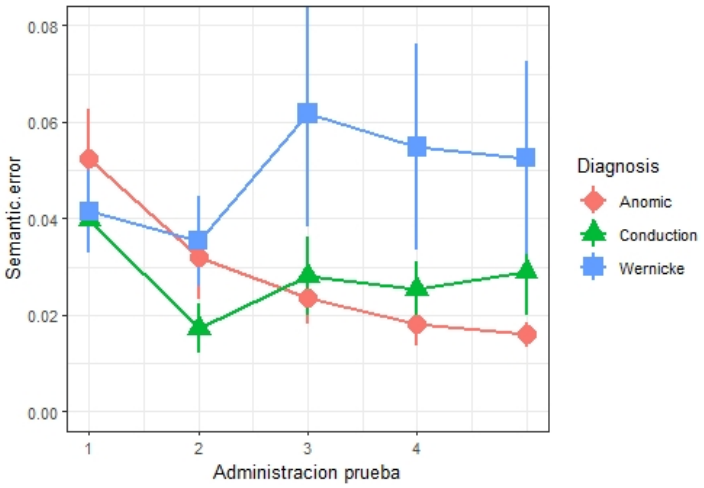

Figura 1: Proporción de errores semánticos en la denominación de imágenes para tres grupos de pacientes afásicos en cinco administraciones de prueba.

A simple vista se puede concluir que la proporción de errores semánticos tiende a disminuir para los pacientes afásicos de diagnostico Anomic, se mantiene casi igual para los pacientes afásicos de diagnostico conducción y aumenta para los pacientes afásicos de diagnostico Wernicke.

Se ajustará un modelo usando el análisis de curva de crecimiento. Dado que solo se tienen cinco observaciones por participante, el uso de un polinomio de orden superior podría realizar un sobre ajuste de los datos, por lo que se usara un análisis de curva de crecimiento de primer orden es decir de forma lineal.

	Estimate	Std. Erros	t value
(Intercept)	0.040867500	0.006734605	6.0682847
TestTime	0.004168750	0.003051837	1.3659804
DiagnosisWAnomic	0.004899167	0.010287279	0.4762354
DiagnosisWConduction	-0.010249722	0.009255829	-1.1073802
TestTime:DiagnosisWAnomic	-0.012853750	0.004661758	-2.7572750
TestTime:DiagnosisWConduction	-0.005545417	0.004194349	-1.3221161

De manera general se da la interpretación a la salida anterior tomando como base el diagnostico del paciente con afasia de diagnostico Wernicke estos pacientes tienen un habla fluente, pero su discurso carece de contenido, por una inadecuada selección de las palabras. Las dificultades de estos pacientes se centran en la comprensión del lenguaje, de manera que el resto de las habilidades cognitivas se encuentran conservadas de lo cual se concluye:

Todos los pacientes diagnosticados con Afasia tipo Wernicke, tienen un intercepto de 0.0408 lo cual hace referencia a el error de semántica medio que tiene cada paciente al inicio del proyecto, luego se puede identificar el efecto que tiene el tiempo $TestTime$ que sera cada uno de los tiempos (1-5) en que se le realizara la prueba de denominación de imágenes para cada paciente, este coeficiente es 0.00416 lo cual indica que el error de semántica aumentara respecto al tiempo que transcurra aplicando la prueba, también se identifican los coeficientes para cada uno de los diagnósticos de afasia que presentan los pacientes y sus coeficientes son (Anomic - Wernicke) = 0.00489 por lo cual se afirma que para los pacientes con diagnostico Anomic sera mayor el error de semántica medio y (Conduction - Wernicke) = -0.10249, entonces el diagnostico Conduction disminuirá esta media de error semántico.

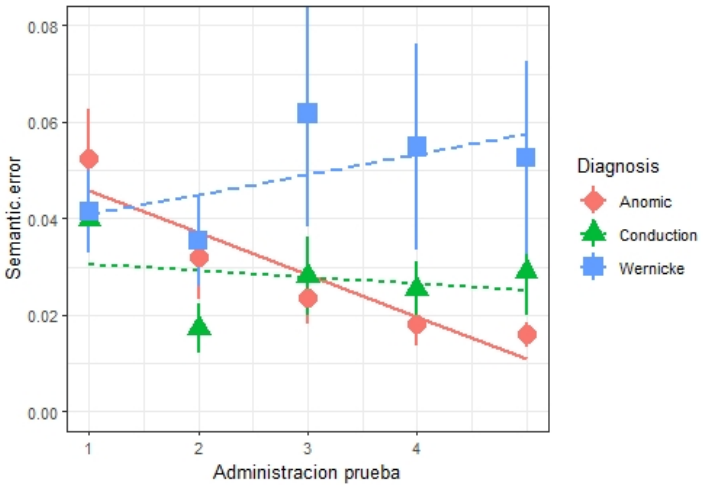

Figura 2: Proporción de errores semánticos en la denominación de imágenes para tres grupos de pacientes afásicos en cinco administraciones de prueba.

Hasta el momento solo se han evaluado las variables de forma independiente pero al revisar la interacción de ellas y tener en cuenta los efectos aleatorios de los pacientes de evidencia que el coeficiente $TestTime : DiagnosisWAnomic$ = -0.010249722, lo cual quiere decir que teniendo en cuenta el tiempo transcurrido en las aplicaciones de la prueba y el diagnostico del paciente, para un paciente con diagnostico Anomic en el transcurso del tiempo el error semántico disminuirá para cada paciente teniendo en cuenta su efecto aleatorio asociado, lo cual se observa claramente en la figura 2.

En la gráfica anterior se evidencia el comportamiento de la curva de crecimiento para el conjunto de datos, notese que ninguna recta logra captar la información para cada uno de los individuos en los diferentes tiempos de la administración de la prueba, en términos generales se puede decir que el ajuste de este modelo para estos datos no es bueno.

Ahora veremos gráficamente el comportamiento de los efectos aleatorios de cada uno de los individuos.

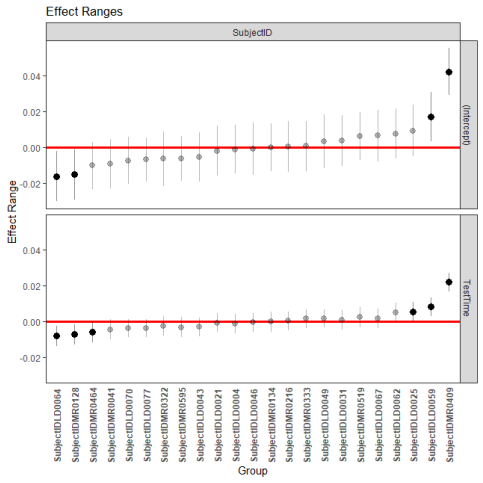

Figura 3: Efectos aleatorios del modelo de curva de crecimiento

En la gráfica anterior se muestra la distribución de los efectos aleatorios de cada uno de los pacientes con afasia, se observa que el paciente identificado con ID MR0595 presenta un efecto positivo muy por encima de los demás y esto hace que su % de semantica.error en la prueba aumente con respecto a los demás manteniendo las demás variables constantes.

Desde la segunda metodología a probar, se realizo un suavizamiento de los datos para convertirlos en funciones y poder ser analizados desde este punto de vista, por lo cual se realizo una prueba ANOVA funcional para evidenciar si hay diferencias entre cada uno de los diagnósticos de los pacientes.

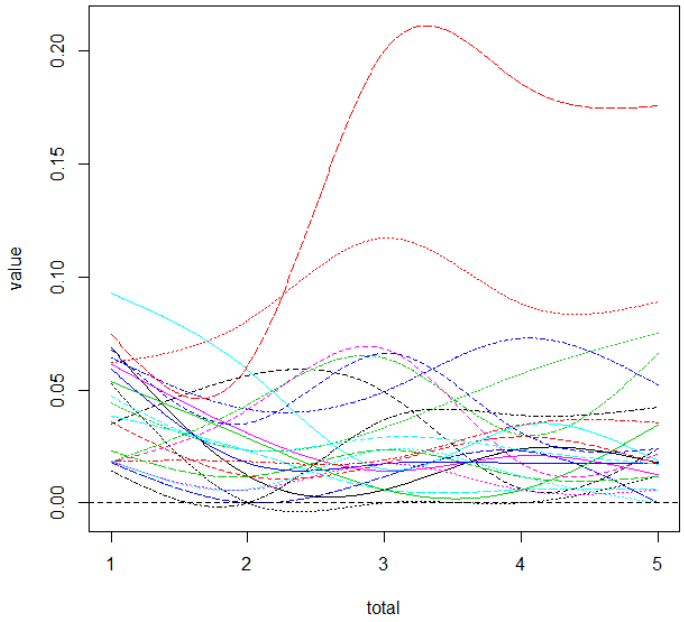

Figura 4: % Semantic.error convertidos a funciones

En el gráfico anterior se evidencian los datos de los pacientes después de realizar el suavizado por medio de splines cubicos, dando lugar a funciones representadas en curvas:

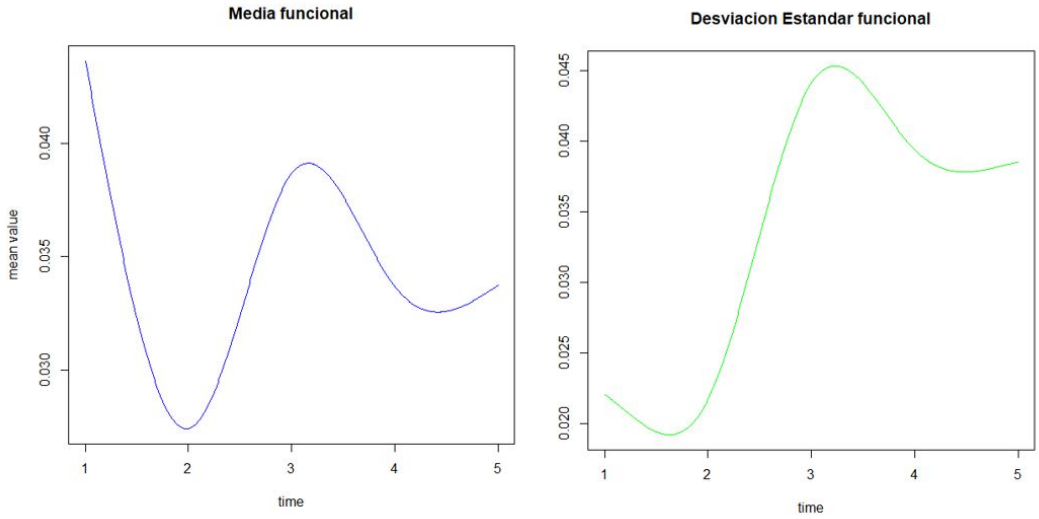

Figura 5: curvas

Para efectos descriptivos se muestra la media funcional, de la cual se puede concluir que para el suministro de prueba n2 los pacientes bajaron considerablemente su % de semantic error pero subieron de nuevo en el suministro de prueba n 3, además se observa que el % mas alto alcanzado estuvo al comienzo de la prueba en su primer suministro. En cuanto a la desviación estándar para la prueba n2 es muy baja es decir la variación de la media registrada no cambia mucho entre individuos y para la prueba n3 pasa lo contrario a pesar de que su media registrada no es tan alta esta puede varias demasiado entre individuos.

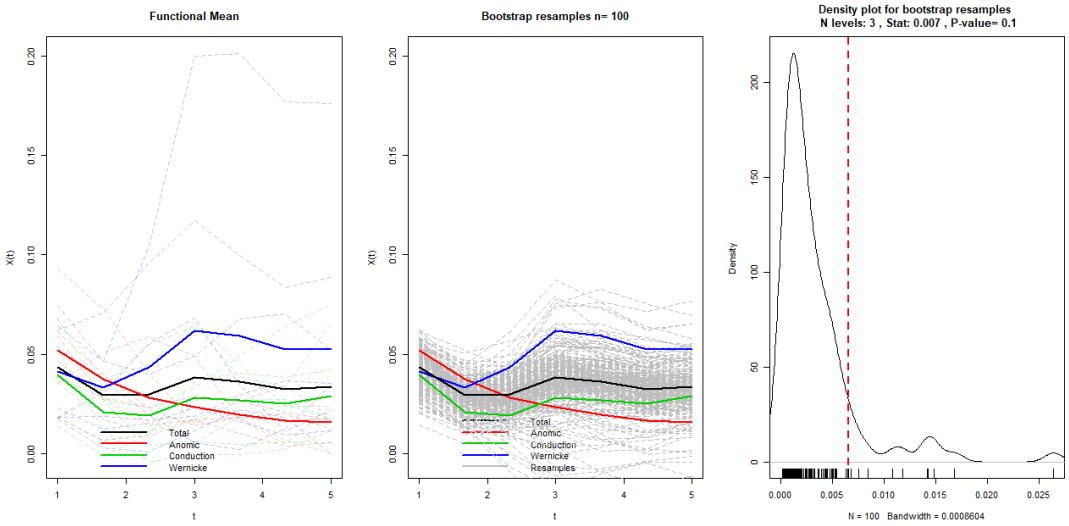

Figura 6: curvas

Por ultimo tenemos la salida del ANOVA funcional la cual se realizo por medio de una distribución bootstrap que se construye por medio de iteraciones dadas, en la cual se evidencia que no hay diferencia significativa en el diagnostico de los pacientes ya que el p-valor = 0.1 por lo tanto no hay suficiente evidencia estadística para rechazar la hipótesis nula, la cual indica que los diagnósticos de los pacientes son iguales entre si.

Como resultado final del ejercicio se puede decir que los dos métodos son efectivos a la hora de evidenciar información que cambia a través del tiempo en este caso los resultados no fueron significativos por ninguno de los métodos sin embargo pudimos comprobar que en próximos estudios es valido analizarlos desde la perspectiva funcional ya que graficamente es mas facil su interpretacion ademas de que la carga computacional es mejor por este metodo.

Referencias

[1] Aristizabal. P (2011)., *Metodologıa estadıstica para el análisis de datos funcionales cerebrales: Una aproximación con potenciales evocados.*

[2] Arnau, J. y Bono, R. (2008)., (1989) *Estudios longitudinales de medidas repetidas. Modelos de diseño y análisis*

[3] Bartholomew DJ, Knott M., *Modelos variables latentes y análisis factorial*, 2da ed. Londres, Reino Unido: Arnold; 1999.

[4] Bollen K. A.(1989) *Structural equations with latent variables. New York: John Wiley a Sons*

[5] Chinn, S., (1989) *Longitudinal studies: objectives and ethical considerations.*

[6] Dunson, D., (2005) *Bayesian Structural Equation Modeling.*

[7] Liang, K. y Zeger, S. (1986) *Longitudinal Data Analysis Using Generalized Linear Models*

[8] Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R.* Chapman and Hall / CRC.

[9] Michael, C., (2018). *Graphical Latent Variable Modeling* Recuperado de https://m-clark.github.io/sem/ .

[10] Mulaik SA. (2009) *Foundations of factor analysis. 2nd ed. Boca Raton: Chapman Hall/CRC*

[11] Nachtigall C, Kroehne U, Funke F, Steyer R. (2003) *should we use SEM? Pros and cons of structural equation modeling. Methods of Psychological Research Online*

[12] Peña, D., (2002) *Análisis de Datos Multivariantes.* España, MCGRAW-HILL / INTERAMERICANA DE ESPAÑA.

[13] ESCUDERO, Amalia I; RECALDE, Celso G; HARO, Silvia M y MENESES, Manuel A. *Spline Cúbico para el Tratamiento Funcional de la Radiación Solar Global.* Ecuador, Escuela Superior Politécnica de Chimborazo. Torres-Saavedra, P. A. (2018). Modelos Estadísticos Avanzados. http://pegasus.uprm.edu/ pedro.torres/book/references.html.

[14] Ferraty, F. and Vieu, P. (2006). *Non parametric functional data analysis.* Theory and practice.

[15] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*