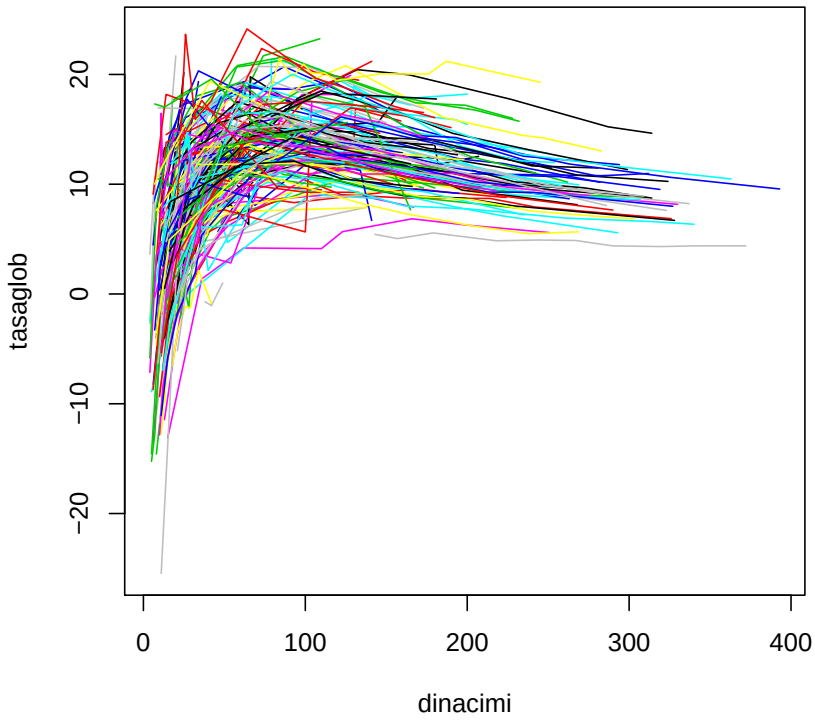
Análisis de datos longitudinales y multivariantes mediante distancias con modelos lineales generalizados

Sandra Esperanza Melo Martínez

Análisis de datos longitudinales y multivariantes mediante distancias con modelos lineales generalizados



Sandra Esperanza Melo Martínez

Análisis de datos longitudinales y multivariantes mediante distancias con modelos lineales generalizados

Memoria presentada por:

Sandra Esperanza Melo Martínez

para optar al título de doctor por la Universidad de Barcelona

Doctorando:

Sandra Esperanza Melo Martínez

Firma:_____

Director:

Carles Maria Cuadras I Avellana

Firma:_____

Universidad de Barcelona
Facultad de Biología
Programa de Doctorado en Estadística
Departamento de Estadística

Barcelona, Mayo de 2012

Dedicatoria

A mi mamá María y a mis hermanos Oscar y Carlos por ser las personas más importantes, valiosas y especiales en mi vida.

Agradecimientos

Agradezco de manera especial a mi director el Doctor Carles Cuadras por ser el mejor profesor que tuve durante mis estudios de doctorado, gracias por sus valiosas enseñanzas durante mis estudios en Barcelona. Por haber aceptado dirigirme la tesis a pesar de su prejubilación, por su valiosa colaboración, amabilidad, disponibilidad, pertinentes observaciones para hacer de este un mejor trabajo y por contribuir a formarme como investigadora.

Quiero expresar mis agradecimientos en especial a mi hermano Oscar quien me ha apoyado constantemente durante toda mi vida, quien siempre ha confiado en mí, me ha dado ánimos y motivación para seguir adelante. Muchas gracias por su colaboración, pertinentes consejos, aclaraciones y paciencia a lo largo de mi vida. Que bueno es haber tenido la fortuna de compartir los estudios de doctorado y algunas clases en la universidad, además de muchas horas de estudio, dedicación, concentración y trabajo.

También, quiero agradecer a mi hermano Carlos con quien iniciamos este proceso de estudios del doctorado y con quien compartimos como compañeros de estudio en todas las materias y nos apoyamos constantemente durante nuestra estadía en Barcelona. Muchas gracias por su colaboración, porque compartimos muchos momentos agradables, porque siempre conté con su apoyo lo que hizo más corta nuestra estadía lejos de la familia. Puedo decir que han sido muchas horas de estudio y trabajo, en el grupo de investigación de la familia Melo (Oscar, Carlos y Sandra), sin este grupo todo habría sido más difícil, me siento afortunada de poder haber estudiado con mis dos hermanos, pues aprendí muchas cosas de ellos durante estos años.

Agradezco a José Enrique Bermúdez a quien conocí en Barcelona, pues ha sido una persona muy importante y especial para mí. Gracias por apoyarme en los momentos difíciles del trabajo, por confiar en mis capacidades, por su grandiosa amistad y consejos que hicieron mi estadía más grata en Barcelona. Mi agradecimiento también para Carlos Gil Bellosta presidente de R en España, por su valiosa colaboración, amistad y solución oportuna de algunas dudas con el R.

El agradecimiento más profundo a mi madre por ser la mejor mamá del mundo, pues siempre nos ha guiado para ser buenas personas, porque con su constante esfuerzo y dedicación logro llevarnos hasta donde hemos llegado hoy en día, sin ella este sueño no habría sido posible. A mi padre, que a pesar de la distancia siempre estuvo atento para saber cómo iba mi proceso de estudios en España. A los profesores del doctorado gracias por sus enseñanzas impartidas, y quienes fueron muy comprensivos en los momentos que tuve dificultades con mi estadía en Barcelona.

Quiero agradecer a la Fundación Carolina por haberme otorgado una beca para realizar mis estudios de doctorado en Barcelona y por sus actividades para darnos a conocer la ciudad. También, agradezco a la Universidad de Barcelona por acogerme durante estos cuatro años de estudio y por las enseñanzas adquiridas durante este tiempo.

Finalmente, no puede faltar mi agradecimiento a la Universidad Nacional de Colombia, donde llevo varios años de formación académica, pues allí realice mis estudios de Estadística, me forme como profesional, también realice el master y es el lugar donde ahora trabajo como docente. Muchas gracias, por darme la oportunidad de ir a otro país a realizar mis estudios de doctorado y por apoyarme durante estos años de estudio.

A todos ellos muchas gracias.

Contenido

Lista de tablas

Lista de figuras

Prefacio

Varias técnicas se han propuesto para el análisis de datos longitudinales y multivariantes. Algunas de éstas exploran la relación entre los datos observados en los puntos del tiempo sucesivos, teniendo en cuenta tales relaciones en los modelos utilizados para la representación de estos datos. Ejemplos de estas técnicas son el análisis de los factores longitudinales, análisis de factores dinámicos, análisis multivariante de series temporales, modelamiento espacio-estado Jørgensen et al. (1996) y análisis de curvas de crecimiento Chaganty & Mav (2007). Estas técnicas están todas basadas en suposiciones de la distribución que no siempre pueden ser plausibles en la práctica.

Desde una perspectiva clásica, los datos longitudinales han sido analizados mediante el modelo del análisis de la varianza (ANOVA) o multivariante (MANOVA) de las medidas repetidas. Sin embargo, recientemente, han emergido una serie de modelos de análisis que superan, en múltiples aspectos, a los modelos clásicos. Todos ellos se subsumen bajo un modelo más amplio, conocido como modelo lineal general mixto.

Estos modelos son abordados en una variedad de áreas, donde las observaciones son tomadas sobre múltiples puntos en el tiempo y medidas en una característica particular, con frecuencia llamada una variable respuesta, para investigar los patrones temporales de cambio en las características. Por ejemplo, a ciertos estudiantes se les puede realizar una prueba estándar repetidamente durante varios meses, la satisfacción de clientes dirigida a una marca en particular se puede medir cada trimestre, el efecto de cierta droga en un grupo de animales, las concentraciones de azúcar en la sangre también se pueden observar a través del tiempo y así en diferentes áreas se pueden encontrar innumerables situaciones similares. Lo cual motiva a abordar este tipo de datos, proponiendo una metodología basada en distancias con algunas variantes para el análisis que permita además tener ciertas ganancias en cuanto a predicción. Por otra parte, se ha visto que datos de este tipo son usualmente analizados por el modelo de curvas de crecimiento, iniciado por Potthoff & Roy (1964), y extensivamente estudiado por numerosos autores como Hwang et al. (2004), Sabo & Chaganty (2009), Chaganty & Mav (2007) entre otros como ya se menciono antes, lo cual también motiva a abordar las curvas de crecimiento

haciendo uso de distancias.

Es importante tener en cuenta que, en los últimos trabajos en el área se supone multinormalidad y en la práctica este supuesto usualmente no es satisfecho en muchos casos, por lo cual en el trabajo se hace una adaptación con modelos lineales generalizados, haciendo uso de las ecuaciones de estimación generalizadas para estimar los parámetros del modelo. Además, la motivación surge por varios problemas que se encuentran de tipo práctico en las diferentes áreas del conocimiento tales como las citadas anteriormente. En la práctica se encuentran varios estudios donde se realizan mediciones sobre un mismo individuo a través del tiempo, y otros con varias variables respuesta en función de ciertas variables independientes que en muchos casos se analizan en forma univariante a través de análisis de varianzas, siendo muy importante para el agrónomo, por ejemplo, ver cuál es el efecto de los tratamientos sobre todas las variables respuesta que fueron medidas y obtener así el mejor tratamiento que produzca el mejor efecto en el experimento, de tal forma que se puedan tomar buenas y mejores decisiones en la práctica.

No solo eso, sino que también en los últimos trabajos que se han abordado a través de distancias se han visto ganancias importantes en predicción con respecto a otro tipo de métodos que emplean otro tipo de criterios para realizar las predicciones, y en muchos casos prácticos resulta importante poder hacer buenas predicciones con el modelo ajustado, por lo cual es interesante abordar el estudio de datos longitudinales y datos multivariantes usando distancias ya que adicionalmente es un campo de trabajo por donde aun se pueden hacer varios aportes teóricos y es posible proponer nuevos métodos de análisis teniendo como soporte teórico algunos de los trabajos que ya se han hecho sobre este campo. Por tal razón, resulta interesante desarrollar este trabajo para ver así las ventajas y desventajas que puede proporcionar esta propuesta respecto a otras metodologías ya existentes, además de proporcionar a los investigadores en el área y usuarios de la estadística otros métodos para el análisis de este tipo de datos, esperando ofrecer alguna ganancia respecto a otro tipo de métodos.

Capítulo 1

Introducción

Son muchos los fenómenos de la vida cotidiana que se salen de las manos al intentar traducirlos a un lenguaje simbólico propio de la disciplina estadística. En consecuencia, se cae en el abismo de ajustar dichos fenómenos a los modelos que se tiene a disposición, en lugar de permitir que los datos "hablen por sí solos". El intento por acercar la teoría a las situaciones reales ha motivado el desarrollo de técnicas estadísticas encaminadas a encontrar modelos cada vez más generales que respondan fielmente a los objetivos del investigador en correspondencia con la realidad.

Por tal razón, en éste trabajo se aborda el análisis de datos longitudinales desde diferentes perspectivas, a través de distancias entre pares de observaciones respecto a las variables explicativas. Inicialmente, se hace una revisión bibliográfica para ver el estado del arte a la fecha. Se inicia la revisión con los datos multidimensionales que surgen cuando un número diferente de variables respuesta son requeridos para medir los resultados de interés. Ejemplos de tales resultados incluyen calidad de vida, capacidad cognitiva, investigaciones biológicas, agronómicas, sociales y de la salud, entre otros, donde se realizan mediciones a lo largo del tiempo sobre una variable dependiente, o sobre varias variables dependientes. El análisis de estos datos resulta algo complejo, debido a la presencia de correlación entre las medidas repetidas en el tiempo, como también entre las variables respuesta.

Existen diferentes métodos para abordar este tipo de datos, tales como el modelo mixto multivariante o modelo doblemente multivariante (Boik 1991), pero ambos enfoques asumen una distribución normal multivariante y homogeneidad de las matrices de covarianzas a través de los niveles del factor de agrupación, además de independencia entre las observaciones de diferentes individuos y covariables independientes del tiempo. No obstante, un problema que se presenta cuando hay desviaciones de uno o más de estos supuestos, conllevan a que no se controle las tasas de error tipo I y por consiguiente

se altera el proceso de inferencia. Sin embargo, algunos autores han estudiado este problema de cómo controlar las tasas de error tipo I para diseños balanceados y desbalanceados cuando la condición de homogeneidad de covarianzas no es sostenible (Keselman & Lix 1997, Kowalchuk et al. 2003, Lix & Algina 2003, Welch 1951). Además, pruebas para efectos de tratamientos en medidas repetidas combinando métodos Bootstrap y medias recortadas son estudiadas por Keselman et al. (2000), donde se muestra que se puede controlar y disminuir el error tipo I.

Cabe agregar que algunos otros autores han abordado el problema de datos longitudinales de respuesta multivariante; es así, el caso del artículo de Gray (2000) donde se desarrolla una metodología para estimar un efecto de tratamiento de datos multidimensionales que han sido recolectados longitudinalmente, usando respuestas en el tiempo al evento, continuas o discretas o una mezcla de este tipo de respuestas. Una transformación de la escala de tiempo que no depende de las unidades de las variables respuesta se utiliza para capturar el efecto de los tratamientos. Esta información permite sobre el efecto de los tratamientos, la combinación a través de las variables respuesta de diferentes tipos. Luego, el modelo se específica usando un par de modelos de regresión, para los primeros dos momentos, y se utilizan ecuaciones de estimación generalizadas para la estimación de los parámetros.

Adicionalmente, datos multidimensionales surgen también en agronomía y otras áreas del conocimiento, cuando un número diferente de variables respuesta son medidas. En muchas instancias, estas variables de respuesta múltiple son destinadas a medir un resultado fundamental de interés que no puede ser capturado por una sola variable respuesta. Una importante complejidad de los datos de respuesta múltiple es que las variables respuestas pueden ser definidas sobre diferentes escalas numéricas. Además de variables continuas, las medidas tomadas sobre el tiempo constan también de variables discretas, variables tiempo a evento o una mezcla de estos tres tipos de variables respuesta (Gray 2000).

De esta manera, datos de respuesta multivariante se han discutido extensivamente en la literatura por varios autores, entre ellos Li et al. (2003), quienes propusieron un método para reducir la dimensionalidad de variables respuestas, haciendo uso de regresión inversa en rodajas (SIR). El enfoque de esta área es reducir la dimensionalidad de las variables regresoras al caso univariante, también hay más trabajos donde se asume que la variable respuesta es univariante. SIR es un método para encontrar vectores en la reducción de la dimensión efectiva (Li 1991), bajo una condición de linealidad para las variables independientes. También, una discusión detallada es dada para el caso donde la respuesta es una curva medida en puntos fijos. El problema de este ajuste es seleccionar funciones base para ajustar un agregado de curvas. Adicionalmente, varios libros sobre este tema han sido publicados, entre los que

se destacan los trabajos de Diggle et al. (1994), Jones (1993), Lindsey (1993), Rencher (2002), Diggle et al. (2002) y Molenberghs & Verbeke (2005). Algunos autores han abordado el problema usando modelos de curvas de crecimiento, iniciado por Potthoff & Roy (1964) y estudiado extensivamente por muchos autores, incluyendo Rao (1965), Khatri (1966), Grizzle & Allen (1969), Laird & Ware (1982), Crowder & Hand (1990), Kshirsagar & Boyce (1995), entre muchos otros.

Asimismo, entre otras metodologías estudiadas para abordar el análisis de datos de respuesta multivariante están el análisis de componentes principales (ACP), éste es quizás un método conocido para reducir la dimensionalidad. Por ejemplo, ACP puede ser aplicado en una regresión para reducir la dimensión de los regresores, pero este procedimiento se lleva a cabo sin usar las variables respuestas, es bastante previsible que las variables regresoras más importantes se pueden perder durante el proceso de reducción (Li et al. 2003). En este caso, se trabaja con regresión inversa en rodajas, que es un método para encontrar vectores en el "Espacio Reducción de Dimensión Efectiva" (EDR) bajo una condición de linealidad que ha sido discutida extensivamente en la literatura por varios autores, (ver Li et al. (2003)).

Entre otros autores que han trabajado en este tema se encuentran Chaganty & Naik (2002), quienes consideran el análisis de datos longitudinales multivariantes asumiendo una escala múltiple de producto Kronecker en la estructura de correlación para la matriz de covarianzas de las observaciones sobre cada sujeto. El método usado para la estimación de los parámetros es el método cuasi-mínimos cuadrados, método desarrollado en los siguientes cuatro artículos: Chaganty (1997), Shults & Chaganty (1998), Chaganty & Shults (1999) y Chaganty & Naik (2002), quienes muestran que las ecuaciones de estimación para los parámetros de correlación en el método cuasi-mínimos cuadrados son óptimas. Además, las ecuaciones de estimación son insesgadas si los datos provienen de una población normal.

Por otro lado, una extensión de los modelos lineales generalizados al análisis de datos longitudinales fue propuesta por Liang & Zeger (1986b), quienes introducen una clase de ecuaciones de estimación generalizadas para analizar datos longitudinales que generan estimaciones consistentes de los parámetros de regresión y de sus varianzas bajo leves condiciones sobre la dependencia del tiempo. Las ecuaciones de estimación se obtienen sin especificar la distribución conjunta de las observaciones de los sujetos; sin embargo, se reducen a las ecuaciones score para resultados gaussianos multivariantes. Estos autores presentan la teoría asintótica para la clase general de estimadores. También, discuten casos donde se asume independencia y dependencia en las estructuras de correlación de cada sujeto. Adicionalmente, Chaganty (1997) muestra un método para estimar los parámetros de correlación el cual supera la propuesta de Crowder (1995), para algunas estructuras de correlación, obteniendo esti-

maciones factibles para los parámetros de correlación.

Además, otro de los modelos usados es el de curva de crecimiento. La idea base de este modelo es introducir algunas funciones conocidas, llamadas funciones base, es decir funciones polinomiales, tal que capturen patrones de cambio para medidas dependientes del tiempo. Sin olvidar que, el modelo de curva de crecimiento tradicional fue diseñado para las situaciones donde los individuos son medidos sobre una sola variable respuesta. En Reinsel (1982) se extiende el modelo curva de crecimiento univariante al caso multivariante, donde varias variables respuesta son medidas sobre múltiples puntos en el tiempo. Hwang et al. (2004) estudian el caso donde las variables respuestas no tienen que ser medidas en los mismos puntos del tiempo y no se debe tener el mismo número de puntos en el tiempo, además muestran que es posible aplicar varias clases de matrices de función base con diferentes rangos a través de las variables respuesta.

No obstante, entre los últimos trabajos en el tema se encuentra el artículo de Sabo & Chaganty (2009) donde adaptan el método cuasi-mínimos cuadrados, proponen un procedimiento robusto para estimar correlación entre variables continuas para el análisis de datos del núcleo familiar en clúster. También, los estimadores que se obtienen en este procedimiento se comparan con máxima verosimilitud tradicional y el estimador de momentos, además del énfasis en la estimación de las correlaciones dentro de una familia nuclear.

Algunos trabajos recientes son el de Genolini & Falissard (2011), quienes desarrollan el paquete KML en R que proporciona una implementación de k-medias, diseñado para trabajar específicamente en datos longitudinales. Puede funcionar k-medias con las distancias diseñadas para datos longitudinales (como la distancia de Frechet o alguna distancia definida por el usuario). La interfaz gráfica permite al usuario elegir el número adecuado de clusters cuando los criterios clásicos no son eficientes.

Entre otros de los trabajos esta el de Liugen & Lixing (2007), quienes proponen hacer inferencia basada en verosimilitud empírica local para un modelo con coeficientes variables en datos longitudinales. Muestran que la razón de verosimilitud empírica es asintóticamente ji-cuadrado estándar cuando se emplea suavizamiento. Además, definen un estimador de máxima verosimilitud empírica con coeficientes variables en el tiempo, muestran la equivalencia asintótica con el estimador de mínimos cuadrados ponderados y la normalidad asintótica.

Adicionalmente, Geraci & Bottai (2007) proponen un nuevo modelo lineal para regresión por cuantiles para datos longitudinales que incluye efectos aleatorios, con el fin de dar cuenta de la dependencia entre las observaciones seriales sobre el mismo sujeto. La noción de regresión por cuantiles es sinónimo de un análisis robusto de la distribución condicional de la variable respuesta.

También, presentan una aproximación basada en verosimilitud para la estimación de los cuantiles de la regresión que utiliza la densidad de Laplace asimétrica.

Otro trabajo por resaltar es el de Yao et al. (2005), donde se propone un método no paramétrico para llevar a cabo el análisis de componentes principales funcional para el caso de escasos datos longitudinales. El método tiene por objeto datos longitudinales espaciados irregularmente, donde el número de medidas repetidas disponible por sujeto es pequeño. En contraste, el análisis de datos funcional clásico requiere un número grande de medidas espaciadas regularmente por sujeto. En Yao et al. (2005) se asume que las medidas repetidas son localizadas aleatoriamente con un número aleatorio de repeticiones para cada sujeto y son determinadas por un suavizamiento aleatorio (especificado por sujeto) más la trayectoria de los errores de medición. También, se realiza una estimación parsimoniosa de la estructura de covarianza y la estimación de la varianza de los errores de medición.

Para desarrollar las metodologías propuestas en la tesis, se tiene como soporte teórico los trabajos usando distancias, los cuales son estudiados por Cuadras (1989), donde se hace la aplicación de funciones distancia, junto con análisis de coordenadas principales, para algunos problemas multivariantes, a saber regresión múltiple, MANOVA y análisis discriminante. También, el análisis discriminante con variables continuas y discretas, así como con datos ordinales, binarios y cualitativos se estudia; asimismo expresiones de distancias entre individuos se proponen y discuten.

Además, pruebas multimuestra basadas en distancias para datos multivariantes son estudiadas por Cuadras (2008). En otro trabajo el mismo autor muestra como relacionar dos conjuntos de datos, cuando las observaciones son tomadas sobre los mismos individuos, estudiando algunas medidas de asociación multivariante basadas solamente sobre distancias entre individuos y mostrando una prueba de permutación para decidir si la asociación es significante; en otros trabajos recientes se estudia la regresión multivariante basada en distancias (ver Cuadras (2011)).

Cabe agregar que, en muchos métodos de estadística y análisis de datos se utiliza el concepto geométrico de distancia entre individuos o poblaciones, estos métodos son aplicados en campos tales como la agronomía, antropología, biología, genética, psicología, entre otros (Arenas & Cuadras (2002)). Las distancias, aparecen en muchos aspectos de la estadística: contraste de hipótesis, estimación, regresión, análisis discriminante, etc. (ver Cuadras (2007)). También, Cuadras & Arenas (1990) proponen un método de regresión múltiple basado en el análisis de distancias utilizando diferentes métricas para el trabajo con variables explicativas continuas y categóricas. No obstante, Cuadras et al. (1996) presentan algunos resultados adicionales del modelo basado en distancias (DB)

para la predicción de variables mezcladas (continuas y categóricas) y exploran el problema de información faltante dando una solución utilizando DB. Uno de los trabajos más recientes es el de Esteve et al. (2010), quienes desarrollan un método donde incluyen términos polinomiales y de interacción en la regresión basada en distancias, bajo las propiedades de un producto de matrices semi-Hadamard o Khatri-Rao. Además, Boj et al. (2010), consideran el problema de predecir no-paramétricamente una variable respuesta escalar de un predictor funcional. También la implementación de mínimos cuadrados parciales para regresión basada en distancias es estudiada por Boj, Grané, Fortiana & Claramunt (2007). Incluso Boj, Claramunt & Fortiana (2007) proponen una solución al problema de la selección del predictor definiendo una prueba estadística generalizada y adaptando un método bootstrap no-paramétrico para estimar sus p-valores.

Este trabajo se desarrolla en siete capítulos: en el Capítulo 1 se presenta una introducción donde se referencian varios autores que han trabajado en el campo de los datos longitudinales y otros que han usado distancias para el ajuste de los modelos. Muestra lo que se ha hecho en este campo de investigación históricamente. En el Capítulo 2 se propone una metodología para el análisis de datos longitudinales en aproximación multivariante basado en distancias. Se plantea el modelo, se hace la estimación de los parámetros y las pruebas de hipótesis, por lo tanto se realiza la inferencia correspondiente con el modelo propuesto. En el Capítulo 3 se presenta una metodología para analizar datos longitudinales en aproximación univariante, se presenta el modelo y se muestra como realizar la estimación de los parámetros.

En el Capítulo 4 se presenta una metodología para la aproximación univariante a las curvas de crecimiento mediante distancias entre pares de observaciones respecto a las variables explicativas y distancias entre tiempos. Además, se presenta la inferencia correspondiente para el modelo propuesto. En el Capítulo 5 se presentan algunos aspectos de los modelos lineales generalizados, esenciales para la comprensión de la metodología propuesta en el Capítulo 6; en el cual se muestra una metodología para analizar datos longitudinales con respuesta no normal haciendo uso de distancias entre pares de observaciones con respecto a las variables explicativas, estimando los parámetros del modelo por medio de las ecuaciones de estimación generalizadas. En la parte final del Capítulo se muestra una aplicación real con sobredispersión donde se puede ver como funciona la metodología propuesta. Finalmente, en el Capítulo 7 se presentan algunas conclusiones y recomendaciones de este trabajo.

En cada uno de los capítulos se presenta una aplicación y en los Capítulos 2 y 3 se muestran los resultados de la simulación usando la distancia de Gower con datos mixtos, mediante MANOVA y aproximación univariante en datos longitudinales, donde se encuentran ganancias en el método DB con respecto al clásico. También, se desarrollan los programas en R para el análisis correspon-

diente en cada capítulo, con las diferentes metodologías propuestas bajo DB y las clásicas. Además, en el Capítulo 6 se utiliza el procedimiento GENMOD del SAS, estos programas se anexan en un CD dentro de la tesis, junto con el programa usado para la simulación de los Capítulos 2 y 3.

De modo que, del trabajo se puede ver que los métodos propuestos para modelar problemas de este tipo producen resultados igualmente de robustos que las estrategias clásicas de modelamiento de esta misma clase de problemas. En este sentido, se adaptaron las metodologías existentes en datos longitudinales y multivariantes bajo la estrategia de modelos lineales generalizados mediante distancias. Aunque este tema se ha abordado por otros métodos, no se ha estudiado a través de distancias lo cual conlleva a tener una ganancia en las predicciones, ya que es posible agregar mas componentes al modelo mejorando así la calidad de las predicciones.

En los Capítulos 2, 3 y 4 de las metodologías propuestas para el análisis de datos longitudinales se hace uso de distancias entre pares de observaciones con respecto a las variables explicativas, mediante variables explicadas continuas, y funcionan también en casos con variables explicativas categóricas, binarias, mixtas y continuas. Se demuestra que las predicciones generadas son las mismas bajo el modelo propuesto y el clásico en los Capítulos 2 y 3. Pero cuando se tienen datos mixtos usando la distancia de Gower se observa que no se obtienen las mismas predicciones, resultado que se puede ver de la simulación.

Objetivos

Objetivos generales

- Proponer una metodología para analizar datos longitudinales mediante distancias entre pares de observaciones con respecto a las variables explicativas.
- Formular un método para analizar curvas de crecimiento mediante distancias entre pares de observaciones con respecto a las variables explicativas y distancias entre los tiempos.
- Plantear un método de análisis longitudinal con respuesta no normal mediante distancias entre pares de observaciones con respecto a las variables explicativas usando modelos lineales generalizados.

Objetivos específicos

- Aplicar las metodologías propuestas a un caso práctico en datos longitudinales y curvas de crecimiento.
- Por medio de "Métodos Montecarlo", comparar la metodología propuesta basada en distancias (DB) con respecto al método clásico en datos longitudinales mixtos usando la distancia de Gower.
- Realizar la inferencia para el método DB propuesto en datos longitudinales.

Capítulo 2

Inferencia en la aproximación basada en distancias en el análisis de datos longitudinales

Desde una perspectiva clásica, los datos longitudinales han sido analizados usando el modelo de análisis de varianza (ANOVA) o multivariante (MANOVA) con medidas repetidas. Sin embargo, recientemente, han emergido un número de modelos estadísticos que superan en muchos aspectos los modelos clásicos. Todos ellos englobados bajo un modelo más amplio, conocido como el modelo mixto lineal general. El análisis multivariante de varianza, aplicado a datos longitudinales asume que medidas multiples son variables dependientes que están correlacionadas en los mismos sujetos. Cuando hay medidas repetidas, MANOVA es una buena alternativa al análisis univariante. Los estudios longitudinales están caracterizados por los registros de datos que contienen medidas repetidas por sujeto, medidas en varios puntos sobre un eje de tiempo adecuado. El objetivo es con frecuencia estudiar el cambio en el tiempo o la dinámica del tiempo de fenómenos biológicos tales como crecimiento, fisiología, fisiopatología y patogenia (Müller 2009). También, el interés es relacionar estas dinámicas sobre el tiempo para ciertos predictores o respuestas. El análisis clásico de los estudios longitudinales esta basado en modelos paramétricos los cuales con frecuencia contienen efectos aleatorios como el modelo mixto lineal generalizado (MMLG) de métodos marginales tales como las ecuaciones de estimación generalizadas (EEG).

El análisis longitudinal multivariante ha sido también estudiado por Gray (2000) quien propone una metodología para estimar un efecto de tratamiento de datos longitudinales multidimensionales donde las variables respuesta pueden ser alguna mezcla de variables continuas, discretas y respuestas tiempo a evento. La idea que el efecto de tratamiento puede ser capturado con una

transformación de la escala de tiempo se muestra naturalmente extendido a variables respuesta discretas y tiempo a evento. La ventaja de usar esta aproximación es que los parámetros de los tratamientos no dependen de la escala de la variable respuesta. A causa de esto, se puede combinar información sobre el efecto de los tratamientos a través de la variable respuesta en diferentes tiempos. Se utilizan ecuaciones de estimación generalizadas para la estimación de los parámetros.

Por otro lado, Laird & Ware (1982) discutieron la ventaja de trabajar con modelos de efectos aleatorios a dos vías para datos longitudinales, incluyendo modelos de curvas de crecimiento y medidas repetidas como casos especiales. También, se combina el método de máxima verosimilitud y estimación Bayesiana empírica de los parámetros del modelo y el uso del algoritmo EM, Jennrich & Schluchter (1986) trabajaron estimación de máxima verosimilitud bajo un modelo muy general para medidas repetidas. Además presentaron la estimación usando los algoritmos iterativos de Newton Raphson, Fisher Scoring y una combinación de EM con Scoring.

Subsecuentemente, Laird et al. (1987) presentaron una aplicación del uso del algoritmo EM para encontrar estimaciones de los parámetros de un conjunto de medidas repetidas bajo un modelo lineal mixto, a través del método de máxima verosimilitud y máxima verosimilitud restringida. Específicamente el modelo teórico, las ecuaciones iterativas que definen el algoritmo, discuten la existencia de soluciones explícitas en casos de datos balanceados y el cálculo de valores iniciales para el proceso iterativo. Andreoni (1989) presentó un estudio de modelos de efectos aleatorios para el análisis de datos longitudinales desbalanceados en relación al tiempo. Además, Andreoni (1989) mostró varios modelos para la estructura media y matrices de covarianza, y presentó una comparación del método de estimación de máxima verosimilitud y máxima verosimilitud restringida usando los algoritmos de Newton Raphson, Scoring Fisher y EM. Adicionalmente, Davis (2002) presenta una descripción de los métodos estadísticos desarrollados para el análisis de medidas repetidas, muestra diferentes alternativas para analizar un conjunto de datos, desde el punto de vista descriptivo para modelos mixtos con variable respuesta continua.

En la práctica varios conjuntos de datos se ajustan a la estructura de un análisis multivariante de varianza (MANOVA) pero no están en correspondencia con las condiciones de MANOVA (Gower & Krzanowski 1999). Para establecer una base para el análisis, Gower & Krzanowski (1999) examinaron la estructura de matrices distancia en la presencia a priori de la agrupación de unidades y mostraron como la distancia de cuadrados total entre las unidades de un conjunto de datos multivariantes puede ser particionada de acuerdo a los factores de una clasificación externa. La partición es exactamente análoga a la del análisis univariante de varianza, proporciona un marco de trabajo para el análisis de algún conjunto de datos cuya estructura conforma un MANOVA,

pero el cual por varias razones no puede ser analizado por esta técnica.

En este capítulo se propone la extensión de los métodos de estimación basados en distancias en aproximación multivariante a los datos longitudinales, usando distancias entre pares de observaciones con respecto a las variables explicativas en variables respuesta continuas. Se estudian datos balanceados, donde el número de veces que cada individuo se mide es el mismo, y los tiempos se consideran igualmente espaciados. Se encontraron algunas ventajas en el uso de los métodos basados en distancias con aproximación multivariante, tales como: las componentes de la matriz del ACP son independientes, donde las variables originales usualmente no lo son. En las circunstancias donde los investigadores están principalmente interesados en hacer predicciones, la metodología propuesta es también útil ya que arroja un mejor ajuste que en los modelos clásicos cuando componentes adicionales se agregan. También, al ser un análisis de datos longitudinales, permite a los investigadores hacer predicciones en cada punto del tiempo, lo cual resulta útil para estimar datos faltantes. Además, se encontró que el uso de esta estrategia para modelar problemas de está clase produce resultados igualmente de robustos que la estrategia de modelamiento tradicional y trabaja en casos con variables explicativas categóricas, binarias, mixtas y continuas. Adicionalmente, se probó que las predicciones generadas son las mismas bajo el modelo propuesto y el modelo clásico, excepto en datos mixtos usando la distancia de Gower, este resultado puede verse en la simulación.

Este capítulo es desarrollado en cuatro secciones: en la Sección 2.1 es construido el modelo DB con datos longitudinales en aproximación multivariante. Además, es presentando el ajuste del modelo, la estimación de los parámetros, pruebas de hipótesis en el caso multivariante y cómo realizar la selección de las dimensiones principales. En la Sección 2.2 se presentan las medidas de asociación multivariante y cómo hacer la predicción de un nuevo individuo. La Sección 2.3 muestra los resultados de la simulación para la aproximación multivariante y la Sección 2.4 presenta una aplicación de la metodología propuesta.

2.1 Modelo multivariante: aspectos inferenciales

Sea y_{ir} que denota la respuesta del individuo i ésimo para la r-ésima condición de evaluación, con $i = 1, \ldots, n$ y $r = t_1, \ldots, t_m$. También se asume que y_{ir} es descrito por un modelo lineal general

$$y_{ir} = v_i' \beta_r + e_{ir}$$

donde $v_i = (v_{i1}, \ldots, v_{ip})'$ es un vector de p coeficientes específicos conocidos para el i-ésimo individuo y $\beta_r = (\beta_{1r}, \ldots, \beta_{pr})'$ es un vector de p parámetros desconocidos.

Sea $e_i = (e_{it_1}, \ldots, e_{it_m})'$ que denota un vector de m residuales del i-ésimo sujeto, con distribución $e_i \sim NM(0_m, \Sigma)$. El vector $nm \times 1$ es

$$\underset{\sim}{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

tiene distribución normal $NM(0_{nm}, I_n \otimes \Sigma)$, donde 0_{nm} denota un vector de ceros de tamaño $nm \times 1$, I_n denota la matriz identidad de dimensión $n \times n$ y el operador $\otimes$ denota el producto Kronecker. Entonces, los y_i son vectores aleatorios independientes con distribución $NM(\mu_i, \Sigma)$ donde

$$\mu_i = \begin{bmatrix} \mu_{i1} \\ \vdots \\ \mu_{im} \end{bmatrix} = \begin{bmatrix} v_i'\beta_1 \\ \vdots \\ v_i'\beta_m \end{bmatrix}$$

Para garantizar que la matriz de covarianza Σ de los y_i sea definida positiva, es decir, todos los valores propios de Σ sean positivos, se debe tener que $p \leq n - m$. Con la finalidad de expresar el modelo en forma matricial se definen las siguientes matrices

$$Y_{n\times m} = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix} = \begin{bmatrix} y_1' \\ \vdots \\ y_n' \end{bmatrix}, \quad V_{n\times p} = \begin{bmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{np} \end{bmatrix} = \begin{bmatrix} v_1' \\ \vdots \\ v_n' \end{bmatrix},$$

$$\boldsymbol{\beta}_{p\times m} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pm} \end{bmatrix} = [\beta_1, \ldots, \beta_m]$$

y

$$\mathbf{e}_{n\times m} = \begin{bmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{bmatrix} = \begin{bmatrix} e_1' \\ \vdots \\ e_n' \end{bmatrix}$$

donde Y es la matriz de datos, V es una matriz diseño de rango $p \leq (n-m)$, $\boldsymbol{\beta}$ es la matriz de parámetros desconocidos y $\mathbf{e}$ es la matriz de errores aleatorios. Entonces, el modelo matricialmente puede ser escrito como

$$Y = V\boldsymbol{\beta} + \mathbf{e} \tag{2.1}$$

donde $E(Y) = V\boldsymbol{\beta}$ y $Var(Y) = I_n \otimes \Sigma$.

2.1.1 Aproximación basada en distancias en el modelo longitudinal

Sea $\Omega = \{\omega_1, \ldots, \omega_n\}$ un conjunto con n individuos. Sea $\delta_{ii'} = \delta(\omega_i, \omega_{i'}) = \delta(\omega_{i'}, \omega_i) \geq \delta(\omega_i, \omega_i) = 0$ una función de distancia (o disimilaridad) definida sobre Ω. Supóngase que la matriz de distancias con dimensión $n \times n$, $\Delta = (\delta_{ii'})$ es Euclidiana. Entonces existe una configuración de puntos $v_1, \ldots, v_n \in \Re^p$, con $v_i = (v_{i1}, \ldots, v_{ip})'$, $i = 1, \ldots, n$, tal que

$$\delta_{ii'}^2 = \sum_{j=1}^{p} (v_{ij} - v_{i'j})^2 = (v_i - v_{i'})'(v_i - v_{i'}) \tag{2.2}$$

Estas coordenadas constituyen la matriz $V = (v_{ij})$ (definida en el modelo (2.1) de dimensión $n \times p$ tal que la distancia Euclidiana entre dos individuos i e i' es igual a $\delta_{ii'}$ (Cuadras 2008).

La distancia definida en (2.2) puede utilizarse cuando todas las variables en la matriz V sean continuas. En tal caso, esta puede ser reemplazada por la distancia valor absoluto que es bastante eficiente

$$\delta_{ii'}^2 = \sum_{h=1}^{p} |v_{ih} - v_{i'h}| \tag{2.3}$$

la cual cumple las condiciones de una distancia Euclidiana.

Por otro lado, en el modelo (2.1) la matriz V se puede particionar como $V = (V_1 \quad V_2)$ donde V_1 es una submatriz de variables continuas y V_2 una submatriz de variables cualitativas. De acuerdo a Cuadras & Arenas (1990) se puede definir la similaridad como

$$s_{ii'} = \frac{\sum_{h=1}^{p_1} \left(\frac{1 - |v_{ih} - v_{i'h}|}{G_h}\right) + a + \alpha}{p_1 + (p_2 - d) + p_3} \tag{2.4}$$

donde p_1 es el número de variables continuas, a y d son el número de coincidencias y no coincidencias para las p_2 variables binarias, respectivamente, y α es el número de coincidencias de las p_3 variables cualitativas. G_h es el rango (o recorrido) de la h-ésima variable cuantitativa. La similaridad (2.4) es conocida como distancia de Gower (1968).

La distancia al cuadrado entre los individuos i y i' es

$$d_{ii'}^2 = 1 - s_{ii'} \tag{2.5}$$

Ahora es definido $\Delta^{(1)} = (d_{ii'})$ como una matriz de distancias Euclidiana sobre el conjunto de n individuos.

En el caso que todas las variables explicativas en el modelo (2.1) sean cualitativas una medida bastante utilizada de similaridad entre dos individuos i y i' es $m_{ii'}$, el número de coincidencias en i y en i'. Ya que $m_{ii'} \leq p$, una medida de distancia puede ser definida como

$$\delta^2_{ii'} = 2(p - m_{ii'}) \tag{2.6}$$

Una vez seleccionada alguna de las distancias presentadas anteriormente es definido $A_x = -\frac{1}{2}\Delta_x^{(2)}$ y $F_x = \mathcal{H}A_x\mathcal{H}$, donde $\Delta_x^{(2)} = (\delta^2_{ii'})$ y $\mathcal{H} = I - \frac{1}{n}\mathbf{1}\mathbf{1}' = I - \frac{1}{n}J$ es la matriz centrada, con $\mathbf{1}$ un vector de unos de longitud $n \times 1$ y $J = \mathbf{1}\mathbf{1}'$. Además, F_x es una matriz semi-definida positiva (Mardia et al. 1979) de rango p. De este modo, se tiene la descomposición espectral

$$\begin{aligned} F_x = &\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) A_x \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = U_x\Lambda_x^2U_x' \\ =&XX' \end{aligned} \tag{2.7}$$

donde $X = U_x\Lambda_x$ es una matriz de $n \times p$ de rango p, Λ_x es la matriz de valores propios positivos de F_x y U_x contiene las coordenadas estandarizadas. Además, las filas $x_1', \ldots, x_n'$ de la matriz X son las coordenadas principales de F_x. Así, si un individuo i es similar a un individuo i' en (2.1) entonces $v_i \cong v_{i'}$, y por lo tanto $x_i \cong x_{i'}$.

El modelo que finalmente es propuesto esta dado por

$$Y = \mathbf{1}B_0 + XB + \Xi \tag{2.8}$$

donde $\mathbf{1}$ es el vector de unos de orden $n \times 1$, Y es igual que en el modelo (2.1), $X_{n\times s}$ es conocida de $rang(X) = s$, $B_{s\times m}$ es una matriz de parámetros desconocidos, B_0 es el vector de interceptos desconocidos de orden $1 \times m$ y Ξ es una matriz de errores aleatorios de orden $n \times m$. Obsérvese que como $F_x\mathbf{1} = \mathbf{0}$, tanto $\mathbf{1}$ como las columnas $X_1, X_2, \ldots, X_s$ de X, son vectores propios de F_x.

El modelo (2.8) se puede escribir como

$$Y = \mathbf{1}B_0 + \sum_{i=1}^{s} X_iB_i + \Xi$$

donde $s = rang(F_x)$ y $X_1, X_2, \ldots, X_s$, juegan el papel de variables predictoras.

De acuerdo a Cuadras (2007) a veces $s = rang(F_x)$ crece con n (incluso puede darse el caso en que $s = n - 1$). Entonces, el número de variables $X_1, X_2, \ldots, X_s$ (las columnas de X) puede resultar excesivo y de esta manera se puede encontrar un modelo ajustado arbitrariamente . Para evitar este problema es conveniente partir X en dos partes, $X = (X_{(k)} \quad L)$ donde

$X_{(k)}$ contiene un subconjunto de k columnas de X y L contiene el restante subconjunto de columnas de X. De esta manera, es definido el modelo DB en dimensión k, el cual puede ser expresado de dos maneras equivalentes

$$\begin{aligned} Y =& \mathbf{1}B_0 + X_{(k)}B_{(k)} + \Xi_k \\ =& \mathbf{1}B_0 + \sum_{i=1}^{k} X_i B_i + \Xi_k \end{aligned} \tag{2.9}$$

donde $X_{(k)} = (X_1, \ldots, X_k)$ y cada X_r, con $r = 1, \ldots, k$ es una columna de X (cada X_i es una componente principal).

Los supuestos sobre el modelo (2.9) son

i) $E(Y) = \mathbf{1}B_0 + X_{(k)}B_{(k)}$ (o $E(\Xi_k) = \mathbf{0}$).

ii) $cov(y_i) = \Sigma$ para todo $i = 1, ..., n$ donde y_i' es la i-ésima fila de Y.

iii) $cov(y_i, y_j) = \mathbf{0}$ para todo $i \neq j$.

El supuesto i) establece que el modelo lineal propuesto es el correcto y no son necesarios v's adicionales para predecir los y's. El supuesto ii) afirma que cada uno de los n vectores observados (filas) en Y tienen la misma matriz de covarianza. Mientras que el supuesto iii) afirma que los vectores observados (filas de Y) no están correlacionados entre sí. Por lo tanto, se asume que los y's dentro de un vector de observación (filas de Y) están correlacionados entre sí pero son independientes entre los diferentes individuos observados.

Selección de las dimensiones principales

Inicialmente, el número de variables explicativas puede ser elegido como k. Una buena selección de las columnas $X_1, \ldots, X_k$ de X consiste en escogerlas por orden de coeficiente de correlación múltiple con Y, es decir,

$$R^2(X_1, Y) > R^2(X_2, Y) > \cdots > R^2(X_k, Y)$$

Otra selección consiste en ordenarlas de acuerdo con la variabilidad explicada en los predictores (o columnas de X): $\lambda_1 > \cdots > \lambda_k$, es decir seleccionar los k primeros ejes principales. Pero si la variable X_{k+1} tiene una correlación $R^2_{k+1} = R^2(X_{k+1}, Y)$, relativamente alta, se podría haber perdido una variable predictiva importante (véase Cuadras & Fortiana (1993) para una discusión de este problema).

Cuando n es muy grande, la selección de coordenadas puede volverse en un cálculo muy arduo. Un procedimiento que requiere solo calcular los primeros k vectores propios adecuados, es el siguiente

Se particiona X en $X = (X_{(i)}\ L_i)$, donde $X_{(i)}$ contiene las primeras columnas de X y L_i las restantes, es decir los primeros vectores propios de F_x ordenados de acuerdo con sus valores propios. Por otro lado, considerando las distancias entre los individuos de las matriz de observaciones Y en el modelo (2.9) y realizando un proceso similar al realizado con la matriz V en el modelo (2.1). La descomposición espectral $F_w = U_w\Lambda_w^2U_w'$, las coordenadas estándar U_w y las coordenadas principales $W = U_w\Lambda_w$ se pueden obtener.

Luego como $X_{(i)}$ y W son matrices cuantitativas centradas de dimensiones $n \times i$ y $n \times m$ las siguientes medidas de asociación se pueden definir

1. Escoufier (1973) introdujo la correlación generalizada dada por

$$RV\left(X_{(i)}, W\right) = \frac{tr(S_{12}S_{21})}{\sqrt{tr(S_{11}^2)tr(S_{22}^2)}} \tag{2.10}$$

   donde $S_{11} = X_{(i)}'X_{(i)}$, $S_{22} = W'W$, $S_{12} = X_{(i)}'W$ y $S_{21} = W'X_{(i)}$. Esta correlación es muy relacionada con las estadísticas Procustes (Cox & Cox 2001),

$$R_{(i)}^2 = 1 - \left\{tr\left(X_{(i)}'W'WX_{(i)}\right)^{1/2}\right\}^2 \Big/ \left\{tr\left(X_{(i)}'X_{(i)}\right) tr(W'W)\right\} \tag{2.11}$$

2. Yanai et al. (2006) emplean determinantes de matrices rectangulares para introducir la medida

$$Re\left(X_{(i)}, W\right)^2 = \frac{\begin{vmatrix} X_{(i)}'X_{(i)} & X_{(i)}'W \\ W'X_{(i)} & W'W \end{vmatrix}}{\left|X_{(i)}'X_{(i)}\right| |W'W|} \tag{2.12}$$

3. Cuadras (2008) define la asociación como

$$\eta^2\left(X_{(i)}, W\right) = \left|U_{x_{(i)}}'U_yU_y'U_{x_{(i)}}\right| \tag{2.13}$$

Algunas propiedades de las medidas anteriores son: $RV\left(X_{(i)}, W\right) = 1$, $R_{(i)}^2 = Re\left(X_{(i)}, W\right)^2 = 0$ si $X_{(i)} = TW$ (T ortogonal) y $RV\left(X_{(i)}, W\right) = 0$, $R^2 = Re\left(X_{(i)}, W\right)^2 = 1$, si $X_{(i)}W = 0$. En el caso de la medida de asociación presentada por Cuadras (2008) se satisfacen las siguientes propiedades

a. $0 \leq \eta^2\left(X_{(i)}, W\right) = \eta^2\left(W, X_{(i)}\right) \leq 1$.

b. $\eta^2\left(X_{(i)}, W\right) = \dfrac{\left|X_{(i)}'WW'X_{(i)}\right|}{\left|X_{(i)}'X_{(i)}\right||W'W|}$.

c. $\eta^2\left(X_{(i)}, W\right)$ no depende de la configuración de las matrices $X_{(i)}$ y W.

d. Si w es un vector y $X_{(i)}$ es una matriz, los dos cuantitativos, entonces $R^2\left(w, X_{(i)}\right) = \eta^2\left(w, X_{(i)}\right)$, donde R es el coeficiente de correlación múltiple.

e. Si r_j, $j = 1, \ldots, m$ son los coeficientes de correlación canónica entre $X_{(i)}$ y W, entonces

$$\eta^2\left(X_{(i)}, W\right) = \prod_{j=1}^{T} r_j^2$$

Utilizando la medida de asociación dada por Cuadras (2008), se define la secuencia

$$c(i) = \frac{\eta^2\left(X_{(i)}, W\right)}{\eta^2(X, W)} \text{ con } i = 1, 2, \ldots, p \tag{2.14}$$

Cada $c(i)$ mide la predictibilidad de las primeras i dimensiones. Es de notar aquí, que se podrían utilizar en la anterior ecuación cualquiera de las otras medidas de asociación presentadas anteriormente.

Finalmente, la selección de k en el modelo (2.9) debe ser realizada representando gráficamente los puntos

$$(i, 1 - c(i)) \quad i = 0, 1, \ldots, p^* < p$$

donde p^* es tal que $1 - c(i)$ esté muy próximo a 0. Esto es, el corte óptimo en p^* es tal que, a la derecha de p^* el gráfico está muy próximo al eje horizontal, indicando que las dimensiones superiores no deben ser tenidas en cuenta. La dimensión principal $1 \leq i \leq p^*$ debe ser seleccionada si se aprecia una caída entre el punto $(i-1, 1-c(i-1))$ y el $(i, 1-c(i-1))$. Entonces la dimensión i es aceptada o rechazada según si r_i^2 o λ_i sean grandes o pequeños.

2.1.2 Estimación de parámetros

El modelo presentado en (2.9) se puede escribir como

$$\begin{aligned} Y =& \begin{pmatrix} \mathbf{1} & X_{(k)} \end{pmatrix} \begin{pmatrix} B_0 \\ B_{(k)} \end{pmatrix} + \Xi_k \\ =& \boldsymbol{X}\boldsymbol{B} + \Xi_k \end{aligned} \tag{2.15}$$

donde $\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & X_{(k)} \end{pmatrix} = (\mathbf{1}, X_1, \ldots, X_k)$ y $\boldsymbol{B} = \begin{pmatrix} B_0' & B_{(k)}' \end{pmatrix}'$.

El estimador de mínimos cuadrados (MC) de $\boldsymbol{B}$ es $\widehat{\boldsymbol{B}}$ tal que minimiza la traza de

$$tr\left(\widehat{\Xi}_k' \widehat{\Xi}_k\right) = tr\left[\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)'\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)\right]$$

donde $\widehat{\Xi}_k = Y - \boldsymbol{X}\widehat{\boldsymbol{B}}$.

La matriz de residuos es la matriz $\boldsymbol{R}_0 = (R_0(i,j))$ de orden $m \times m$

$$\boldsymbol{R}_0 = \widehat{\Xi}_k'\widehat{\Xi}_k = \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)'\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)$$

Las estimaciones de MC de los parámetros $\boldsymbol{B}$ verifican las ecuaciones normales (EN)

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{B}} = \boldsymbol{X}'Y \tag{2.16}$$

y vienen dadas por la expresión

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y \tag{2.17}$$

ya que el modelo es de rango máximo $k = rang(\boldsymbol{X})$. Además, se tiene que si Y_j denota la j-ésima columna de Y, entonces

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[Y_1, \ldots, Y_m] = \left[\widehat{\boldsymbol{B}}_1, \ldots, \widehat{\boldsymbol{B}}_m\right],$$

donde $\widehat{\boldsymbol{B}}_j = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y_j$ $(j = 1, \ldots, m)$ es el estimador univariante considerando cada columna de Y como una variable separada.

El estimador presentado en (2.17) es insesgado ya que

$$E\left(\widehat{\boldsymbol{B}}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(Y) = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B} = \boldsymbol{B}$$

Por otro lado, obsérvese que

$$\begin{aligned}\boldsymbol{R}_0 =& \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)'\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)\\ =& Y'Y - Y'\boldsymbol{X}\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}'\boldsymbol{X}'Y + \widehat{\boldsymbol{B}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{B}}\end{aligned}$$

como $\widehat{\boldsymbol{B}}'\boldsymbol{X}'Y = \widehat{\boldsymbol{B}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{B}}$, entonces

$$\begin{aligned}\boldsymbol{R}_0 =& Y'Y - Y'\boldsymbol{X}\widehat{\boldsymbol{B}}\\ =& Y'\left(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)Y\end{aligned} \tag{2.18}$$

***Teorema* 2.1.** *Bajo las condiciones del modelo* (2.15)*, con* $k = rang(\boldsymbol{X})$ *una estimación centrada de la matriz de covarianzas* Σ *es*

$$\hat{\Sigma} = \boldsymbol{R}_0/(n-k)$$

Para la demostración de este teorema se siguieron los pasos de Cuadras (2010). Veamos la prueba a continuación

Demostración: Sea $T = [t_1, \ldots, t_k, t_{k+1}, \ldots, t_n]$ una matriz ortogonal, tal que sus columnas forman una base ortonormal de $\Re^n$, de manera tal que las primeras k columnas generen el mismo subespacio $C_k(\boldsymbol{X})$ generado por las columnas de $\boldsymbol{X}$. Por lo tanto, las otras $n-k$ columnas serán ortogonales a $C_k(\boldsymbol{X})$. Es decir

$$t_i'\boldsymbol{X} = \begin{cases} * & \text{si} \quad i \leq k \\ 0 & \text{si} \quad i > k \end{cases}$$

donde $*$ indica que es posiblemente un valor no-nulo. Considérese ahora $Z = T'Y$, entonces la esperanza matemática de Z es

$$E(Z) = E(T'Y) = T'E(Y) = T'\boldsymbol{X}\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\eta}_k \\ \mathbf{0}_{n-k} \end{pmatrix}$$

donde $\boldsymbol{\eta}$ tiene k filas y $\mathbf{0}$ es de $n-k$ filas.

Considérese los residuos $\widehat{\Xi}_k = Y - \boldsymbol{X}\widehat{\boldsymbol{B}}$. Además, $\boldsymbol{X}'\widehat{\Xi}_k = \boldsymbol{X}'(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}) = \boldsymbol{X}'Y - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y = \mathbf{0}$, esto se tiene ya que es proyector el espacio sobre el mismo. Por lo tanto, $\widehat{\Xi}_k$ es ortogonal a $\boldsymbol{X}$, en el sentido que

$$T'\widehat{\Xi}_k = \begin{pmatrix} \mathbf{0} \\ Z_{(n-k)\times m} \end{pmatrix}$$

ya que

$$T'\widehat{\Xi}_k = T'\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right) = T'Y - T'\boldsymbol{X}\widehat{\boldsymbol{B}} = Z - \begin{pmatrix} * \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ Z_{n-k} \end{pmatrix}$$

Es decir, las últimas $n-k$ filas de Z y $T'\widehat{\Xi}_k$ coinciden. Entonces como $T'T = I$, se tiene que

$$\boldsymbol{R}_0 = \widehat{\Xi}_k'\widehat{\Xi}_k = \widehat{\Xi}_k'TT'\widehat{\Xi}_k = \begin{pmatrix} \mathbf{0}' & Z_{n-k}' \end{pmatrix}\begin{pmatrix} \mathbf{0} \\ Z_{n-k} \end{pmatrix} = Z_{n-k}'Z_{n-k}$$

Haciendo $Z_{n-k}' = [z_1, \ldots, z_{n-k}]$ donde $z_1', \ldots, z_{n-k}'$ son las filas independientes de Z_{n-k}. Entonces cada z_i es un vector de media cero y matriz de covarianzas Σ. Luego $E(z_i z_i') = \Sigma$ y

$$Z_{n-k}'Z_{n-k} = \begin{bmatrix} z_1 & \ldots & z_{n-k} \end{bmatrix}\begin{bmatrix} z_1' \\ \vdots \\ z_{n-k}' \end{bmatrix} = \sum_{i=1}^{n-k} z_i z_i'$$

Por lo tanto, se tiene que

$$\begin{aligned} E(\boldsymbol{R}_0) =& E(Z'_{n-k}Z_{n-k}) = E\left(\sum_{i=1}^{n-k} z_i z_i'\right) = \sum_{i=1}^{n-k} E(z_i z_i') \\ =& (n-k)\Sigma \end{aligned}$$

Luego $\widehat{\Sigma} = \frac{\boldsymbol{R}_0}{n-k} = \frac{\widehat{\Xi}'_k \widehat{\Xi}_k}{n-k}$. □

Del anterior teorema se tiene que $E(Z_{n-k}) = \mathbf{0}$. Así todas las $n-k$ filas de Z_{n-k} son $N_T(\mathbf{0}, \Sigma)$ independientes, entonces $\boldsymbol{R}_0 = Z'_{n-k}Z_{n-k} \sim Wishart_T(\Sigma, n-k)$ ya que cumple las condiciones de una matriz $m \times m$ que sigue la distribución de Wishart (ver mayores detalles en Cuadras (2010) y Mardia et al. (2002)).

2.1.3 Modelo restringido

Utilizando el modelo (2.15), la matriz de sumas de cuadrados total se puede expresar como

$$Y'Y = Y'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y + Y'\left[I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]Y$$

al centrar la información se encuentra que

$$\begin{aligned} Y'\left(I - \frac{1}{n}J\right)Y =& Y'\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \frac{1}{n}J\right]Y + \boldsymbol{R}_0 \\ T_{CM} =& M_{CM} + \boldsymbol{R}_0 \end{aligned}$$

donde $T_{CM} = Y'\left(I - \frac{1}{n}J\right)Y$ es la matriz de las sumas de cuadrados total corregida por la media y $M_{CM} = Y'\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \frac{1}{n}J\right]Y$ es la matriz de sumas de cuadrados del modelo corregido por la media. Además se puede comprobar que

$$\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \frac{1}{n}J\right]\left[I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] = 0$$

lo cual también sucede en el caso univariante, es decir que las matrices de sumas de cuadrados del modelo y de los residuos son ortogonales.

Bajo los supuestos del modelo (2.15) se tiene que $M_{CM} \sim Wishart_m(\Sigma, k)$.

Por otro lado, considérese el modelo (2.15) sujeto a la restricción

$$H\boldsymbol{B} = D \tag{2.19}$$

donde H, $\boldsymbol{B}$ y D tienen dimensiones $s \times (k+1)$, $(k+1) \times m$ y $(s \times m)$ respectivamente.

Entonces haciendo la minimización bajo la restricción, utilizando la matriz de multiplicadores de Lagrange Λ, se encuentra

$$\begin{aligned} L_1 =& tr\left[(Y - \boldsymbol{X}\boldsymbol{B})'(Y - \boldsymbol{X}\boldsymbol{B}) - 2\Lambda(H\boldsymbol{B} - D)\right] \\ =& tr\left[Y'Y - \boldsymbol{B}'\boldsymbol{X}'Y - Y'\boldsymbol{X}\boldsymbol{B} + \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B} - 2\Lambda H\boldsymbol{B} + 2\Lambda D\right] \end{aligned}$$

Derivando parcialmente con respecto a $\boldsymbol{B}$ y Λ, se obtiene

i. $\frac{\partial L_1}{\partial \boldsymbol{B}} = -2\boldsymbol{X}'Y + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B} - 2H'\Lambda' = 0$

ii. $\frac{\partial L_1}{\partial \Lambda} = -2(H\boldsymbol{B} - D)' = 0 \Rightarrow H\boldsymbol{B} = D$

Del ítem i. se tiene que

$$\begin{aligned} \widehat{\boldsymbol{B}}_{r_1} =& (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'Y + H'\Lambda') \\ =& (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y + (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda' \\ =& \widehat{\boldsymbol{B}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda' \end{aligned} \tag{2.20}$$

reemplazando esta última expresión en ii. se encuentra

$$H\widehat{\boldsymbol{B}}_{r_1} = H\widehat{\boldsymbol{B}} + H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda' = D$$

entonces

$$\begin{aligned} H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda' =& D - H\widehat{\boldsymbol{B}} \\ \Lambda' =& \left[H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\right]^{-1}\left(D - H\widehat{\boldsymbol{B}}\right) \end{aligned} \tag{2.21}$$

Por lo tanto, reemplazando (2.21) en (2.20), se obtiene

$$\widehat{\boldsymbol{B}}_{r_1} = \widehat{\boldsymbol{B}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\left[H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\right]^{-1}\left(D - H\widehat{\boldsymbol{B}}\right) \tag{2.22}$$

En muchas otras situaciones es de mayor interés trabajar con el modelo (2.15) bajo la restricción

$$H\boldsymbol{B}A = G \tag{2.23}$$

H es de orden $s \times (k+1)$ (de rango $s \leq k+1$), La matriz A es $m \times c$ (con rango $c \leq m \leq n-k-1$) y G es una matriz de orden $s \times c$ de constantes.

Al igual que en el modelo (2.15) bajo la restricción (2.23). Haciendo la minimización utilizando multiplicadores de Lagrange Λ_1 y Λ_2, se encuentra

$$\begin{aligned} L_2 =& tr\left[(Y - \boldsymbol{X}\boldsymbol{B})'(Y - \boldsymbol{X}\boldsymbol{B}) - 2\Lambda_1(H\boldsymbol{B}A - G)\Lambda_2\right] \\ =& tr\left[Y'Y - \boldsymbol{B}'\boldsymbol{X}'Y - Y'\boldsymbol{X}\boldsymbol{B} + \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B} - 2\Lambda_1 H\boldsymbol{B}A\Lambda_2 + 2\Lambda_1 G\Lambda_2\right] \end{aligned}$$

Derivando parcialmente con respecto a $\boldsymbol{B}$, Λ_1 y Λ_2, se obtiene

i. $\frac{\partial L_2}{\partial \boldsymbol{B}} = -2\boldsymbol{X}'Y + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B} - 2H'\Lambda_1'\Lambda_2'A' = 0$

ii. $\frac{\partial L_2}{\partial \Lambda_1} = -2\Lambda_2'(H\boldsymbol{B}A - G)' = 0 \Rightarrow H\boldsymbol{B}A = G$

iii. $\frac{\partial L_2}{\partial \Lambda_2} = -2(H\boldsymbol{B}A - G)'\Lambda_1' = 0 \Rightarrow H\boldsymbol{B}A = G$

Del ítem i. y haciendo $\Lambda_3 = \Lambda_2\Lambda_1$ se tiene que

$$\begin{aligned}\widehat{\boldsymbol{B}}_{r_2} =& (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'Y + H'\Lambda_3'A') \\ =& (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y + (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda_3'A' \\ =& \widehat{\boldsymbol{B}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda_3'A' \end{aligned} \tag{2.24}$$

reemplazando esta última expresión en ii. o iii., se encuentra

$$H\widehat{\boldsymbol{B}}_{r_2}A = H\widehat{\boldsymbol{B}}A + H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda_3'A'A = G$$

entonces

$$\begin{aligned} H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\Lambda_3'A'A =& G - H\widehat{\boldsymbol{B}}A \\ \Lambda_3' =& \left[H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\right]^{-1}\left(G - H\widehat{\boldsymbol{B}}A\right)(A'A)^{-1} \end{aligned} \tag{2.25}$$

Por lo tanto, al reemplazar (2.25) en (2.24), se llega a

$$\widehat{\boldsymbol{B}}_{r_2} = \widehat{\boldsymbol{B}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\left[H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\right]^{-1}\left(G - H\widehat{\boldsymbol{B}}A\right)(A'A)^{-1}A' \tag{2.26}$$

Obsérvese que si en (2.26), $A = I_m$ y $G = D$, entonces se obtiene la expresión (2.22).

2.1.4 Pruebas de hipótesis lineales

Retomando el modelo (2.15) y haciendo $Y_c = \left(I - \frac{1}{n}J\right)Y$, se obtiene el modelo

$$Y_c = X_{(k)}B + \Xi_k$$

La estimación mínimos cuadrados de B es $\hat{B} = \left(X_{(k)}'X_{(k)}\right)^{-1}X_{(k)}'Y_c$ y la matriz de predicción es $\hat{Y}_c = X_{(k)}\hat{B} = PY_c$ donde $P = X_{(k)}\left(X_{(k)}'X_{(k)}\right)^{-1}X_{(k)}'$ es la matriz sombrero. Claramente, no hay relación sí $B = 0$. Asumiendo $X_{(k)}$, Y_c centrados, una estadística apropiada para decidir sobre esta hipótesis nula está basada en

$$F = \frac{tr\left(Y_c'PY_c\right)/k}{tr\left(Y_c'\left(I - P\right)Y_c\right)/\left(n - k\right)}$$

donde $X_{(k)}$, ha sido obtenida por escalamiento métrico de una matriz de distancias, entonces $X_{(k)} = U_{x_{(k)}}\Lambda_{x_{(k)}}$ y $P = U_{x_{(k)}}U_{x_{(k)}}'$.

Como $P = P^2$ se tiene

$$tr\left(Y_c'PY_c\right) = tr\left(PY_cY_c'P\right)$$

Y similarmente $tr\left[Y_c'\left(I-P\right)Y_c\right] = tr\left[\left(I-P\right)Y_cY_c'\left(I-P\right)\right]$. Además, la razón F puede ser formulada en términos de distancias así

$$\begin{aligned} F &= \frac{tr\left(PY_cY_c'P\right)/k}{tr\left[\left(I-P\right)Y_cY_c'\left(I-P\right)\right]/\left(n-k\right)} \\ &= \frac{tr\left(F_x^-F_xY_cY_c'F_x^-F_x\right)/k}{tr\left[\left(I-F_x^-F_x\right)Y_cY_c'\left(I-F_x^-F_x\right)\right]\Big/\left(n-k\right)} \end{aligned}$$

donde $F_x^- = U_{x_{(k)}}\Lambda_{x_{(k)}}^{-2}U_{x_{(k)}}'$ es una g-inversa de F_x.

La prueba F se puede usar cuando hay un único vector de parámetros B y una sola columna Y_c con distribución normal. La prueba F puede aún ser usada cuando las filas de Y_c son multinormales con matriz de covarianzas $\Sigma = \sigma^2 I$.

En el caso general, para probar $B = 0$ se puede llevar a cabo una prueba de permutación. Para ejecutar esta prueba, se mantiene Y_c fijo, luego son halladas las $n!$ permutaciones de las filas de $X_{(k)}$ y es obtenida la distribución de aleatorización de F. Hay evidencia en contra de $B = 0$ si el F observado esta en el extremo de la cola. Si n es grande, se puede elegir una submuestra (con repetición) de las $n!$ permutaciones (Cuadras 2011).

Algunas pruebas basadas en la F cuando solamente Y_c viene de una distancia, han sido usadas por McArdle & Anderson (2001) relacionando datos ecológicos, y Wessel & Schork (2006) en estudios de asociación multilocus a gran escala. En Cuadras (2011) esta prueba ha sido adaptada a dos matrices de distancia. Sin embargo, esta aproximación F tiene cuatro inconvenientes. Primero, esta depende de $F_w = U_w\Lambda_w^2U_w'$ (como en la Subsección 2.1.1), es decir, sobre la matriz diagonal Λ_w, cuyas entradas son proporcionales a las varianzas de las columnas de W. Segundo, Λ_w puede tener entradas negativas si la matriz de distancias no es Euclidiana, como puede ocurrir en la presencia de datos faltantes en forma aleatoria. Tercero, si F es significante, se muestra una relación de dependencia pero no se sabe el grado de asociación entre ambos conjuntos de datos. Además, F no es simétrica en $X_{(k)}$ y W. Aquí la prueba F ha sido adaptada para cuando $X_{(k)}$ proviene de una matriz de distancias y Y_c no, ya que el interés en este trabajo es mirar el efecto del tiempo en los Y's. Sin embargo, es de notar que la propuesta hecha por Cuadras (2011) también se puede utilizar aquí porque se tendrían variables W que provienen de las distancias de la matriz Y, y estas nuevas variables serían independientes, dándole un mejor soporte al estadístico F.

Hipótesis de la forma $H\boldsymbol{B} = \mathbf{0}$

Una hipótesis lineal demostrable de rango s y matriz H es

$$H_0 : H\boldsymbol{B} = \mathbf{0} \tag{2.27}$$

donde las filas de H son combinación lineal de las filas de $\boldsymbol{X}$.

La ecuación (2.22) proporciona la estimación de $\boldsymbol{B}$ para el modelo (2.15) bajo la restricción (2.19). Restringido a la hipótesis (2.27), se puede mostrar que

$$\widehat{\boldsymbol{B}}_{r_1} = \widehat{\boldsymbol{B}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}H'\left[H(\boldsymbol{X}'\boldsymbol{X})^{-1}H'\right]^{-1}H\widehat{\boldsymbol{B}}$$

y la matriz residual es

$$\boldsymbol{R}_1 = \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}_{r_1}\right)'\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}_{r_1}\right)$$

***Teorema* 2.2.** *Sea el modelo lineal multivariante* (2.15)*, donde las filas de* Ξ_k *son* $NM_m(0, \Sigma)$ *independientes,* $\boldsymbol{R}_0$ *la matriz de residuos,* $H_0 : H\boldsymbol{B} = \mathbf{0}$*, una hipótesis lineal demostrable y* $\boldsymbol{R}_1$ *la matriz de residuos bajo* H_0*. Se verifica*

1. $\boldsymbol{R}_0 \sim W_m(\Sigma, n-k)$.
2. *Si* H_0 *es cierta, las matrices* $\boldsymbol{R}_0$ *y* $\boldsymbol{R}_1 - \boldsymbol{R}_0$ *siguen la distribución de Wishart con* $\boldsymbol{R}_1 \sim W_m(\Sigma, n-k')$, $\boldsymbol{R}_1 - \boldsymbol{R}_0 \sim W_m(\Sigma, s)$ *siendo* $s = rang(H)$ *y* $k' = k - s$.
3. *Si* H_0 *es cierta, las matrices* $\boldsymbol{R}_0$ *y* $\boldsymbol{R}_1 - \boldsymbol{R}_0$ *son estadísticamente independientes.*

Para la demostración de este teorema se siguieron los pasos de Cuadras (2010). Veamos la prueba a continuación

Demostración: Si la hipótesis nula H_0 es cierta, el subespacio generado por las filas de H está contenido en el generado por las filas de $\boldsymbol{X}$. Se puede construir una base ortogonal de $\Re^m$

$$[u_1, \ldots, u_s, u_{s+1}, \ldots, u_k, u_{k+1}, \ldots, u_k]$$

tal que $[u_1, \ldots, u_s]$ generen H, y $[u_1, \ldots, u_s, u_{s+1}, \ldots, u_k]$ generen $\boldsymbol{X}$.

Considérese ahora la matriz C de orden $m \times (k-s)$ generada por $[u_{s+1}, \ldots, u_k]$. Entonces $HC = 0$ y el modelo $Y = \boldsymbol{X}\boldsymbol{B} + \Xi_k$ se convierte en $Y = \tilde{X}\Theta + \Xi_k$, siendo $\tilde{X} = \boldsymbol{X}C$ y $C\Theta = \boldsymbol{B}$ pues $H\boldsymbol{B} = HC\Theta = 0$. Así la matriz $\boldsymbol{X}$ se transforma en $\tilde{X} = \boldsymbol{X}C$, donde las columnas de $\boldsymbol{X}C$ son combinación lineal de las columnas de $\boldsymbol{X}$.

Es posible construir una matriz ortogonal

$$T = [t_1, \ldots, t_{k'}, t_{k'+1}, \ldots, t_k, t_{k+1}, \ldots, t_n]$$

tal que las $k' = k - s$ primeras columnas generen $\boldsymbol{X}C$ y las k primeras generen $\boldsymbol{X}$,

$$C_{k'}(\boldsymbol{X}C) = [t_1, \ldots, t_{k'}] \subset C_k(\boldsymbol{X}) = [t_1, \ldots, t_k]$$

Siguiendo los mismos argumentos de la prueba del teorema 2.1 se tiene que

$$T'\widehat{\Xi}_k = \begin{bmatrix} \mathbf{0} \\ Z_{n-k'} \end{bmatrix}$$

donde las $n - k'$ filas de $Z_{n-k'}$ son $NM_m(\mathbf{0}, \Sigma)$ independientes. Por tanto,

$$\boldsymbol{R}_1 = \left(Y - \tilde{X}\widehat{\Theta}\right)' \left(Y - \tilde{X}\widehat{\Theta}\right) = Z'_{n-k'} Z_{n-k'} \sim W_m(\Sigma, n - k')$$

Por otro lado, se puede escribir

$$T' \left(Y - \tilde{X}\widehat{\Theta}\right) = \begin{bmatrix} \mathbf{0} \\ Z_{n-k'} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ Z_s \\ Z_{n-k} \end{bmatrix}$$

donde las $s = k - k'$ filas de Z_s son independientes de las $n - k$ filas de Z_{n-k}. Entonces $\boldsymbol{R}_1 = Z'_s Z_s + Z'_{n-k} Z_{n-k}$ es decir,

$$\boldsymbol{R}_1 - \boldsymbol{R}_0 = Z'_s Z_s \sim W_m(\Sigma, n - k')$$

e independiente de $\boldsymbol{R}_0 = Z'_{n-k} Z_{n-k}$. □

La consecuencia más importante del teorema 2.2 es que, si la hipótesis nula H_0 es cierta, entonces $\boldsymbol{R}_0$ y $\boldsymbol{R}_1 - \boldsymbol{R}_0$ son Wishart independientes y

$$\Lambda_{W_1} = \frac{|\boldsymbol{R}_0|}{|(\boldsymbol{R}_1 - \boldsymbol{R}_0) + \boldsymbol{R}_0|} = \frac{|\boldsymbol{R}_0|}{|\boldsymbol{R}_1|} \sim \Lambda(m; n - k, s)$$

Así $0 \leq \Lambda_{W_1} \leq 1$ sigue la distribución de Wilks. No se rechazará H_0 si Λ no es significativo y se rechazará H_0 si Λ es pequeño y significativo.

Otra forma de expresar el criterio de Wilks es

$$\Lambda_{W_1} = \frac{|\boldsymbol{R}_0|}{|\boldsymbol{R}_1|} = \mu_1 \times \cdots \times \mu_m$$

donde $\mu_1 \geq \cdots \geq \mu_m$ son los valores propios de $\boldsymbol{R}_0$ respecto de $\boldsymbol{R}_1$.

Este criterio es especialmente útil, teniendo en cuenta que si λ es la razón de verosimilitud en la prueba de hipótesis, entonces $\lambda = \Lambda_{W_1}^{n/2}$ (Cuadras 2010).

Es posible mostrar que cualquier estadístico que sea invariante por cambios de origen y de escala de los datos, debe ser función de estos valores propios (Anderson 2003). Así otros estadísticos propuestos son

1. Traza de Lawley-Hotelling:

$$tr\left((\boldsymbol{R}_1 - \boldsymbol{R}_0)\boldsymbol{R}_0^{-1}\right) = \sum_{i=1}^{m} \frac{1-\mu_i}{\mu_i}$$

2. Traza de Pillai:

$$tr\left((\boldsymbol{R}_1 - \boldsymbol{R}_0)\boldsymbol{R}_1^{-1}\right) = \sum_{i=1}^{m} (1-\mu_i)$$

3. Raíz mayor de Roy:

$$\frac{1-\mu_m}{\mu_m}$$

Hipótesis de la forma $H\boldsymbol{B}A = G$

Como en los modelos de datos longitudinales es de interés el tiempo, la hipótesis general se puede plantear como

$$H_0 : H\boldsymbol{B}A = G \tag{2.28}$$

La matriz H tiene los coeficientes que permiten probar la hipótesis "dentro de tiempos" (es decir, la hipótesis en los elementos dentro de las columnas dadas de $\boldsymbol{B}$). La matriz A permite probar la hipótesis "entre tiempos" (es decir, la hipótesis en los elementos dentro de las filas dadas de $\boldsymbol{B}$). Finalmente G es una matriz de constantes. Esta forma de escribir las hipótesis es muy general, un caso particular es cuando $H = I_k$, $A = I_m$ y todos los elementos de G son iguales a cero.

En este caso, si el modelo (2.15) es restringido con la condición (2.23) se obtiene la estimación de $\boldsymbol{B}$ mediante la ecuación (2.26). La matriz de residuales viene dada por

$$\begin{aligned}
\boldsymbol{R}_2 &= \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}_{r_2}\right)' \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}_{r_2}\right) \\
&= \left[\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right) + \boldsymbol{X}\left(\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}_{r_2}\right)\right]' \left[\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right) + \boldsymbol{X}\left(\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}_{r_2}\right)\right] \\
&= \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)' \left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right) + \left(\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}_{r_2}\right)' \boldsymbol{X}'\boldsymbol{X} \left(\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}_{r_2}\right)
\end{aligned}$$

ya que $\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right)' \boldsymbol{X} = 0$ y $\boldsymbol{X}'\left(Y - \boldsymbol{X}\widehat{\boldsymbol{B}}\right) = 0$. Por lo tanto,

$$\boldsymbol{R}_2 = \boldsymbol{R}_0 + \boldsymbol{R}_2^* \tag{2.29}$$

donde $\boldsymbol{R}_0$ fue presentado en (2.18) y,

$$\begin{aligned}\boldsymbol{R}_2^* =& \left(\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}_{r_2}\right)' \boldsymbol{X}'\boldsymbol{X} \left(\widehat{\boldsymbol{B}} - \widehat{\boldsymbol{B}}_{r_2}\right) \\ =& \left\{\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H' \left[H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H'\right]^{-1} \left(H\widehat{\boldsymbol{B}}A - G\right)\left(A'A\right)^{-1} A'\right\}' \left(\boldsymbol{X}'\boldsymbol{X}\right) \\ & \left\{\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H' \left[H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H'\right]^{-1} \left(H\widehat{\boldsymbol{B}}A - G\right)\left(A'A\right)^{-1} A'\right\} \\ =& A\left(A'A\right)^{-1} \left(H\widehat{\boldsymbol{B}}A - G\right)' \left[H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H'\right]^{-1} \left(H\widehat{\boldsymbol{B}}A - G\right) \\ & \left(A'A\right)^{-1} A' \end{aligned} \tag{2.30}$$

Sustituyendo (2.17) en (2.30), expandiendo y tomando valor esperado, se obtiene

$$\begin{aligned}E\left(\boldsymbol{R}_2^*\right) =& E\Big\{A\left(A'A\right)^{-1} \left(H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'YA - G\right)' \left[H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H'\right]^{-1} \\ & \left(H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'YA - G\right)\left(A'A\right)^{-1} A'\Big\}\end{aligned}$$

y reemplazando por (2.15), se llega a

$$\begin{aligned}E\left(\boldsymbol{R}_2^*\right) =& A\left(A'A\right)^{-1} \left(HBA - G\right)' \left[H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H'\right]^{-1} \left(HBA - G\right)\left(A'A\right)^{-1} A' \\ & + L'\Xi_k' K \Xi_k L \end{aligned} \tag{2.31}$$

donde $L = A\left(A'A\right)^{-1} A'$ y haciendo $K = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H' \left[H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} H'\right]^{-1} H\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'$, se cumple que $L^2 = L = L'$ y $K^2 = K = K'$, por lo que L y K son proyecciones. Adicionalmente, la j-ésima entrada en la matriz $\Xi_k' K \Xi_k$ de orden $m \times m$ en (2.31) es la forma cuadrática

$$\Xi_{(j)}' K \Xi_{(j')} = \sum_u \sum_v K_{uv} \Xi_{ju} \Xi_{j'v}$$

donde $\Xi_{(j)} = (\Xi_{ju})$ es la j-ésima fila de Ξ_k. De este modo, su valor esperado es

$$E\left(\Xi_{(j)}' K \Xi_{(j')}\right) = \sum_u K_{uu} (\Sigma)_{jj'} = (\Sigma)_{jj'} tr(K)$$

entonces

$$E(\Xi' K \Xi) = s\Sigma$$

porque $tr(K) = tr(I_s) = s$.

Finalmente, bajo $H_0 : H\boldsymbol{B}A = G$, $E(\boldsymbol{R}_2^*/s) = L'\Sigma L$ y además $E(\boldsymbol{R}_0/(n-k)) = \Sigma$. Una prueba de significancia para juzgar $H_0 : H\boldsymbol{B}A = G$ vs $H_a : H\boldsymbol{B}A \neq G$ a través de una función (es decir, determinante, traza, o máximo valor propio) de la cantidad

$$L'\boldsymbol{R}_2^* L \left(L'\boldsymbol{R}_0 L\right)^{-1}$$

donde se hace uso del hecho que L es una matriz de proyección.

Las cuatro pruebas empleadas para este tipo de análisis, al igual que antes, son

1. Lambda de Wilks. La estadística de razón de verosimilitud es

$$\Lambda_{W_2} = \frac{|\boldsymbol{R}_0|}{|\boldsymbol{R}_2^* + \boldsymbol{R}_0|} = \frac{|\boldsymbol{R}_0|}{|\boldsymbol{R}_2|} = \mu_1^* \times \cdots \times \mu_m^*$$

   donde $\mu_1^* \geq \cdots \geq \mu_m^*$ son los valores propios de $\boldsymbol{R}_0$ respecto de $\boldsymbol{R}_2$. Se rechaza la hipótesis nula si $\Lambda_{W_2} < \Lambda_{(\alpha,m,n-k,s)}$.

2. Traza de Lawley-Hotelling

$$tr\left((\boldsymbol{R}_2 - \boldsymbol{R}_0)\boldsymbol{R}_0^{-1}\right) = \sum_{i=1}^{m} \frac{1-\mu_i^*}{\mu_i^*}$$

   Se rechaza la hipótesis nula si este valor es más grande que una cantidad que depende de m, $n-k$ y s. La distribución exacta de la anterior estadística no es sencilla, razón por la cual se utiliza la aproximación a la estadística F.

3. Traza de Pillai

$$tr\left((\boldsymbol{R}_2 - \boldsymbol{R}_0)\boldsymbol{R}_2^{-1}\right) = \sum_{i=1}^{m}(1-\mu_i^*)$$

   Asintóticamente se ha demostrado que esta estadística tiene distribución chi-cuadrado con s grados de libertad para tomar la decisión de rechazar o no H_0.

4. Raíz mayor de Roy

$$\frac{1-\mu_m^*}{\mu_m^*}$$

En muchos de los casos las distribuciones nulas exactas para estas estadísticas no se pueden calcular, por lo tanto, se requiere de pruebas aproximadas. En la mayoría de los programas estadísticos, tales como R, SAS y SPSS la aproximación a la estadística F se utiliza.

2.2 Aproximación basada en distancias en asociación multivariante

Muchos coeficientes han sido propuestos para medir la asociación multivariante entre dos vectores aleatorios o dos conjuntos de datos tomados sobre los mismos

individuos. En ecología se tiene un ejemplo claro, donde datos ambientales están relacionados a especies. En datos genómicos se pueden ver relaciones entre genotipos (es decir, datos de DNA) a fenotipos de interés. Se pueden también encontrar muchos ejemplos en biometría y psicología, es decir, para relacionar características físicas con pruebas mentales (Cuadras 2011).

En el caso de este trabajo la idea es relacionar un conjunto de datos tomados en el tiempo Y con un conjunto de variables explicativas que son fijas V, como se muestra en el modelo (2.1). En este sentido algunas medidas de dependencia basadas en correlaciones canónicas pueden ser usadas. Sin embargo, si los conjuntos de datos $X_{(k)}$ y W no son cuantitativos (binarios, categóricos y nominales), la información puede alternativamente ser dada por una similaridad o una matriz de distancias. Esta aproximación basada en distancias, originada en Cuadras (1989), ha sido usada como una herramienta en predicción y análisis multivariante (ver (Bartkowiak & Jakimiec 1994) y (Boj, Claramunt & Fortiana 2007)).

La medida de asociación multivariante entre $X_{(k)}$, W esta dada por

$$\eta^2(X_{(k)}, W) = \left|U'_{x_{(k)}} U_w U'_w U_{x_{(k)}}\right| = \left|U'_w U_{x_{(k)}} U'_{x_{(k)}} U_w\right|$$

Estas medidas satisfacen $0 \leq \eta(X_{(k)}, W) = \eta(W, X_{(k)}) \leq 1$ y se reducen al coeficiente de correlación múltiple cuando W es un vector de datos. Sea $S_{11} = X'_{(k)}X_{(k)}$, $S_{12} = X'_{(k)}W$, $S_{22} = W'W$. Como las correlaciones canónicas positivas r_i, $i = 1, \ldots, s = \min\{k, t\}$ entre $X_{(k)}$ y W son los valores singulares de $S_{11}^{-1/2} S_{12} S_{22}^{-1/2} = U'_{x_{(k)}} U_w$, este coeficiente puede ser expresado como

$$\eta(X_{(k)}, W) = \prod_{i=1}^{s} r_i$$

Ya que $U'_{x_{(k)}} U_w$ es una matriz de Gram, la medida η puede ser interpretada como el coseno del ángulo entre dos subespacios expandidos por $U_{x_{(k)}}$ y U_w.

2.2.1 Medidas de asociación multivariante

Otro criterio usado para juzgar $\boldsymbol{B} = 0$ en el modelo de regresión lineal multivariante es el criterio de razón de verosimilitud o lambda de Wilks, la cual es bien conocida en análisis multivariante (Mardia et al. 1979). La lambda de Wilks es

$$\Lambda_{W_1} = \frac{|\boldsymbol{R}_0|}{|(\boldsymbol{R}_1 - \boldsymbol{R}_0) + \boldsymbol{R}_0|} = \frac{|\boldsymbol{R}_0|}{|\boldsymbol{R}_1|} \sim \Lambda(m; n-k, s)$$

donde $\boldsymbol{R}_0 = \left(Y - \boldsymbol{X}\hat{\boldsymbol{B}}\right)' \left(Y - \boldsymbol{X}\hat{\boldsymbol{B}}\right) = \Lambda_w U'_w \left(I - U_{x_{(i)}} U'_{x_{(i)}}\right) U_w \Lambda_w$ y $\boldsymbol{R}_1 = \left(Y - \boldsymbol{X}\hat{\boldsymbol{B}}_{r_1}\right)' \left(Y - X\hat{\boldsymbol{B}}_{r_1}\right) = \Lambda_w U'_w U_w \Lambda_w$, con $\hat{\boldsymbol{B}}_{r_1} = \hat{\boldsymbol{B}} -$

$(\boldsymbol{X}'\boldsymbol{X})^{-1}H'[H(\boldsymbol{X}'\boldsymbol{X})^{-1}H']^{-1}H\hat{\boldsymbol{B}}$. Por lo tanto,

$$\Lambda_{W_1} = \left|I - U_w' U_{x_{(i)}} U_{x_{(i)}}' U_w\right|$$

La lambda de Wilks no depende de Λ_w y puede expresarse en términos de correlaciones canónicas, es decir

$$\Lambda_{W_1} = \prod_{i=1}^{s} \left(1 - r_i^2\right)$$

Claramente $A_W = 1 - \Lambda_{W_1}$ es también una medida de asociación tal que se aproxima a 0 si $\boldsymbol{X}$, W son independientes y se aproxima a 1 si $\boldsymbol{X}$, W son linealmente dependientes.

Para juzgar $\boldsymbol{B} = 0$ se puede emplear otros criterios tales como Lawley-Hotelling y Pillai (Cuadras 2011). Si $(\boldsymbol{R}_1 - \boldsymbol{R}_0)u_{w_i} = \mu_i \boldsymbol{R}_0 u_{w_i}$, arroja los valores propios de $\boldsymbol{R}_0^{-1}(\boldsymbol{R}_1 - \boldsymbol{R}_0)$, siendo $\mu_i = r_i^2/\left(1 - r_i^2\right)$, los criterios $U_{x_{(i)}}$ de Lawley-Hotelling y el criterio U_w de Pillai's son

$$\mathsf{U}_x = tr\left[\boldsymbol{R}_0^{-1}(\boldsymbol{R}_1 - \boldsymbol{R}_0)\right] = \sum_{i=1}^{s} \frac{r_i^2}{1 - r_i^2}$$

$$\mathsf{U}_y = tr\left[(\boldsymbol{R}_1)^{-1}(\boldsymbol{R}_1 - \boldsymbol{R}_0)\right] = \sum_{i=1}^{s} r_i^2$$

Entonces dos medidas de asociación multivariante pueden estar basadas en $A_{LH} = (\mathsf{U}_x/s)\,\big/\,(1 + \mathsf{U}_x/s)$ y $A_p = \mathsf{U}_w/s$. Para la derivación de U_x, U_w y Λ_{W_1} (ver Anderson (2003) y Rao (1973)).

La medida A_p también surge aplicando la correlación vectorial entre dos vectores aleatorios $\boldsymbol{X}$, W, definida por (ver Escoufier (1973))

$$RV = \frac{tr\left(S_{12}S_{21}\right)}{\sqrt{tr(S_{11}^2)tr(S_{22}^2)}}$$

donde $S_{11} = X_{(k)}'X_{(k)}$, $S_{12} = X_{(k)}'W$, etc. Si se toman las columnas estandarizadas, entonces $S_{xy} = U_{x_{(k)}}'U_w$, $S_{11} = U_{x_{(k)}}'U_{x_{(k)}} = I_k$, $S_{22} = U_w'U_w = I_t$ y esta correlación se reduce a $RV = \sum_{i=1}^{s} r_i^2 \Big/ \sqrt{kt}$. Claramente $RV = A_p$ si $k = t$. Sin embargo, si $k < t$ hay $t-k$ correlaciones cero, además $A_p = \sum_{i=1}^{s} r_i^2 \Big/ s$ es mejor. En general $RV \neq A_p$.

Otra medida de asociación, la cual generaliza el coeficiente de correlación múltiple, está dada por

$$R^2 = \frac{\left|S_{21}S_{11}^{-1}S_{12}\right|}{|S_{22}|}$$

Simplificando, se obtiene $R^2 = \left|U'_w U_{x_{(k)}} U'_{x_{(k)}} U_w\right| = \prod_{i=1}^{s} r_i^2$, la cual se indica por A_{HC} (ver Cramer & Nicewander (1979)).

También, es posible relacionar X_k y W vía el estadístico Procrustes (ver Cox & Cox (2001))

$$P^2 = 1 - \frac{\left[tr\left(X'_{(k)} W W' X_{(k)}\right)^{1/2}\right]^2}{tr\left(X'_{(k)} X_{(k)}\right) tr\left(W'W\right)}$$

En el contexto basado en distancias se obtiene,

$$P^2 = 1 - \frac{\left[tr\left(\Lambda_x U'_{x_{(k)}} U_w \Lambda_w^2 U'_w U_{x_{(k)}}\right)^{1/2} \Lambda_x\right]^2}{tr\left(\Lambda_x^2\right) tr\left(\Lambda_w^2\right)}$$

Estandarizando las variables esta ecuación se reduce a

$$P^2 = 1 - \left[tr\left(U'_{x_{(k)}} U_w U'_w U_{x_{(k)}}\right)^{1/2}\right]^2 \Big/ (kt) = 1 - \left(\sum_{i=1}^{s} r_i\right)^2 \Big/ (kt)$$

Esta medida sugiere $A_{PR} = \left(\sum_{i=1}^{s} r_i\right)^2 \Big/ s^2$.

Arenas & Cuadras (2004) proponen la medida de asociación

$$A_{AC} = \frac{tr\left(G_x^{1/2} G_w^{1/2} + G_w^{1/2} G_x^{1/2}\right)}{tr\left(G_x + G_w\right)}$$

la cual cae entre 0 y 1. Donde $G_x = U_{x_{(k)}} \Lambda_x^2 U'_{x_{(k)}}$ y $G_w = U_w \Lambda_w^2 U'_w$. Estandarizando, $tr\left(U_{x_{(k)}} U'_{x_{(k)}} U_w U'_w\right) = tr\left(U'_{x_{(k)}} U_w U'_w U_{x_{(k)}}\right)$, esta ecuación se reduce a

$$\begin{aligned} A_{AC} &= tr\left(U_{x_{(k)}} U'_{x_{(k)}} U_w U'_w + U_w U'_w U_{x_{(k)}} U'_{x_{(k)}}\right) / (k+t) \\ &= 2tr\left(U'_{x_{(k)}} U_w U'_w U_{x_{(k)}}\right) / (k+t) \\ &= 2\sum_{i=1}^{s} r_i^2 \Big/ (k+t) \end{aligned}$$

lo cual sugiere $A_{AC} = \sum_{i=1}^{s} r_i^2 \Big/ s = A_p$.

2.2.2 Predicción de un nuevo individuo

Si se supone que en las variables mixtas explicativas se ha observado un nuevo individuo $n+1$ del que se conoce las observaciones sobre las variables independientes, $v_{n+1} = (v_{(n+1)0}, v_{(n+1)1}, v, v_{(n+1)p})'$; tales observaciones permiten calcular las distancias entre el individuo $n+1$ y cada uno de los individuos que intervinieron en el modelo planteado en (2.1), es decir,

$$\delta_{(n+1)i} = \delta(v_{n+1}, v_i), \quad v_i \in \Omega, \quad i = 1, \ldots, n$$

A partir de estas distancias se puede hacer una predicción empleando el siguiente resultado (Gower 1968, Cuadras & Arenas 1990), que relaciona el vector $d = (\delta^2_{(n+1)1}, \ldots, \delta^2_{(n+1)n})'$ de los cuadrados de estas distancias con el vector $x_{n+1} = (x_{(n+1)1}, \ldots, x_{(n+1)p})'$ de las coordenadas principales atribuibles al nuevo individuo.

$$\begin{aligned} \delta^2_{(n+1)i} =& (x_{n+1} - x_i)'(x_{n+1} - x_i) \\ =& x'_{n+1}x_{n+1} + x'_i x_i - 2x'_{n+1}x_i \end{aligned} \tag{2.32}$$

Sumando para i de 1 a n, y teniendo en cuenta que las columnas de la matriz X de coordenadas suman 0, se obtiene

$$\sum_{i+1}^{n} \delta^2_{(n+1)i} = n x'_{n+1}x_{n+1} + tr(\mathtt{B})$$

Sustituyendo esta última ecuación en (2.32), se tiene

$$2x_{n+1}x_i = \frac{1}{n}\left(\sum_{i=1}^{n} \delta^2_{(n+1)i} - tr(\mathtt{B})\right) + b_{ii} - \delta^2_{(n+1)i}$$

Al considerar las diferencias con los n individuos de forma matricial

$$2Xx_{n+1} = \frac{1}{n}\left(\sum_{i=1}^{n} \delta^2_{(n+1)i} - tr(\mathtt{B})\right) 1_n + (b - d)$$

donde $b = (b_{11}, \ldots, b_{nn})$ con $b_{ii} = x'_i x_i$, $i = 1, \ldots, n$.

Premultiplicando por X' y dado que $X'1_n = 0$, se encuentra que

$$\begin{aligned} 2X'Xx_{n+1} =& X'(b - d) \\ x_{n+1} =& \frac{1}{2}(X'X)^{-1}X'(b - d) \\ =& \frac{1}{2}\Lambda_x^{-1}X'(b - d) \end{aligned} \tag{2.33}$$

La predicción es entonces

$$\widehat{Y}_{(n+1)} = \widehat{B}_0 + x'_{n+1}\widehat{B}$$

Si se considera ahora el modelo DB en dimensión k y se hace la partición

$$x_{n+1} = \begin{pmatrix} x_{(k)} \\ l \end{pmatrix}, \quad X = \begin{pmatrix} X_{(k)} & L \end{pmatrix} \quad \text{y} \quad \Lambda_x = \begin{pmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{p-k} \end{pmatrix}$$

donde $x_{(k)} = (x_1, \ldots, x_k)'$ son las k coordenadas relativas de las k-dimensiones predictivas asociadas al $n+1$ individuo y la diagonal Λ_k contiene los valores propios, así se obtiene

$$\widehat{Y}_{(n+1)} = \widehat{B}_0 + x'_{(k)}\widehat{B}_{(k)} + l'\widehat{B}_{(p-k)}$$

como l contiene las coordenadas menos correlacionadas en el nuevo individuo $n+1$, entonces

$$\widehat{Y}_{(n+1)}(k) = \widehat{B}_0 + x'_{(k)}\widehat{B}_{(k)} \tag{2.34}$$

Obsérvese que si l es muy grande, el nuevo individuo es un atípico, entonces $x_{(k)}$ y esta predicción puede no ser precisa.

Por otro lado, en Cuadras (2011), se muestra que las predicciones del modelo lineal clásico y el modelo basado en distancias coinciden. En la siguiente sección se extiende este resultado al modelo longitudinal multivariante con DB.

2.2.3 Relación con el modelo longitudinal clásico

El modelo (2.8) depende de la distancia elegida δ_{ij}. Por lo tanto, cuando las variables explicativas son continuas y la distancia Euclidiana se utiliza, el DB es compatible con el modelo clásico. Esta equivalencia también se muestra que se mantiene para variables cualitativas cuando un método basado en distancias con el coeficiente de coincidencias se utiliza.

Variables continuas

Si todas las variables explicativas en (2.1), $V = (V_1, \ldots, V_p)$, son continuas, la distancia Euclidiana esta dada por (2.2). Ahora, se prueba que el espacio generado por las columnas de X es el mismo que el generado por las columnas de V y por lo tanto, los modelos DB y clásico producen las mismas predicciones. Esto es, la relación entre F_x y $\Delta_v^{(2)}$ es

$$\begin{aligned} F_x =& \mathcal{H}A_v\mathcal{H} = -\frac{1}{2}\mathcal{H}\Delta_v^{(2)}\mathcal{H} = -\frac{1}{2}\mathcal{H}\left(\mathbf{1}f_v' + f_v\mathbf{1}' - 2VV'\right)\mathcal{H} \\ =& -\frac{1}{2}\left(\mathcal{H}\mathbf{1}f_v'\mathcal{H} + \mathcal{H}f_v\mathbf{1}'\mathcal{H} - 2\mathcal{H}VV'\mathcal{H}\right) \\ =& \mathcal{H}VV'\mathcal{H} = XX' \end{aligned} \tag{2.35}$$

porque $\mathcal{H}\mathbf{1}f_v'\mathcal{H} = 0$ y $\mathcal{H}f_v\mathbf{1}'\mathcal{H} = 0$, donde f_v es un vector de longitud n que contiene la diagonal de VV'. Cabe señalar que si $S_v = (s_{ii'})$ es la matriz de similaridad y la función distancia seleccionada es $\delta^2_{ii'} = s_{ii} + s_{i'i'} - 2s_{ii'}$, entonces $F_x = \mathcal{H}S_v\mathcal{H}$. Esto corresponde a escalamiento multidimensional clásico o ACP (Cuadras 1989).

Entonces, el modelo DB introducido en (2.1) es un modelo centrado (2.8), es decir, este produce las mismas predicciones en el modelo propuesto en dimensión p que el modelo dado en (2.1). Sin embargo, no sería necesario considerar una distancia Euclidiana p-dimensional. Sea E_{p^*} el espacio generado por las columnas de X, donde X es una solución del escalamiento métrico obtenido de una distancia aplicada a los mismos datos. Entonces, tomando $k > p$, es decir, las columnas más adecuadas de X, el modelo DB mejora el modelo longitudinal clásico cuando $\left(Y - \mathbf{1}\widehat{B}_0\right) \in E_{p^*}$. Teniendo en cuenta que esto es siempre verdadero para $p^* = n - 1$ con $p^* > p$.

Variables cualitativas

Supóngase ahora que todas las variables explicativas $V = (V_1, \ldots, V_p)$ son cualitativas en (2.1), donde ahora todos los V_j son variables en los q_j estados, $j = 1, \ldots, p$. Una medida de similaridad entre individuos i y i' es el número de coincidencias $m_{ii'}$ para las variables cualitativas involucradas en el modelo. Tener en cuenta que $0 \leq m_{ii'} \leq p$, y también $m_{ii'}/p$ es el coeficiente de coincidencias si las variables son binarias. En este caso, se eligen las distancias al cuadrado, es decir, $\delta^2_{ii'} = 2(p - m_{ii'})$. Sin embargo, como V_j se puede representar por q variables binarias, respectivamente, las cuales son codificadas como 0 y 1, $\delta^2_{ii'}$ es la distancia euclídea al cuadrado. Por lo tanto, el modelo DB se reduce al modelo longitudinal clásico para variables cualitativas al anotar los estados como 0 (ausente) y 1 (presente). No hay ninguna ventaja sobre el modelo longitudinal clásico, excepto que el problema de multicolinealidad puede resolverse automáticamente usando distancias. Por supuesto, la solución es diferente si otro tipo de distancia se elige.

Los resultados anteriores muestran que las predicciones son las mismas para ambos modelos, longitudinal clásico y DB. Sin embargo, hay diferencias para datos mixtos, como se muestra en las simulaciones mas adelante. Por lo tanto, en la siguiente sección se presenta una simulación donde se compara el método DB con el método clásico para analizar datos longitudinales en aproximación multivariante.

2.3 Simulación

2.3.1 Detalles de la simulación

Se crearon tres variables explicativas para la simulación de los modelos, una de tipo continuo, otra categórica y la última binaria. Estas variables fueron creadas generando muestras aleatorias a partir de tres distribuciones diferentes, para tamaños n de 50, 100 y 200 respectivamente. La variable continua fue muestreada de una distribución normal con media de 100 y varianza de 100; la variable categórica fue muestreada de una distribución multinomial asumiendo tres valores con probabilidades de 0.27, 0.53, 0.2 y la variable binaria de una distribución binomial con probabilidad de 0.4.

La matriz de parámetros $\boldsymbol{B}$ se obtuvo generando una muestra aleatoria a partir de una distribución normal con media 25 y varianza de 36 cuyo número de filas es igual al número de variables independientes más el intercepto, es decir, p y cuyo número de columnas es m, es decir el número de tiempos. La variable respuesta es continua y se generó a partir de los errores del modelo. Los errores del modelo fueron creados generando muestras aleatorias de una distribución normal multivariante con un vector de medias de ceros de tamaño igual al número m de tiempos y una matriz de covarianzas Ψ_0 generada a partir de dos estructuras de correlación; autorregresiva de orden uno AR(1) y compuesta simétrica (ver mayores detalles en la Subsección 3.1.1 y Rencher (2002) y Diggle et al. (2002)). Como los datos contienen variables mixtas, se utilizó la distancia de Gower para el método DB.

Por otro lado, para la simulación de Monte Carlo cada escenario se repite $N = 100$ veces y el software utilizado para el efecto es R versión 2.15 (R Development Core Team 2012). Los criterios de información AIC y BIC de cada modelo fueron también calculados.

2.3.2 Resultados y Discusión

Se simularon en total 126 escenarios con 4, 7 y 10 tiempos (m), varianzas $\sigma^2 = 10, 50$ y correlaciones $\rho = -0.5, 0, 0.5, 0.9$, las estructuras de correlación consideradas fueron dos AR(1) y compsymm, para tamaños de muestra n de 50, 100 y 200. En esta sección se presenta un resumen de los resultados obtenidos.

En las Tablas 2.1 y 2.2 se muestran los diferentes niveles de varianza σ^2, correlación ρ y estructuras de autocorrelación, donde el método MANOVA DB tiene valores de AIC mas bajos para muestras pequeñas ($n = 50$). En ambos métodos son similares los valores similares de BIC con muestras pequeñas, independiente del número de tiempos, varianza, correlación y estructura de autocorrelación que se tome. En muestras grandes, los AIC y BIC en el método

clásico MANOVA presentan buenos resultados bajo algunos escenarios de varianza, correlacción y estructura de correlacción. Además, resultados similares fueron obtenidos bajo los otros escenarios, los cuales se pueden ver en las Tablas A.6 a A.10, presentadas en el anexo A del trabajo y cuya interpretación es similar. Es de aclarar aquí que La estructura de correlación compuesta simétrica con valor ρ=-0.5 no se consideró ya que la matriz resulta ser singular.

Parámetros			$m = 7$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	-0.5	50	1754.87	1793.45	1765.43	1788.58
		100	3778.78	3824.29	3572.62	3599.92
		200	7322.05	7374.49	7174.46	7205.93
	0	50	1757.65	1796.23	1769.13	1792.28
		100	3777.65	3823.16	3574.70	3602.00
		200	7323.99	7376.44	7175.62	7207.08
	0.5	50	1755.87	1794.45	1765.43	1788.58
		100	3773.12	3818.63	3572.62	3599.92
		200	7325.77	7378.21	7174.46	7205.93
	0.9	50	1750.96	1789.54	1760.92	1784.07
		100	3764.79	3810.30	3567.25	3594.55
		200	7326.97	7379.41	7176.27	7207.73
50	-0.5	50	2306.33	2344.91	2328.73	2351.88
		100	4727.86	4773.37	4699.22	4726.53
		200	9441.99	9494.43	9427.67	9459.14
	0	50	2309.98	2348.56	2332.44	2355.58
		100	4728.89	4774.40	4701.30	4728.61
		200	9443.66	9496.10	9428.83	9460.29
	0.5	50	2306.83	2345.41	2328.73	2351.88
		100	4724.86	4770.37	4699.22	4726.53
		200	9443.96	9496.40	9427.67	9459.14
	0.9	50	2301.57	2340.15	2324.22	2347.37
		100	4717.53	4763.04	4693.85	4721.16
		200	9444.65	9497.09	9429.48	9460.95

Tabla 2.1: Simulación con estructura de correlación AR(1)

Los resultados que se presentan en las Tablas 2.1 y 2.2 representan un subconjunto de los resultados obtenidos en los diagramas de cajas de las Figuras 2.1 y 2.2. Los diagramas de caja de las dos figuras muestran los criterios de información AIC y BIC bajo ambas aproximaciones (métodos DB y clásico). En todos los diagramas de cajas, los números pares corresponden a la aproximación basada en distancias, mientras los números impares corresponden al método clásico. En la primera columna de gráficos de cajas en las dos figuras, las dos primeras cajas de izquierda a derecha corresponden a muestras de tamaño 50, las siguientes dos a muestras de tamaño 100, y las últimas dos, a muestras de tamaño 200. Además en todos los gráficos de esta columna se tiene una estructura de correlación AR(1). En las otras tres columnas de gráficos de

Parámetros			$m = 10$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	0	50	2502.51	2544.66	2523.18	2548.47
		100	5370.69	5419.768	5101.64	5131.09
		200	10429.75	10485.76	10241.09	10274.7
	0.5	50	2505.28	2547.43	2527.48	2552.76
		100	5363.78	5412.86	5094.74	5124.19
		200	10421.74	10477.74	10248.13	10281.74
	0.9	50	2504.25	2546.39	2527.78	2553.07
		100	5356.89	5405.97	5086.1	5115.55
		200	10418.79	10474.8	10249.33	10282.94
50	0	50	3291.77	3333.91	3327.90	3353.19
		100	6745.77	6794.84	6711.08	6740.52
		200	13472.72	13528.73	13459.97	13493.57
	0.5	50	3294.78	3336.92	3332.20	3357.48
		100	6738.33	6787.40	6704.18	6733.63
		200	13471.97	13527.98	13467.01	13500.61
	0.9	50	3293.92	3336.07	3332.50	3357.79
		100	6729.88	6778.96	6695.54	6724.99
		200	13470.7	13526.71	13468.21	13501.81

Tabla 2.2: Simulación con estructura de correlación compuesta simétrica

las dos figuras, las cuatro primeras cajas de izquierda a derecha corresponden a muestras de tamaño 50, las siguientes cuatro a muestras de tamaño 100, y las últimas cuatro, a muestras de tamaño 200. Las dos primeras cajas en cada bloque corresponden a estructuras de correlación AR(1) y las dos siguientes a la estructura de correlación compuesta simétrica.

Por ejemplo para los valores de $\rho = 0.5$, $m = 4$, y $\sigma^2 = 10$, en las Figuras 2.1 y 2.2, se muestra que el método basado en distancias es más eficiente en muestras pequeñas (tamaño 50) independiente de la estructura de autocorrelación, correlación y varianza, aunque cuando el tamaño de muestra crece, el modelo clásico presenta mejor ajuste. Otro ejemplo es para el caso que $m = 10$, $\sigma^2 = 50$ y $\rho = 0.9$ en donde ambos métodos (DB y clásico) son similares, teniendo en cuenta los diferentes tamaños de muestra, con varianza grande, correlación alta y sin importar la estructura de correlación que se escoja. Se pueden ver diferencias pequeñas con respecto al método clásico en los diferentes tamaños de muestra usando los criterios de AIC y BIC. Sin embargo, para muestras pequeñas ($n = 50$), hay una diferencia pequeña como se menciono antes, en la cual el método basado en distancias es más eficiente. Resultados similares se obtienen con otras configuraciones de m, σ^2 y ρ, utilizando los dos criterios de información AIC y BIC.

En la siguiente sección se presenta una aplicación, en donde se puede visualizar la metodología propuesta y su comparación con el método clásico para análisis de datos longitudinales MANOVA.

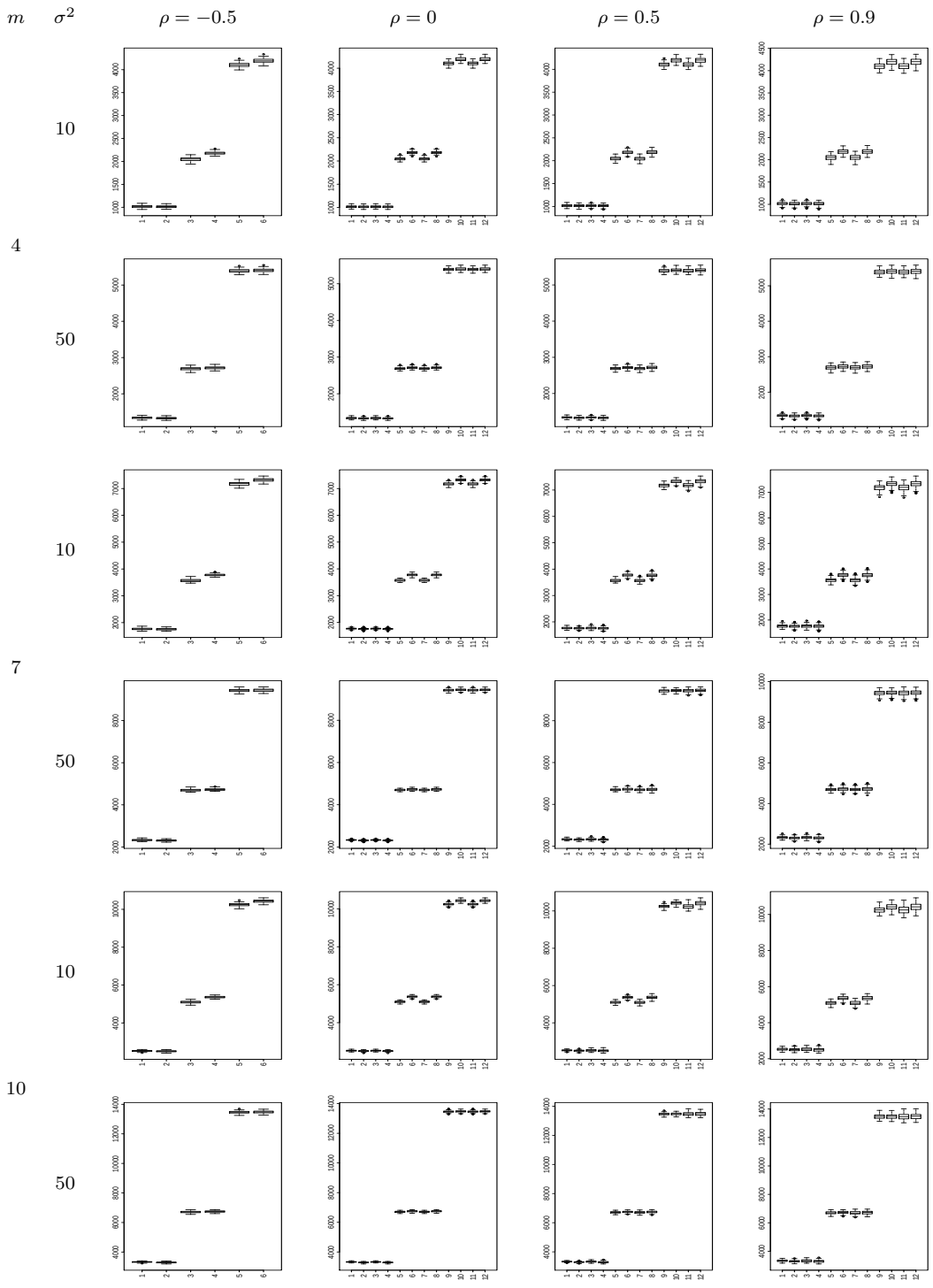


FIGURA 2.1: AIC para DB y análisis clásico por tamaño de muestra, en estructuras de autocorrelación AR(1) y compuesta simétrica

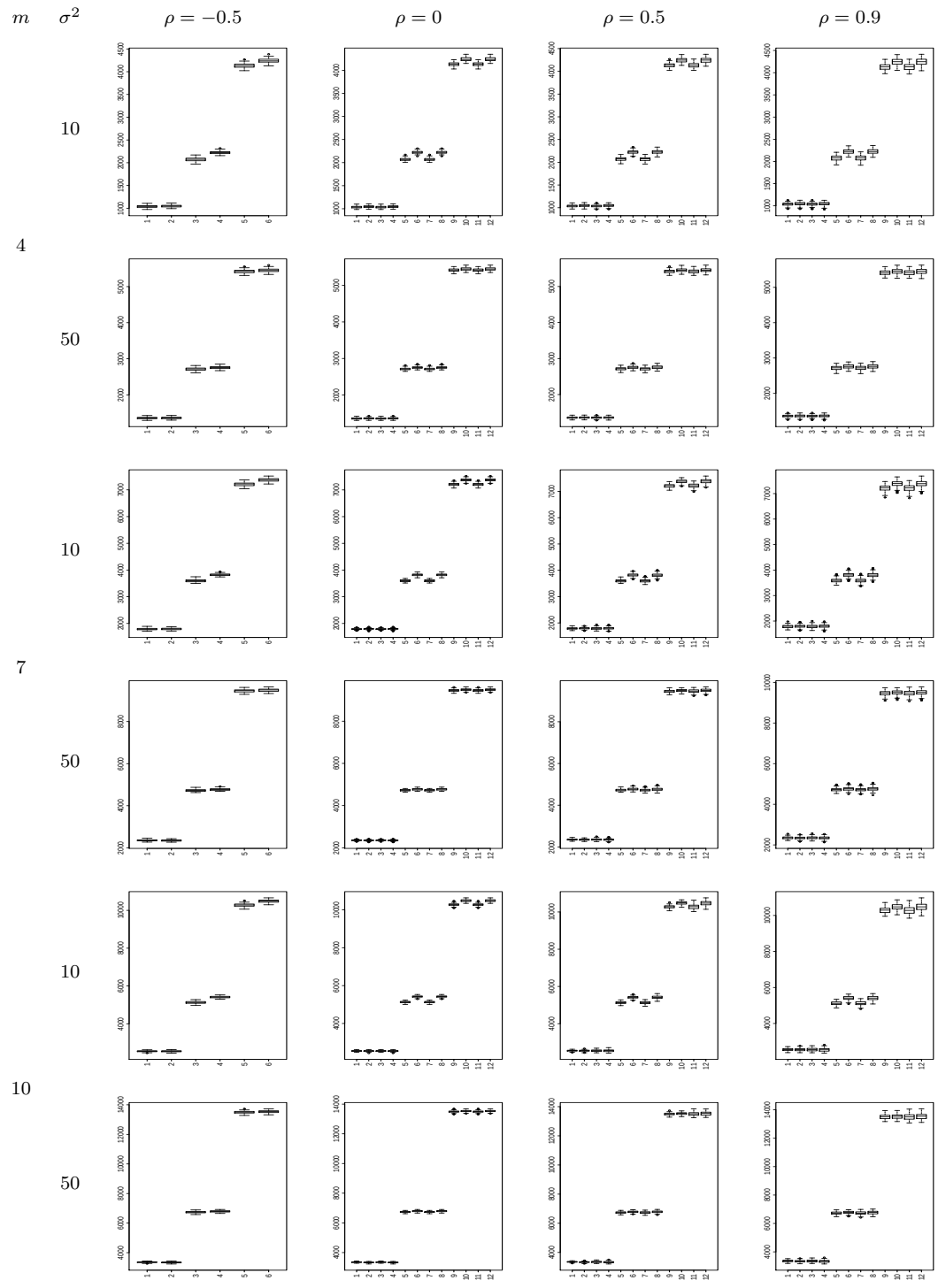


FIGURA 2.2: BIC para DB y análisis clásico por tamaño de muestra, en estructuras de autocorrelación AR(1) y compuesta simétrica

2.4 Aplicación

Se ilustran los desarrollos teóricos a través de los datos de la National Youth Survey (Raudenbush & Chan 1992, Raudenbush & Chan 1993). El estudio se realizó haciendo mediciones cada año, cuando los participantes tenían edades de 11, 12, 13, 14 y 15 años, estos completaron un instrumento de nueve elementos diseñado para evaluar su tolerancia respecto a la desviación de la conducta. Utilizando una escala de cuatro puntos (1=muy mala, 2=mala, 3=un poco mal, 4= no está mal del todo), se indicó si estaba mal para alguien de sus edades: (a) hacer trampa en las pruebas, (b) La destrucción deliberada de la propiedad de otros, (c) el consumo de marihuana, (d) robar algo que vale menos de cinco dólares, (e) golpear o amenazar a alguien sin razón, (f) el uso de alcohol, (g) entrar en un edificio o vehículo para robar, (h) vender drogas fuertes, o (i) robar algo que vale más de cincuenta dólares. En cada ocasión, el resultado, es decir la tolerancia, se calcula como el promedio de encuestados a través de las nueve respuestas. En los datos también se encuentran las variables explicativas de cambio en la tolerancia: género de los encuestados y exposición, la evaluación del entrevistado auto reporte, exposición a la desviación del comportamiento a los 11 años de edad. Para obtener los valores de esta última variable, los participantes estimaron la proporción de sus amigos cercanos que participaron en cada una de las mismas nueve actividades en otros cuatro puntos de escala (de 0 = ninguno, a 4 = todos). Al igual que la variable tolerancia, el valor de exposición de cada uno de los encuestados es el promedio de sus nueve respuestas. Se tomo una muestra aleatoria de 16 participantes de los más grandes estados de Nueva York (ver Singer & Willett (2003)).

Para ilustrar el método propuesto en este trabajo inicialmente se realizó un análisis exploratorio con los datos, donde se puede observar el crecimiento empírico, se puede apreciar que los cambios difieren considerablemente a través de los adolescentes, tal como se muestra en la Figura 2.3. Aunque la mayoría se vuelven más tolerantes de la conducta desviada con el tiempo (por ejemplo, los individuos 514 y 1653), muchos se mantienen relativamente estables(por ejemplo, los individuos 569 y 624), ninguno de los 16 se vuelve con el tiempo mucho menos tolerante (aunque en el sujeto 949 disminuye por un tiempo antes de aumentar).

Se observa de los gráficos de adolescentes que fueron más tolerantes a la desviación de la conducta en un tiempo que estos tienden a ser más tolerantes en la próxima medición. Esto indica que el orden de clasificación de los adolescentes sigue siendo relativamente estable en ocasiones. Si el puntaje de todos los adolescentes se redujo en un punto entre las edades 11 y 12 años, la correlación entre los tiempos es positiva. Se puede inferir una relación directa entre las correlaciones en los tiempos y el cambio. El análisis exploratorio de los datos muestra que no hay diferencia por género, pero en lo que respecta

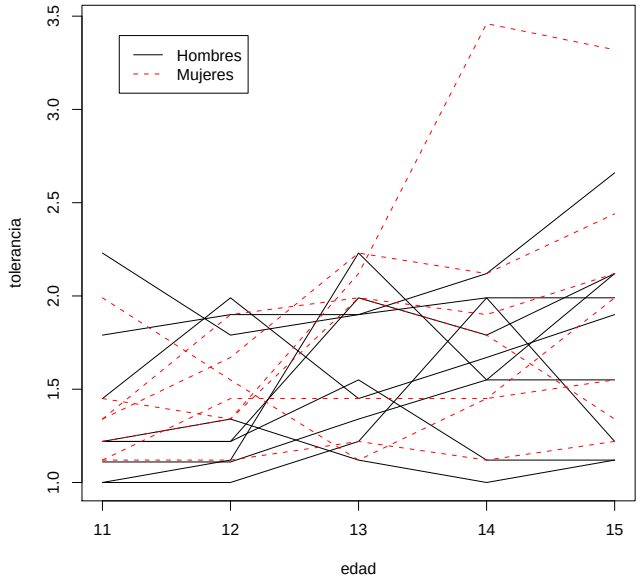


FIGURA 2.3: Gráfico de Tolerancia en función de la edad por género

a la exposición, parece que los adolescentes con mayor exposición temprana a la desviación del comportamiento se vuelven más tolerantes a un ritmo más rápido que sus compañeros que estaban menos expuestos.

Se realizó el análisis de los datos en forma multivariante, el modelo ajustado después de realizar la validación de supuestos del modelo en forma multivariante, usando el método clásico sin hacer distancias es un modelo logarítmico en Y obteniéndose así

$$ln\hat{Y}_{16\times 5} = V_{16\times 3}\hat{\boldsymbol{\beta}}_{3\times 5}$$

donde la matriz diseño V contiene un vector de unos asociado al intercepto y las variables independientes son género y exposición, donde $\hat{\boldsymbol{\beta}}$ es la matriz de parámetros estimados, reemplazando por los valores estimados obtenidos a través de mínimos cuadrados ponderados se obtiene

$$ln\hat{Y}_{16\times 5} = V_{16\times 3}\begin{pmatrix} 0.2598 & 0.2987 & 0.4472 & 0.4323 & 0.4807 \\ 0.0530 & 0.1005 & 0.0924 & 0.1956 & 0.2065 \\ 0.0763 & 0.0846 & 0.1053 & 0.2069 & 0.2241 \end{pmatrix}_{3\times 5}$$

Este modelo ajustado tiene un valor de AIC de -14.294 y un BIC de -4.766. Después de realizar el análisis se encuentra que la variable género no resulta ser significativa en el MANOVA, y al hacer el análisis en cada uno de los

cinco tiempos se observa que la variable exposición a la conducta desviada es altamente significativa en los tiempos 4 y 5.

El modelo ajustado después de realizar la validación de supuestos del modelo y usando distancias es un modelo logarítmico en Y y se presenta a continuación

$$ln\hat{Y}_{16\times 5} = \mathbf{1}_{16\times 1}\widehat{\boldsymbol{B}_{0(1\times 5)}} + \boldsymbol{X}_{16\times 2}\widehat{\boldsymbol{B}}_{2\times 5}$$

donde $\boldsymbol{X}_{16\times 2}$ corresponde a la matriz de coordenadas principales reducida, $\widehat{\boldsymbol{B}}_0$ contiene las estimaciones asociadas al intercepto y $\widehat{\boldsymbol{B}}$ las estimaciones asociadas a las dos primeras coordenadas principales. Al reemplazar por las estimaciones obtenidas por mínimos cuadrados ponderados del análisis MANOVA se obtiene el siguiente modelo ajustado

$$\begin{aligned} ln\hat{Y}_{16\times 5} =& \mathbf{1}_{16\times 1}\begin{pmatrix} 0.2829 & 0.3427 & 0.4876 & 0.5179 & 0.5710 \end{pmatrix} \\ &+\boldsymbol{X}_{16\times 2}\begin{pmatrix} -0.0497 & -0.1057 & -0.0565 & -0.0971 & -0.0943 \\ -0.3791 & -0.3455 & -0.4187 & -0.8119 & -0.9048 \end{pmatrix} \end{aligned}$$

El AIC y BIC para este modelo fueron -14.548 y -5.0204 respectivamente. La diferencia con respecto al MANOVA clásico es pequeña, arrojando valores casi similares, siendo el ajuste prácticamente igual con ambos métodos. Del análisis MANOVA se encuentra que la primera componente no tiene efecto significativo sobre la tolerancia desviada del comportamiento, recordemos que esta componente recoge la variabilidad de la variable género por lo cual se puede decir que el género no tiene efecto significativo en la tolerancia y al realizar el análisis en cada uno de los tiempos por medio de ANOVAS se encuentra que la segunda componente es altamente significativa en los tiempos 4 y 5. Además, el cuadrado medio del error resulta ser más bajo para el método DB siendo de 0.01167 y para el método clásico de 0.012, pero son prácticamente iguales.

La Figura 2.4 permite ver el ajuste del modelo con distancias y sin distancias en comparación con los valores observados de tolerancia (los círculos), donde se puede apreciar que las predicciones usando distancias mediante MANOVA (los triángulos), con respecto al método clásico mediante MANOVA (las cruces) son casi iguales excepto para algunos individuos y en los últimos tiempos observados, pero difieren en muy poco. Se puede ver que ambos métodos son similares usando los criterios de información AIC y BIC.

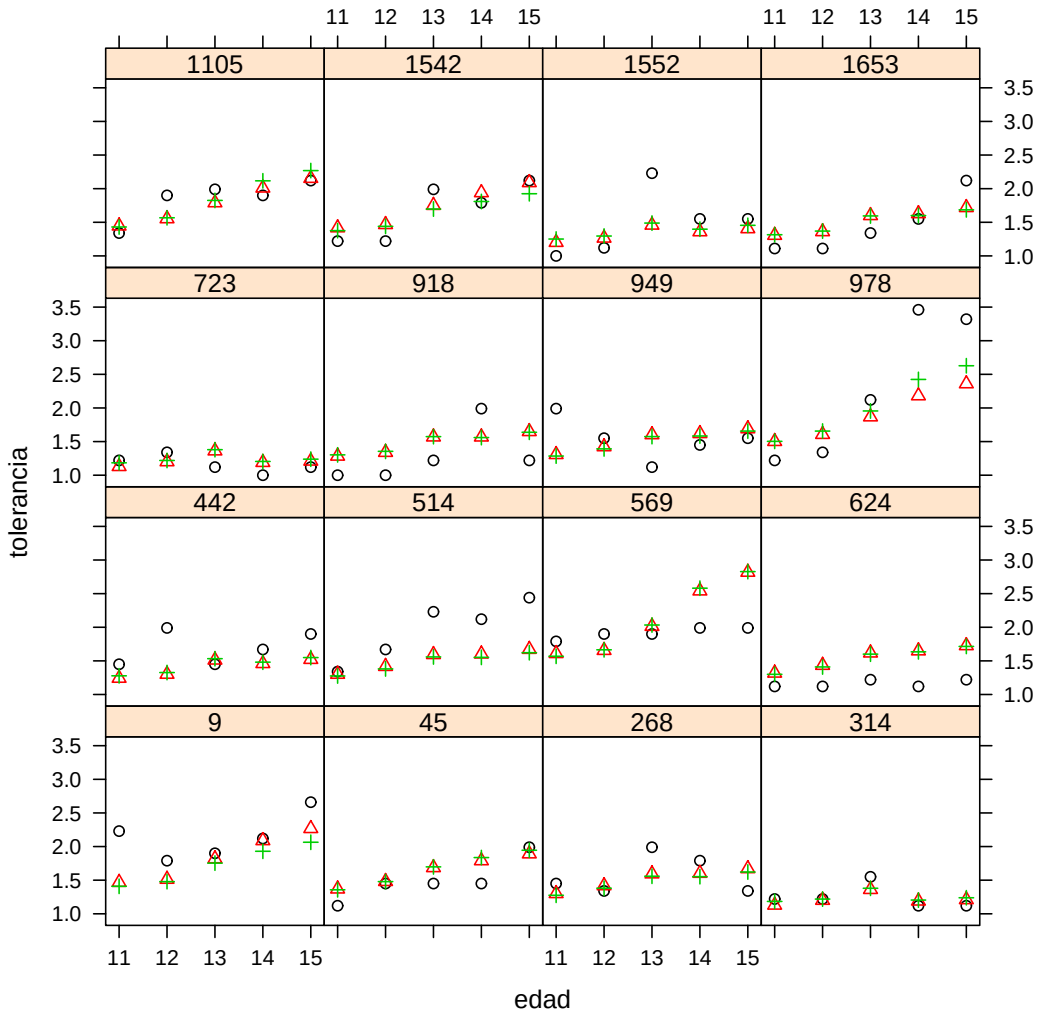


FIGURA 2.4: Tolerancia vs predicciones utilizando los métodos DB y clásico en función de la edad por individuo mediante MANOVA

Capítulo 3

Aproximación basada en distancias en análisis de datos longitudinales univariantes

En este capítulo se propone la extensión de los métodos de estimación basados en distancias a esta clase de problemas. Se estudian datos desbalanceados, donde el número de veces que cada individuo es medido puede ser diferente, y los tiempos también pueden ser desigualmente espaciados. Se encontraron algunas ventajas en el uso de los métodos basados en distancias, tales como: las componentes de la matriz de ACP son independientes, donde las variables originales usualmente no lo son. En estas circunstancias es donde los investigadores están principalmente interesados en hacer predicciones. La metodología propuesta es también útil ya que produce un mejor ajuste que con el modelo clásico cuando se agregan más componentes. También, al ser un análisis longitudinal permite al investigador hacer predicciones en cada tiempo, lo cual resulta útil para estimar datos faltantes.

Existen diferentes técnicas para el análisis de datos de medidas repetidas. Algunas de estas exploran la relación entre los datos en diferentes tiempos, teniendo en cuenta la estructura del tiempo de los datos, tales como análisis factor longitudinal, análisis factor dinámico, series de tiempo multivariante, modelos espacio-estado (Jørgensen et al. 1996) y análisis de curvas de crecimiento (Chaganty & Mav 2007). Métodos tales como el multivariante mixto o el doblemente multivariante (Boik 1991) asumen distribución normal y homogeneidad en la matriz de covarianza respecto a cierto factor de agrupación (usualmente es la variable tiempo), así como independencia entre las observaciones y las variable dependiente del tiempo.

Entre otros estudios usando distancias esta el trabajo de Peng & Müller (2008), quienes proponen una distancia entre dos realizaciones de un proceso

aleatorio, donde para cada realización solamente medidas espaciadas irregularmente y escasas con errores de medición adicional están disponibles. Tales datos ocurren comúnmente en estudios longitudinales. Una medida de distancias entonces hace posible aplicar análisis basado en distancias tales como clasificación, escalamiento multidimensional y clustering para datos longitudinales muestreados irregularmente. Además, Müller & Yao (2010) muestra que los procesos subyacentes precios de las ofertas en linea, subastas y muchos otros datos longitudinales pueden ser representados por una ecuación diferencial ordinaria estocástica de primer orden empírica con los coeficientes del tiempo y un proceso de desplazamiento suave.

Teniendo en cuenta los estudios realizados a la fecha usando distancias. En este capítulo se propone una metodología para el análisis de datos longitudinales a través de una aproximación univariante usando distancias entre pares de observaciones con respecto a las variables explicativas con la variable respuesta continua. Se encuentra que el uso de esta estrategia para modelar problemas de este tipo arroja resultados igualmente de robustos que la estrategia de modelamiento tradicional y funciona también en casos donde se tiene en las variables explicativas: datos categóricos, binarios, mixtos y continuos. Además, se prueba que las predicciones generadas son las mismas bajo el modelo propuesto y el clásico.

Este capítulo se desarrolla en cuatro secciones: en la Sección 3.1 se presentan algunos aspectos inferenciales para el modelo de datos longitudinales univariante basado en distancias. Se realiza la estimación de los parámetros por mínimos cuadrados generalizados y máxima verosimilitud restringida. En la Sección 3.2 se presentan los modelos paramétricos para la estructura de covarianza y cómo hacer la predicción de un nuevo individuo. En la Sección 3.3 se muestran los resultados de la simulación para la aproximación univariante, y finalmente, en la Sección 3.4 se presenta una aplicación de la metodología propuesta.

3.1 Modelos de covarianza, estimación de parámetros y aspectos inferenciales

Otra aproximación al problema de datos longitudinales esta basado en un enfoque univariante. Debido a la naturaleza de la información, cada individuo a través del tiempo se encuentra correlacionado, no conociéndose la estructura de la matriz de covarianza. Por esta razón, la estimación de los parámetros se hace usualmente por medio de mínimos cuadrados generalizados.

En primer lugar, se expresa el modelo (2.15) en forma univariante aplicando

el operador *vec*, obteniendo

$$\begin{aligned} vec(Y') =& vec(\boldsymbol{B}'\boldsymbol{X}') + vec(\Xi_k') \\ y =& (\boldsymbol{X} \otimes I_m)\theta + \epsilon \end{aligned} \tag{3.1}$$

donde $y = vec(Y')$, $\theta = vec(\boldsymbol{B}')$ y $\epsilon = vec(\Xi_k')$. Además, $Var(y) = Var(\epsilon) = I_n \otimes \Sigma = \sigma^2\Psi = \sigma^2 I_n \otimes \Psi_0$ con $\Psi = I_n \otimes \Psi_0$ y $\Sigma = \sigma^2\Psi_0$.

Antes de hacer las estimaciones de los valores de θ, σ^2 y Ψ_0, se considera cómo puede ser la forma de los bloques diferentes de cero, $\sigma^2\Psi_0$, en el modelo (3.1).

3.1.1 Patrones de covarianza

Se dispone de una gran selección de patrones de covarianza en el modelamiento de datos longitudinales y medidas repetidas. Muchos de estos patrones son dependientes de la medición tomada en tiempos fijos y algunos son fáciles de justificar cuando las observaciones son igualmente espaciadas. Hay también patrones donde las covarianzas se basan sobre el valor exacto del tiempo (por ejemplo, el cambio del número de visitas), y éstos son más útiles en situaciones donde los intervalos del tiempo son irregulares. A continuación se presentan algunos de los más utilizados.

Modelo de covarianza compuesta simétrica

En este modelo se asume que hay una correlación positiva, ρ, entre dos mediciones sobre el mismo sujeto. En términos matriciales, ésta corresponde a

$$\Psi_0 = (1-\rho)I + \rho J \tag{3.2}$$

donde I representa una matriz identidad y J es una matriz de unos, ambas de orden $m \times m$.

Una justificación del uso del modelo de *correlación uniforme* en el modelamiento clásico es la siguiente: considérese el modelo

$$y_{ij} = x_i'\theta + U_i + Z_{ij}, \quad i = 1, ..., n; \ j = 1, ..., m \tag{3.3}$$

con x_i' una fila de la matriz $(\boldsymbol{X} \otimes I_m)$, y donde $x_i'\theta = E(y_{ij})$, los U_i son variables aleatorias mutuamente independientes $N(0, \nu^2)$, los Z_{ij} son variables aleatorias mutuamente independientes $N(0, \tau^2)$ y, los U_i y Z_{ij} son independientes uno del otro. Entonces la estructura de covarianza de los datos corresponde a (3.2) con $\rho = \nu^2/(\nu^2 + \tau^2)$ y $\sigma^2 = \nu^2 + \tau^2$. Se puede observar que el modelo (3.3) da una interpretación de la correlación del modelo uniforme en el que un modelo

lineal para la respuesta media incorpora un término de intercepto aleatorio con varianza ν^2 entre sujetos. Además, el modelo (3.3) proporciona una justificación basada en modelos para una aproximación del análisis de varianza al análisis de datos longitudinales llamado análisis de parcelas divididas.

Modelo con correlación exponencial

En este modelo, Ψ_0 tiene el j-ésimo elemento, $\psi_{jk} = Cov(y_{ij}, y_{ik})$, de la forma

$$\psi_{jk} = \sigma^2 \exp\left(-\phi|t_j - t_k|\right) \tag{3.4}$$

En contraposición al modelo de correlación uniforme, la correlación entre un par de mediciones sobre la misma unidad decae hacia cero a medida que el tiempo de separación es mas lejano. La razón de decrecimiento es mas rápida para valores muy grandes de ϕ. Es de notar que si los tiempos de observación, t_j, son igualmente espaciados ($t_{j+1} - t_j = d$, para todo j) entonces (3.4) se puede expresar como

$$\psi_{jk} = \sigma^2 \rho^{|j-k|} \tag{3.5}$$

donde $\rho = \exp(-\phi d)$ es la correlación entre observaciones consecutivas sobre el mismo sujeto.

Una justificación de (2.18) es representar las variables aleatorias y_{ij} como

$$y_{ij} = x_i'\theta + \xi_{ij}, \quad i = 1, ..., n;\ j = 1, ..., m \tag{3.6}$$

donde

$$\xi_{ij} = \rho\xi_{i(j-1)} + Z_{ij} \tag{3.7}$$

y Z_{ij} son variables aleatorias $N\left(0, \sigma^2\left(1 - \rho^2\right)\right)$ mutuamente independientes, tal que $Var(y_{ij}) = Var(\xi_{ij}) = \sigma^2$. En vista de (3.6) y (3.7), el modelo de correlación exponencial es a veces llamado el modelo *autorregresivo de primer orden*, debido a que (3.7) es la definición estándar de un proceso autorregresivo discreto de primer orden (Diggle et al. 2002). Esto ha llevado a algunos autores a generalizar el modelo asumiendo las secuencias, $\xi_{ij},\ j = 1, \ldots, m$ en (3.6) como realizaciones parciales mutuamente independientes de un proceso general autorregresivo estacionario de media móvil,

$$\xi_{ij} = \sum_{r=1}^{p} \rho_r \xi_{i(j-r)} + Z_{ij} + \sum_{s=1}^{q} \alpha_s Z_{i(j-s)} \tag{3.8}$$

En el caso de datos longitudinales la expresión (3.8) de tiempo discreto es poco común, de hecho es más natural generalizar el modelo (3.6) escribiéndolo de la forma

$$y_{ij} = x_i'\theta + \xi_i(t_j), \quad i = 1, ..., n;\ j = 1, ..., m \tag{3.9}$$

donde las secuencias $\xi_i(t_j)$, $j = 1, \ldots, m$ son realizaciones mutuamente independientes, tiempo continuo, del proceso estacionario Gaussiano $\{\xi_i(t), t \in \Re\}$ con estructura de covarianza común, $\gamma(u) = Cov(\xi_i(t), \xi_i(t-u))$. Se debe tener en cuenta que (3.9) se adapta automáticamente a mediciones irregularmente espaciadas, t_j, y de paso, a diferentes momentos de medición de las unidades; mientras que en los modelos de tiempo discreto, (3.7) o (3.8), sería muy poco natural para datos irregularmente espaciados t_j.

Otros patrones de covarianza

Algunos patrones de covarianza para la matriz Ψ_0 en un ensayo con m puntos en el tiempo se presentan a continuación

- **No estructurada o patron general.** La varianza de la respuesta, σ_i^2, difiere para cada período de tiempo i, y las covarianzas, σ_{ij}, difieren entre cada par de períodos j y k.

$$\Psi_0 = \frac{1}{\sigma^2} \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$

  La anterior estructura produce los mismos resultados que ajustar un modelo con la distribución normal multivariante, es decir como el modelo propuesto en la Subsección 2.1.2. El problema es que este modelo necesita muchos parámetros de covarianzas.

- **Toeplitz.** Esta matriz usa covarianzas separadas para cada nivel de separación entre los puntos del tiempo. Este modelo es conocido como el modelo autorregresivo general.

$$\Psi_0 = \frac{1}{\sigma^2} \begin{pmatrix} \sigma^2 & \theta_1 & \theta_2 & \cdots & \theta_{m-1} \\ \theta_1 & \sigma^2 & \theta_1 & \cdots & \theta_{m-2} \\ \theta_2 & \theta_1 & \sigma^2 & \cdots & \theta_{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{m-1} & \theta_{m-2} & \theta_{m-3} & \cdots & \sigma^2 \end{pmatrix}$$

- **Heterogénea no estructurada.** Los puntos en el tiempo tienen diferentes varianzas, pero las observaciones sobre el mismo individuo están no correlacionadas. Ésta debería solamente ser utilizada si los análisis

preliminares con patrones con más parámetros indican la correlación entre las observaciones repetidas o longitudinales.

$$\Psi_0 = \frac{1}{\sigma^2}\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m^2 \end{pmatrix}$$

- **Heterogénea compuesta simétrica.** Tiene una forma similar al modelo compuesto simétrico incluyendo diferentes varianzas en el tiempo.

$$\Psi_0 = \frac{1}{\sigma^2}\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \cdots & \rho\sigma_1\sigma_m \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho\sigma_2\sigma_m \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma_1\sigma_m & \rho\sigma_2\sigma_m & \cdots & \sigma_m^2 \end{pmatrix}$$

- **Heterogénea autorregresiva de primer orden.** Tiene una forma similar al modelo autorregresivo de primer orden incluyendo diferentes varianzas en el tiempo.

$$\Psi_0 = \frac{1}{\sigma^2}\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \cdots & \rho^{m-1}\sigma_1\sigma_m \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho^{m-2}\sigma_2\sigma_m \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1}\sigma_1\sigma_m & \rho^{m-2}\sigma_2\sigma_m & \cdots & \sigma_m^2 \end{pmatrix}$$

- **Heterogénea Toeplitz.** Tiene una forma similar al modelo de Toeplitz excepto porque se consideran diferentes varianzas para cada punto del tiempo.

$$\Psi_0 = \frac{1}{\sigma^2}\begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \cdots & \rho_{m-1}\sigma_1\sigma_m \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{m-2}\sigma_2\sigma_m \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m-1}\sigma_1\sigma_m & \rho_{m-2}\sigma_2\sigma_m & \cdots & \sigma_m^2 \end{pmatrix}$$

3.1.2 Estimación de parámetros

Dado que un objetivo de este análisis es ajustar un modelo para los datos, es necesario determinar los valores de θ, σ^2 y Ψ_0. Por lo tanto, en la siguiente sección se presenta el cálculo de los estimadores para los parámetros y las pruebas de hipótesis sobre ellos.

Estimación por mínimos cuadrados ponderados

El estimador de mínimos cuadrados ponderados (MCP) de θ, utilizando una matriz simétrica de pesos, Υ, es $\hat{\theta}$ el cual minimiza

$$[y - (\boldsymbol{X} \otimes I)\,\theta]' \,\Upsilon\, [y - (\boldsymbol{X} \otimes I)\,\theta]$$

Las estimaciones de MCP de los parámetros θ verifican las ecuaciones normales (EN)

$$(\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)'\,\theta = (\boldsymbol{X} \otimes I)\,\Upsilon y \tag{3.10}$$

y vienen dadas por la expresión

$$\hat{\theta}_{\Upsilon} = \left[(\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,\Upsilon y \tag{3.11}$$

ya que el modelo es de rango máximo $k^* = rang\,(\boldsymbol{X} \otimes I)$.

El estimador presentado en (3.11) es insesgado ya que

$$\begin{aligned} E\left(\hat{\theta}_{\Upsilon}\right) &= \left[(\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,\Upsilon E(y) \\ &= \left[(\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)\,\theta = \theta \end{aligned}$$

Además, ya que $Var(y) = \sigma^2\Psi$, entonces

$$\begin{aligned} Var\left(\hat{\theta}_{\Upsilon}\right) =& \sigma^2 \left[(\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,\Upsilon\Psi\Upsilon \\ & (\boldsymbol{X} \otimes I) \left[(\boldsymbol{X} \otimes I)' \,\Upsilon\, (\boldsymbol{X} \otimes I)\right]^{-1} \end{aligned} \tag{3.12}$$

Si $\Upsilon = I$, la matriz identidad, (3.11) se reduce a el estimador de mínimos cuadrados ordinarios (MCO)

$$\hat{\theta}_I = \left[(\boldsymbol{X} \otimes I)' \,(\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,y \tag{3.13}$$

con

$$\begin{aligned} Var\left(\hat{\theta}_I\right) =& \sigma^2 \left[(\boldsymbol{X} \otimes I)' \,(\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,\Psi\, (\boldsymbol{X} \otimes I) \\ & \left[(\boldsymbol{X} \otimes I)' \,(\boldsymbol{X} \otimes I)\right]^{-1} \end{aligned} \tag{3.14}$$

Si $\Upsilon = \Psi^{-1}$, el estimador toma la forma

$$\hat{\theta} = \left[(\boldsymbol{X} \otimes I)' \,\Psi^{-1}\, (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \,\Psi^{-1} y \tag{3.15}$$

con

$$Var\left(\hat{\theta}\right) = \sigma^2 \left[(\boldsymbol{X} \otimes I)' \,\Psi^{-1}\, (\boldsymbol{X} \otimes I)\right]^{-1} \tag{3.16}$$

La notación sombrero anticipa que $\hat{\theta}$ es el estimador máximo verosimilitud de θ bajo la condición Gaussiana multivariante $y \sim NM((\boldsymbol{X} \otimes I)\,\theta, \sigma^2\Psi)$. Esto último sugiere que el estimador de mínimos cuadrados ponderados mas eficiente para θ utiliza $\Upsilon = \Psi^{-1}$. Sin embargo, para identificar esta matriz ponderada óptima se necesita saber la estructura de correlación completa de los datos, no se necesita saber σ^2, porque $\hat{\theta}_\Upsilon$ no se altera por cambios proporcionales en todos los elementos de Υ. Debido a que la estructura de correlación puede ser difícil de identificar en la práctica, es de interés preguntarse cuánta eficiencia se pierde al utilizar un Υ diferente. Es de notar que la eficiencia relativa de $\hat{\theta}_\Upsilon$ y $\hat{\theta}$ se pueden calcular de sus respectivas matrices de varianza (3.12) y (3.16).

La eficiencia relativa de MCO depende de la interacción precisa entre las matrices $(\boldsymbol{X} \otimes I)$ y Ψ, como lo describió Bloomfield & Watson (1975). Sin embargo, bajo las condiciones encontradas en un amplio rango de aplicaciones de datos longitudinales, la eficiencia relativa es con frecuencia bastante buena (Diggle et al. 2002). Incluso cuando MCO es razonablemente eficiente, es claro de la forma de $Var\left(\hat{\theta}\right)$ dada en (3.16) que la estimación por intervalo para θ requiere información sobre $\sigma^2\Psi$, la matriz de varianza de los datos. En particular, la formula usual para la varianza del estimador de mínimos cuadrados es

$$Var\left(\hat{\theta}\right) = \sigma^2 \left[(\boldsymbol{X} \otimes I)' (\boldsymbol{X} \otimes I)\right]^{-1} \tag{3.17}$$

asumiendo que $\Psi = I$, la matriz identidad, lo cual puede ser seriamente engañoso cuando esto no es así.

Un uso ingenuo de los MCO sería ignorar la estructura de correlación de los datos y basar la estimación de θ en la fórmula de varianza (3.17) con σ^2 reemplazado por su estimador usual, el cuadrado medio residual

$$\hat{\sigma}^2 = \frac{1}{nm - k^*} \left[y - (\boldsymbol{X} \otimes I)\,\hat{\theta}\right]' \left[y - (\boldsymbol{X} \otimes I)\,\hat{\theta}\right] \tag{3.18}$$

Hay dos fuentes de error en esta ingenua aproximación cuando $\Psi = I$. Primero, la fórmula (3.17) es incorrecta para $Var\left(\hat{\theta}\right)$ y segundo, $\hat{\sigma}^2$ no es un estimador insesgado para σ^2. Para evaluar el efecto combinado de estas dos fuentes de error, se puede comparar los elementos de la diagonal de $Var\left(\hat{\theta}\right)$ como se presento en (3.14) con los correspondientes elementos de la diagonal de la matriz $E\left(\hat{\sigma}^2\right)\left[(\boldsymbol{X} \otimes I)' (\boldsymbol{X} \otimes I)\right]^{-1}$. Algunos ejemplos numéricos se presentan en el Capítulo 1 de Diggle et al. (2002), donde la conclusión es que en presencia de autocorrelación positiva, el uso ingenuo de MCO puede seriamente sobre o sub estimar la varianza de $\hat{\theta}$ dependiendo de la matriz diseño.

Estimación máxima verosimilitud bajo condiciones Gaussianas

Una estrategia para la estimación de los parámetros en el modelo lineal general (MLG) consiste en considerar la estimación simultánea de los parámetros de interés, θ, y de los parámetros de covarianza, σ^2 y Ψ_0, usando la función de verosimilitud. Recuérdese que Ψ es una matriz de bloques diagonal con bloques comunes diferentes de cero Ψ_0. Bajo la condición Gaussiana $y \sim NM\left((\boldsymbol{X} \otimes I)\,\theta, \sigma^2\Psi\right)$, la log-verosimilitud para los datos observados y es

$$\begin{aligned} L\left(\theta, \sigma^2, \Psi_0\right) = &-\frac{nm}{2}\log(\sigma^2) - \frac{n}{2}\log(|\Psi_0|) \\ &-\frac{1}{2\sigma^2}\left[y - (\boldsymbol{X} \otimes I)\,\theta\right]' \Psi^{-1}\left[y - (\boldsymbol{X} \otimes I)\,\theta\right] \end{aligned} \tag{3.19}$$

Para Ψ_0 dado, el estimador de máxima verosimilitud para θ es el estimador de mínimos cuadrados ponderado dado por

$$\hat{\theta}(\Psi_0) = \left[(\boldsymbol{X} \otimes I)' \Psi^{-1} (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \Psi^{-1} y \tag{3.20}$$

Sustituyendo en (3.19), se obtiene

$$L\left(\hat{\theta}(\Psi_0), \sigma^2, \Psi_0\right) = -\frac{1}{2}\left\{nm\log\sigma^2 + n\log(|\Psi_0|) + \sigma^{-2} SCR(\Psi_0)\right\} \tag{3.21}$$

donde

$$SCR(\Psi_0) = \left[y - (\boldsymbol{X} \otimes I)\,\hat{\theta}(\Psi_0)\right]' \Psi^{-1}\left[y - (\boldsymbol{X} \otimes I)\,\hat{\theta}(\Psi_0)\right]$$

Ahora diferenciando (3.21) con respecto a σ^2, manteniendo Ψ_0 fijo, se obtiene el estimador de máxima verosimilitud para σ^2,

$$\hat{\sigma}^2(\Psi_0) = \frac{SCR(\Psi_0)}{nm} \tag{3.22}$$

De esta manera, al sustituir (3.20) y (3.22) en (3.19) se obtiene la log-verosimilitud reducida para Ψ_0, omitiendo el termino constante,

$$L_r(\Psi_0) = L\left(\hat{\theta}(\Psi_0), \hat{\sigma}^2(\Psi_0), \Psi_0\right) = -\frac{1}{2} n\left[m\log SCR(\Psi_0) + \log(|\Psi_0|)\right] \tag{3.23}$$

Finalmente, la maximización de $L_r(\Psi_0)$ produce $\widehat{\Psi}_0$ y, por sustitución en (3.20) y (3.22), los estimadores de máxima verosimilitud, $\hat{\theta} \equiv \hat{\theta}\left(\widehat{\Psi}_0\right)$ y $\hat{\sigma}^2 \equiv \hat{\sigma}^2\left(\widehat{\Psi}_0\right)$.

En general, la maximización de (3.23) con respecto a los distintos elementos de Ψ_0 requiere de técnicas de optimización numérica. La dimensionalidad del

problema de optimización para Ψ_0 es $\frac{1}{2}m(m-1)$ si se asume una varianza común σ^2 en cada uno de los m puntos del tiempo, y $\frac{1}{2}m(m+1)-1$ en otro caso. También, el mayor trabajo computacional en evaluar $L(\Psi_0)$, consiste en calcular el determinante e inversa de una matriz simétrica definida positiva, $m \times m$.

Nótese que usando máxima verosimilitud para la estimación simultánea de θ, σ^2 y Ψ_0, la forma de la matriz de coordenadas principales $(\boldsymbol{X} \otimes I)$ esta involucrada explícitamente en la estimación de σ^2 y Ψ_0. Esta estimación necesita la utilización de una matriz de coordenadas con un gran número de columnas para obtener estimaciones consistentes de la estructura de covarianzas, mientras que una estimación aproximadamente insesgada requiere un número pequeño de columnas. Si se esta seguro de que un número pequeño de columnas definirá un modelo adecuado, el problema no es grave. De lo contrario, se debe tener en cuenta otros métodos de estimación. Uno de ellos es el método de máxima verosimilitud restringida (REML).

Estimación máxima verosimilitud restringida

El método de estimación REML, fue introducido por Patterson & Thompson (1971) como un camino para estimar componentes de varianza en un MLG. EL problema con el procedimiento de máxima verosimilitud estándar es que éste produce estimadores sesgados de los parámetros de covarianza. Es bien conocido que en el MLG con errores independientes se asume

$$y \sim NM\left((\boldsymbol{X} \otimes I_m)\theta, \sigma^2 I_{nm}\right) \tag{3.24}$$

En este caso, el estimador máximo verosimilitud para σ^2 es $\hat{\sigma}^2 = SCR/(nm)$, donde SCR representa la suma de cuadrados residual, el estimador insesgado usual es $\hat{\sigma}^2 = SCR/(nm-k^*)$, con k^* el número de columnas de $(\boldsymbol{X} \otimes I)$. De hecho, $\hat{\sigma}^2$ es el estimador REML para σ^2 en el modelo (3.24).

En el caso de el MLG con errores dependientes,

$$y \sim NM((\boldsymbol{X} \otimes I_m)\theta, \sigma^2\Psi) \tag{3.25}$$

El estimador REML es definido como un estimador máximo verosimilitud basado en un conjunto de datos transformado linealmente $Y^* = A_1 Y$ tal que la distribución de Y^* no depende de θ. Un camino para alcanzar esto, es tomando A_1 como la matriz que convierte a Y en residuales MCO,

$$A_1 = I - (\boldsymbol{X} \otimes I)\left[(\boldsymbol{X} \otimes I)'(\boldsymbol{X} \otimes I)\right]^{-1}(\boldsymbol{X} \otimes I)' \tag{3.26}$$

Entonces Y^* tiene una distribución Gaussiana multivariante con media cero, independiente del valor de θ. Para obtener una matriz no singular, se

puede usar solamente $nm - k^*$ filas de la matriz, A_1, definida en (3.26). Lo anterior hace que los estimadores resultantes para σ^2 y Ψ no dependan de las filas que se utilicen, ni tampoco de la elección particular de A_1: alguna matriz de rango completo con la propiedad que $E(Y^*) = 0$ para todo θ dará la misma respuesta. Por otra parte, desde un punto de vista operativo, no es necesario realizar la transformación de Y a Y^* explícita. Los detalles algebraicos son los siguientes.

Para el desarrollo teórico es conveniente que no fuera el factor σ^2 dentro de Ψ, y escribir el modelo para el vector respuesta y como

$$y \sim NM((\boldsymbol{X} \otimes I_m)\theta, \Psi_1),$$

donde $\Psi_1 \equiv \Psi_1(\alpha)$ y la matriz de varianza de y depende de un vector de parámetros α. Sea A_1 la matriz definida en (3.26) y B_1 la matriz $nm \times (nm - k^*)$ definida tal que $B_1 B_1' = A_1$ y $B_1' B_1 = I$, donde I denota la matriz identidad $(nm - k^*) \times (nm - k^*)$. Finalmente, sea $Z = B_1' y$.

Ahora, para un α fijo el estimador de máxima verosimilitud para θ es el estimador de mínimos cuadrados generalizados (MCG),

$$\hat{\theta} = \left[(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)' \Psi_1^{-1} y = G_1 y$$

También, las respectivas funciones de densidad de probabilidad (fdp) de y y θ son

$$f(y) = (2\pi)^{-\frac{1}{2}nm} |\Psi_1|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left[y - (\boldsymbol{X} \otimes I)\theta\right]' \Psi_1^{-1} \left[y - (\boldsymbol{X} \otimes I)\theta\right]\right\}$$

y

$$g(\hat{\theta}) = \frac{\exp\left\{-\frac{1}{2}(\hat{\theta} - \theta)' \left[(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right] (\hat{\theta} - \theta)\right\}}{(2\pi)^{\frac{1}{2}k^*} \left|(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right|^{\frac{1}{2}}}$$

Además, $E(Z) = 0$ y, Z y $\hat{\theta}$ son independientes cualquiera que sean los valores de θ como se muestra a continuación.

Primero,

$$E(Z) = B_1' E(y) = B_1' (\boldsymbol{X} \otimes I)\theta = B_1' B_1 B_1' (\boldsymbol{X} \otimes I)\theta$$

ya que $B_1' B_1 = I$. Pero como $B_1 B_1' = A_1$, entonces

$$E(Z) = B_1' A_1 (\boldsymbol{X} \otimes I)\theta$$

y

$$\begin{aligned} A_1(\boldsymbol{X} \otimes I) =& \left\{I - (\boldsymbol{X} \otimes I)\left[(\boldsymbol{X} \otimes I)'(\boldsymbol{X} \otimes I)\right]^{-1} (\boldsymbol{X} \otimes I)'\right\} (\boldsymbol{X} \otimes I) \\ =& (\boldsymbol{X} \otimes I) - (\boldsymbol{X} \otimes I) = 0 \end{aligned}$$

Por consiguiente, $E(Z) = 0$ como se requería.

En segundo lugar,

$$\begin{aligned} Cov(Z, \hat{\theta}) =& E\left\{Z\left(\hat{\theta} - \theta\right)'\right\} \\ =& E\left\{B_1' y \left(y' G_1' - \theta'\right)\right\} \\ =& B_1' E(yy') G_1' - B_1' E(y)\theta' \\ =& B_1' \{Var(y) + E(y)E(y)'\} G_1' - B_1' E(y)\theta' \end{aligned}$$

Ahora, sustituyendo $E(y) = (\boldsymbol{X} \otimes I)\theta$ dentro de esta última expresión se obtiene

$$Cov(Z, \hat{\theta}) = B_1' \left[\Psi_1 + (\boldsymbol{X} \otimes I)\theta\theta'(\boldsymbol{X} \otimes I)'\right] G_1' - B_1'(\boldsymbol{X} \otimes I)\theta\theta'$$

También,

$$(\boldsymbol{X} \otimes I)' G_1' = (\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I) \left[(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right]^{-1} = I$$

y

$$\begin{aligned} B_1' \Psi_1 G_1' =& B_1' \Psi_1 \Psi_1^{-1} (\boldsymbol{X} \otimes I) \left[(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right]^{-1} \\ =& B_1' (\boldsymbol{X} \otimes I) \left[(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right]^{-1} = 0 \end{aligned}$$

ya que $B_1'(\boldsymbol{X} \otimes I) = B_1' A_1 (\boldsymbol{X} \otimes I) = 0$ como en la prueba de $E(Z) = 0$. Se sigue que $Cov(Z, \hat{\theta}) = B_1'(\boldsymbol{X} \otimes I)\theta\theta' - B_1'(\boldsymbol{X} \otimes I)\theta\theta' = 0$. Finalmente, en el ajuste de la distribución Gaussiana multivariante, la covarianza cero es equivalente a independencia. Por lo tanto, la forma algebraica de la *fdp* de Z Gaussiana multivariante (singular) en términos de y es proporcional a la razón $f(y)/g(\hat{\beta})$. De esta forma, se obtiene la forma explícita de esta razón utilizando el siguiente resultado estándar para el MLG.

$$\begin{aligned} [y - (\boldsymbol{X} \otimes I)\theta]' \Psi_1^{-1} [y - (\boldsymbol{X} \otimes I)\theta] = & \left[y - (\boldsymbol{X} \otimes I)\hat{\theta}\right]' \Psi_1^{-1} \left[y - (\boldsymbol{X} \otimes I)\hat{\theta}\right] \\ & + \left(\hat{\theta} - \theta\right)' \left[(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right] \left(\hat{\theta} - \theta\right) \end{aligned}$$

Entonces, la *fdp* de $Z = B_1' y$ es proporcional a

$$\frac{f(y)}{g(\hat{\theta})} = \frac{\exp\left\{-\frac{1}{2}\left[y - (\boldsymbol{X} \otimes I)\hat{\theta}\right]' \Psi_1^{-1} \left[y - (\boldsymbol{X} \otimes I)\hat{\theta}\right]\right\}}{(2\pi)^{\frac{1}{2}(nm - k^*)} |\Psi_1|^{\frac{1}{2}} \left|(\boldsymbol{X} \otimes I)' \Psi_1^{-1} (\boldsymbol{X} \otimes I)\right|^{\frac{1}{2}}} \tag{3.27}$$

donde la constante omitida de proporcionalidad es el Jacobiano de la transformación de y a $(Z, \hat{\theta})$. Harville (1974) muestra que el Jacobiano se reduce

a $|(\boldsymbol{X}\otimes I)'(\boldsymbol{X}\otimes I)|^{-\frac{1}{2}}$, el cual no depende de ninguno de los parámetros en el modelo y puede además ser ignorado para hacer inferencias sobre α o θ. Además, es de notar que el lado derecho de (3.27) es independiente de A_1, y el mismo resultado se mantiene para cualquier Z tal que $E(Z)=0$ y $Cov(Z,\hat{\theta})=0$.

La implicación practica de (3.27) es que el estimador REML, $\hat{\alpha}$, maximiza la log-verosimilitud

$$\begin{aligned}L(\alpha) = &-\frac{1}{2}\log|\Psi_1(\alpha)| - \frac{1}{2}\log\left|(\boldsymbol{X}\otimes I)'\Psi_1(\alpha)^{-1}(\boldsymbol{X}\otimes I)\right| \\ &-\frac{1}{2}\left[y-(\boldsymbol{X}\otimes I)\hat{\theta}(\alpha)\right]'\Psi_1(\alpha)^{-1}\left[y-(\boldsymbol{X}\otimes I)\hat{\theta}(\alpha)\right]\end{aligned}$$

manteniendo en mente que se consideraron n unidades con m medidas por unidad y que $\Psi_1(\alpha)$ es una matriz de bloques diagonal cuyos bloques no cero es la matriz $m\times m$-dimensional $\sigma^2\Psi_0(\alpha)$ que representa la matriz de covarianza entre las medidas observadas. También, para un $\Psi_0(\alpha)$ dado se tiene que

$$\hat{\theta}(\alpha) = \left\{(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\alpha)]^{-1}(\boldsymbol{X}\otimes I)\right\}^{-1}(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\alpha)]^{-1}y \quad (3.28)$$

y

$$\hat{\sigma}^2(\alpha) = SCR(\alpha)/(nm-k^*) \quad (3.29)$$

donde

$$SCR(\alpha) = \left[y-(\boldsymbol{X}\otimes I)\hat{\theta}(\alpha)\right]'[I\otimes\Psi_0(\alpha)]^{-1}\left[y-(\boldsymbol{X}\otimes I)\hat{\theta}(\alpha)\right] \quad (3.30)$$

y k^* es el número de elementos en θ. El estimador REML para α maximiza la log verosimilitud reducida

$$\begin{aligned}L(\alpha) = &-\frac{1}{2}n\left[m\log SCR(\alpha) + \log|\Psi_0(\alpha)|\right] \\ &-\frac{1}{2}\log\left|(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\alpha)]^{-1}(\boldsymbol{X}\otimes I)\right| \quad (3.31)\end{aligned}$$

Finalmente, reemplazando el estimador resultante $\hat{\alpha}$ en (3.28) y (3.29), los estimadores REML $\hat{\theta}=\hat{\theta}(\hat{\alpha})$ and $\hat{\sigma}^2=\hat{\sigma}^2(\hat{\alpha})$ se obtienen.

La naturaleza de los cálculos involucrados en la maximización de $L(\alpha)$ es clara si explícitamente la estructura del bloque de la diagonal $\Psi_1(\alpha)$ es reconocido. Tal que, asumiendo $y_i \sim NM((\boldsymbol{x}_i'\otimes I_{t_{m_i}})\theta, \Psi_i(t_i,\alpha))$, donde $\Psi_i(t_i,\alpha)$ es el bloque no cero $m_i\times m_i$ de $\Psi_1(\alpha)$ correspondiente a la matriz de covarianzas de y_i, las expresiones (3.30) y (3.31) pueden ser escritas como

$$SCR(\alpha) = \sum_{i=1}^{n}\left[y_i-(\boldsymbol{x}_i'\otimes I_{t_{m_i}})\hat{\theta}(\alpha)\right]'\Psi_i(t_i,\alpha)^{-1}\left[y_i-(\boldsymbol{x}_i'\otimes I_{t_{m_i}})\hat{\theta}(\alpha)\right]$$

y

$$L(\alpha) = -\frac{1}{2}\left[N\log SCR(\alpha) + \sum_{i=1}^{n}\log|\Psi_i(t_i,\alpha)|\right] - \frac{1}{2}\sum_{i=1}^{n}\log\left|(\boldsymbol{x}_i'\otimes I_{t_{m_i}})'\Psi_i(t_i,\alpha)^{-1}(\boldsymbol{x}_i'\otimes I_{t_{m_i}})\right| \tag{3.32}$$

En resumen, los estimadores de máxima verosimilitud y REML a menudo darán resultados muy similares. Sin embargo, cuando estos difieren sustancialmente, los estimadores REML pueden ser menos sesgados. En adelante, se usa la notación sombrero para referirse a los estimadores máxima verosimilitud o REML, excepto cuando el contexto no deja en claro que esta previsto.

3.1.3 Inferencia

La inferencia sobre θ puede estar basada en los resultados (3.28) los cuales, en conjunción con (3.25) implican que

$$\hat{\theta}(\alpha) \sim NM\{\theta, \sigma^2[(\boldsymbol{X}\otimes I)'(I\otimes\Psi_0(\alpha))^{-1}(\boldsymbol{X}\otimes I)]^{-1}\} \tag{3.33}$$

Asumiendo que (3.33) sigue siendo valida, como una buena aproximación, si los estimadores REML de $\hat{\sigma}^2$ y $\hat{\alpha}$ son sustituidos por los valores desconocidos de σ^2 y α en (3.33). Esto da

$$\hat{\theta} \sim NM\{\theta, \hat{\sigma}^2[(\boldsymbol{X}\otimes I)'(I\otimes\Psi_0(\hat{\alpha}))^{-1}(\boldsymbol{X}\otimes I)]^{-1}\} \tag{3.34}$$

La aplicación inmediata de (3.34) es un conjunto de errores estándar sobre los elementos individuales de θ. Casi tan inmediato es el cálculo de regiones de confianza para transformaciones lineales generales de la forma $C\theta$, donde C es de rango completo, $r\times k^*$ es una matriz con $r\leq k^*$. Las regiones de confianza para $C\theta$ se siguen del resultado que si $C\hat{\theta}$, entonces

$$C\hat{\theta} \sim NM(C\theta, \hat{\sigma}^2C\{(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\hat{\alpha})]^{-1}(\boldsymbol{X}\otimes I)\}^{-1}C')$$

del que se desprende a su vez

$$(C\hat{\theta}-C\theta)'\{\hat{\sigma}^2C\{(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\hat{\alpha})]^{-1}(\boldsymbol{X}\otimes I)\}^{-1}C'\}^{-1}(C\hat{\theta}-C\theta)$$

se distribuye como χ^2_r. Sea $\vartheta_r(q)$ denota el q-valor critico de χ^2_r, tal que $P(\chi^2_r\geq\vartheta_r(q))=q$. Entonces, una región de confianza del $100(1-q)\%$ para $C\theta$ es

$$(C\hat{\theta}-C\theta)'\{\hat{\sigma}^2C\{(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\hat{\alpha})]^{-1}(\boldsymbol{X}\otimes I)\}^{-1}C'\}^{-1}(C\hat{\theta}-C\theta)\leq\vartheta_r(q)$$

Una prueba de la hipótesis de un valor para $C\theta$, es decir $H_0: C\theta = C\theta_0$, consiste en rechazar H_0 en el nivel $100q\%$ si

$$(C\hat{\theta}-C\theta_0)'\{\hat{\sigma}^2C\{(\boldsymbol{X}\otimes I)'[I\otimes\Psi_0(\hat{\alpha})]^{-1}(\boldsymbol{X}\otimes I)\}^{-1}C'\}^{-1}(C\hat{\theta}-C\theta_0)>\vartheta_r(q)$$

3.2 Modelos paramétricos para la estructura de covarianza

En esta sección, se sigue trabajando con el modelo lineal general (3.1), pero se supone que la estructura de covarianza de la secuencia de medidas sobre cada unidad experimental asume ciertos valores de unos pocos parámetros desconocidos. Algunos ejemplos son: el modelo de correlación compuesta simétrica presentado en (3.2) y el modelo de correlación exponencial (3.4), en cada uno de éstos se utilizan dos parámetros para definir la estructura de covarianza.

Un enfoque del modelamiento paramétrico es particularmente útil para datos en los cuales las medidas sobre diferentes unidades no se hacen en un conjunto común de tiempos. Por esta razón, a continuación se utiliza una notación ligeramente mas general. Sea $\boldsymbol{y}_i = vec(Y_i') = (y_{i1}, \ldots, y_{im_i})$ el vector de m_i medidas sobre la i ésima unidad y $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{im_i})$ el correspondiente conjunto de tiempos en los cuales estas medidas son realizadas. Si hay n unidades en conjunto, se escribe $y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $t = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$ y $N = \sum_{i=1}^{n} m_i$.

Se asume que los $\boldsymbol{y}_i$ son realizaciones Gaussianas mutuamente independientes de vectores aleatorios y_i, con

$$y_i \sim NM((\boldsymbol{x}_i' \otimes I_{t_{m_i}})\theta, \Psi_i(t_i, \alpha)). \tag{3.35}$$

donde $\boldsymbol{x}_i'$ es un vector de componentes principales de orden $1 \times (k+1)$, $I_{t_{m_i}}$ es una matriz identidad de dimensión $m_i \times m_i$ y $\Psi_i(t_i, \alpha)$ es la matriz de covarianzas $m_i \times m_i$ de y_i. Los parámetros desconocidos son θ de dimensión $k+1$ y α de dimensión q. Nótese que la media y estructura de covarianzas son parametrizadas separadamente. Así mismo, cuando sea conveniente se escribirá el modelo para todo el conjunto de datos y como

$$y \sim NM((\boldsymbol{X} \otimes I_m), \Psi(t, \alpha)). \tag{3.36}$$

donde la matriz $\boldsymbol{X}$ de orden $N \times (k+1)$ se obtiene agrupando las matrices $\boldsymbol{x}_i'$ y $\Psi(\cdot)$, esta es una matriz diagonal por bloques de $N \times N$, con bloques diferentes de cero $\Psi_i(\cdot)$.

La notación resalta que el marco natural para la mayoría de los datos longitudinales es el tiempo que es continuo. Debido a esto, se puede obtener modelos específicos asumiendo que las secuencias y_{ij} $(j = 1, \ldots, m_i)$ son muestreadas de replicas independientes de un proceso estocástico subyacente con el tiempo continuo, $\{y(t), t \in \Re\}$. Por consiguiente, $y_{ij} = y_i(t_{ij}), \quad j = 1, \ldots, m_i; \quad i = 1, \ldots, n$. En la siguiente subsección se presentan algunos ejemplos de modelos que caen dentro del marco general de trabajo de (3.35), y se muestran cómo

diferentes clases de comportamientos estacionarios y no estacionarios se dan naturalmente.

La principal herramienta que se puede usar para describir las propiedades de cada modelo en procesos estocásticos $\{y(t)\}$ es la función de covarianza y la variograma. El variograma de un proceso estocástico $\{y(t)\}$, esta dado por la función

$$\gamma(u) = \frac{1}{2}E\left[\{y(t) - y(t-u)\}^2\right], \quad u \geq 0$$

Para un proceso estacionario $y(t)$, si $\rho(u)$ denota la correlación entre $y(t)$ y $y(t-u)$, y $\sigma^2 = Var\{y(t)\}$, entonces

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}$$

3.2.1 Modelos de covarianza

Con el fin de desarrollar un conjunto útil de los modelos que se necesitan comprender, al menos, cualitativamente, cuáles son las fuentes probables de variación aleatoria en datos longitudinales, en los modelos se desea incluir por lo menos tres diferentes fuentes cualitativas de variación aleatoria.

1. *Efectos aleatorios*: se da cuando las unidades son muestreadas en forma aleatoria de una población, varios aspectos de su comportamiento pueden mostrar variación estocástica entre sus unidades. Quizás el ejemplo mas simple es cuando el nivel general del perfil respuesta varia entre unidades, esto es, algunas unidades son intrínsecamente de alta respuesta, otras de baja respuesta.

2. *Correlación serial*: Al menos parte del perfil de cualquier unidad de medición observada puede ser una respuesta de un proceso estocástico que varia en el tiempo de operación dentro de una unidad. Este tipo de variación estocástica resulta de una correlación entre parejas medidas sobre la misma unidad, la cual depende de la separación en el tiempo entre las parejas medidas. Típicamente, la correlación se vuelve más débil cuando la separación en el tiempo aumenta.

3. *Error de medición*: Se da especialmente cuando las medidas individuales involucran alguna clase de muestreo dentro de las unidades, el proceso de medición puede agregar por si mismo una componente de variación a los datos.

Hay diferentes formas en las cuales estas características cualitativas pueden ser incorporadas dentro de un modelo específico. La siguiente formulación aditiva es manejable y útil.

Primero, se hace explícita la separación entre media y estructuras de covarianza (3.36) de la siguiente forma

$$y = (\boldsymbol{X} \otimes I_m)\theta + \epsilon \tag{3.37}$$

se sigue que

$$\epsilon \sim NM\{0, \Psi(t, \alpha)\} \tag{3.38}$$

Ahora, denotando como ϵ_{ij} a la j-ésima medida sobre la i-ésima unidad de ϵ, se asume una descomposición aditiva de ϵ_{ij} dentro de los efectos aleatorios, variación de correlación serial y error de medición, lo cual formalmente se expresa como

$$\epsilon_{ij} = \mathtt{d}'_{ij}\mathtt{U}_i + \mathtt{W}_i(t_{ij}) + \mathtt{Z}_{ij} \tag{3.39}$$

donde los $\mathtt{Z}_{ij}$ son un conjunto de N variables aleatorias Gaussianas mutuamente independientes, cada una con media cero y varianza τ^2. Los $\mathtt{U}_i$ son un conjunto de n vectores aleatorios Gaussianos mutuamente independientes con r elementos, cada uno con media cero y matriz de covarianza $\mathtt{G}$. Los $\mathtt{d}_{ij}$ son vectores de r elementos de variables explicativas adjuntas a la medición individual. Los $\mathtt{W}_i(t_{ij})$ son muestras de n replicas independientes de un proceso Gaussiano estacionario con media cero, varianza σ^2 y función de correlación $\rho(u)$. Además los $\mathtt{U}_i$, los $\{\mathtt{W}_i(t_{ij})\}$ y los $\mathtt{Z}_{ij}$ corresponden a efectos aleatorios, correlación serial y error de medición, respectivamente.

En las aplicaciones, la estructura aditiva que se asume es mas razonable después de una transformación de los datos. Por ejemplo, una transformación logaritmo debería convertir una estructura multiplicativa subyacente en una aditiva.

Escribiendo el vector de variables aleatorias como $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})$, donde ϵ_{ij} esta asociado a la i-ésima unidad. Sea $\mathtt{D}_i$ la matriz de $m_i \times r$ con la j-ésima fila $\mathtt{d}_{ij}$. Sea $\mathtt{H}_i$ una matriz de $m_i \times m_i$ con el (j,k)-ésimo elemento $\mathtt{h}_{ijk} = \rho(|t_{ij} - t_{ik}|)$, es decir $\mathtt{h}_{ijk}$ es la correlación entre $\mathtt{W}_i(t_{ij})$ y $\mathtt{W}_i(t_{ik})$. Finalmente, sea I_i la matriz identidad de orden $m_i \times m_i$. Entonces la matriz de covarianza de ϵ_i es

$$Var(\epsilon_i) = \mathtt{D}_i\mathtt{G}\mathtt{D}'_i + \sigma^2\mathtt{H}_i + \tau^2 I_i \tag{3.40}$$

El modelo (3.40) se puede escribir como

$$Var(\epsilon) = \mathtt{D}\mathtt{G}\mathtt{D}' + \sigma^2\mathtt{H} + \tau^2 I \tag{3.41}$$

donde $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ denota una secuencia genérica de m medidas de una unidad. En lo que sigue, se escribe $t = (t_1, \ldots, t_m)$ para el correspondiente conjunto de veces en que las mediciones se realizan.

Correlación serial pura

Para el primer ejemplo se asume que los efectos aleatorios y el error de medición no están presentes, tal que (3.39) se reduce a

$$\epsilon_j = \mathtt{W}(t_j)$$

y (3.41) en consecuencia se simplifica a

$$Var(\epsilon) = \sigma^2 \mathtt{H}$$

Ahora σ^2 es la varianza de cada ϵ_j y la correlación entre los ϵ_j esta determinada por la función de autocorrelación $\rho(u)$, específicamente,

$$Cov(\epsilon_j, \epsilon_k) = \sigma^2 \rho(|t_j - t_k|).$$

El correspondiente variograma es

$$\gamma(u) = \sigma^2 \{1 - \rho(u)\} \tag{3.42}$$

Así $\gamma(0) = 0$ y $\gamma(u) \to \sigma^2$ cuando $u \to \infty$. Típicamente $\gamma(u)$ es una función creciente de u porque la correlación $\rho(u)$ decrece cuando aumenta la separación del tiempo u.

Una elección popular para $\rho(u)$ es el modelo correlación exponencial,

$$\rho(u) = \exp(-\phi u)$$

para algún valor de $\phi > 0$

En muchas aplicaciones el variograma empírico se diferencia del modelo de correlación exponencial, mostrando un crecimiento inicial lento a medida que aumenta u de cero, después un fuerte aumento y finalmente, un aumento más lento de nuevo, ya que se aproxima a su asíntota. Un modelo que refleja este comportamiento es la función de correlación Gaussiana

$$\gamma(u) = \exp(-\phi u^2)$$

para algún valor de $\phi > 0$.

Correlación serial más error de medición

Estos son modelos para los cuales no hay efectos aleatorios en (3.39), tal que $y_j = \mathtt{W}(t_j) + \mathtt{Z}_j$ y (3.41) se reduce a

$$Var(\epsilon) = \sigma^2 \mathtt{H} + \tau^2 I$$

Por lo tanto, la varianza de cada ϵ_j es $\sigma^2 + \tau^2$ y si los elementos de $\mathtt{H}$ están especificados por una función de correlación $\rho(u)$, tal que $\mathtt{h}_{ij} = \rho(|t_i - t_j|)$, entonces el variograma es

$$\gamma(u) = \tau^2 + \sigma^2 \{1 - \rho(u)\} \tag{3.43}$$

Una propiedad característica de los modelos con error de medida es que $\gamma(u)$ no tiende a cero cuando u tiende a cero. Si los datos incluyen medidas duplicadas en el mismo tiempo, se puede estimar $\gamma(0) = \tau^2$ directamente como la mitad de la diferencia media cuadrática entre tales observaciones duplicadas. De otro modo, la estimación de τ^2 involucra explícitamente o implícitamente extrapolación basada sobre el modelo paramétrico que se asume y la estimación $\gamma(0)$ puede ser fuertemente dependiente del modelo.

Intercepto aleatorio más correlación serial más error de medición

Este es otro caso simple del modelo general (3.39) en el cual las tres componentes de variación están presentes y en la cual $\mathtt{U}$ es una variable aleatoria Gaussiana univariante con media cero, varianza υ y $\mathtt{d}_j = 1$. Entonces, el valor de $\mathtt{U}$ representa un intercepto aleatorio, esto es, una cantidad por la cual todas las mediciones de la unidad en cuestión aumenta o disminuye en relación con el promedio de la población. La matriz de varianza (3.41) se convierte en

$$Var(\epsilon) = \upsilon^2 J + \sigma^2 \mathtt{H} + \tau^2 I$$

donde J es una matriz de dimensión $m \times m$ con todos sus elementos igual a 1. El variograma tiene la misma forma (3.43) como para la correlación serial más modelo de medición de error,

$$\gamma(u) = \tau^2 + \sigma^2 \{1 - \rho(u)\}$$

excepto que ahora la varianza de cada ϵ_j es $Var(\epsilon_j) = \upsilon^2 + \sigma^2 + \tau^2$ y el limite de $\gamma(u)$ cuando $u \to \infty$ es menor que $Var(\epsilon_j)$.

3.2.2 Criterios de selección para la estructura $\sigma^2\Psi$

Existe una relación entre el número de grupos que se consideren y la complejidad de las matrices de covarianzas requeridas. Si se tienen muchos grupos, se puede obtener buenas soluciones imponiendo que las matrices de covarianzas sean iguales e incluso que sean de la forma $\sigma^2 I$ (Peña 2002). Por el contrario, con pocos grupos, típicamente se debe dejar bastante libertad a la matriz de varianzas y covarianzas para obtener un buen ajuste del modelo a los datos. Por esta razón, las condiciones sobre la matriz de covarianzas se

deciden conjuntamente con el número de grupos. Utilizando el criterio BIC como se presenta a continuación

$$BIC = min\sum_{h} n_h \log|S_h| + n(p,G)ln(n)$$

con $S_h = \frac{1}{n_h} = \sum_{i=1}^{n_h}(X_i - \bar{X}_h)(X_i - \bar{X}_h)'$, donde $n(p,G)$ es el número de parámetros del modelo. Conviene indicar que, aunque este criterio parece funcionar bien en la práctica para escoger el número de grupos y las condiciones del modelo. Las hipótesis de regularidad que se efectúan para deducir el BIC como aproximación de la probabilidad a-posteriori no se verifican en algunas combinaciones, por lo cual el criterio puede aplicarse como una guía y no como una regla automática. En este sentido surge el criterio Akaike (AIC) dado por

$$AIC = min\sum n_h \log|S_h| + 2n(p,G)$$

Una vez se ha hecho la conveniente elección de $\sigma^2\Psi$ se procede a estimar los parámetros.

3.2.3 Inferencia sobre la estructura de la matriz de varianzas y covarianzas

La inferencia en un modelo depende de la conveniente elección de $\sigma^2\Psi$. Por esta razón, es indispensable tener en cuenta las siguientes consideraciones planteadas por Singer & Andrade (1994): si el análisis realizado es de tipo multivariante, no es necesario tener restricciones sobre la elección de $\sigma^2\Psi$, pero si el análisis es de tipo univariante, la selección apropiada de $\sigma^2\Psi$ facilita la interpretación de las hipótesis. La elección apropiada de los términos del error y para algunas hipótesis el estadístico F derivado bajo un modelo donde la estructura de la matriz de covarianzas satisface al menos la restricción de esfericidad, haciendo que el estadístico sea exacto. En muchos casos, cuando la condición de esfericidad no se mantiene, el estadístico de prueba para las hipótesis se aproxima a una F.

En datos longitudinales, la herramienta apropiada para el análisis de perfiles es esencialmente determinada por la condición impuesta sobre la matriz de covarianza de las observaciones en el caso univariante. Singer & Andrade (1994) plantean que se podría considerar que $\sigma^2\Psi$ tiene un patrón uniforme, es decir, corresponde a la estructura compuesta simétrica (3.2). Entonces, para el análisis de la estructura, Srivastava (2001) plantea las hipótesis al respecto y sus estadísticos de prueba $H_{01} : \sigma^2\Psi = \sigma^2 I$ para verificar si cumple el supuesto de esfericidad; equivalente a contrastar si los errores son independientes.

El estadístico de prueba esta dado por

$$\lambda_1 = \frac{|S|}{(p^{-1}tra(S))^p} \tag{3.44}$$

donde S corresponde a la matriz de covarianzas muestral, cuando $n \to \infty$. Además, se puede comprobar que $Q_1 = -\left\{(N-1) - (2p^2+p+2)\frac{1}{6p}\right\}\log\lambda_1$ se distribuye asintóticamente como una chi-cuadrado, con $\frac{1}{2}p(p+1)-1$ grados de libertad. Por lo tanto, se rechaza H_{01} si $Q_1 > \chi^2_{\left(\frac{1}{2}p(p+1)-1,\alpha\right)}$.

Teniendo en cuenta que el supuesto de esfericidad no se satisface, entonces se plantea la hipótesis $H_{02} : \Psi = \Psi^*$, con la estructura de Ψ^* indicando que las medidas dentro del mismo individuo podrían estar correlacionadas y el grado de correlación no depende del orden de las medidas a través del tiempo.

El estadístico de prueba obtenido por medio de la razón de verosimilitud, esta basado en el estadístico

$$\lambda_2 = \frac{|S|}{(s^2)^p(1-r)^{p-1}[1+(p-1)r]} \tag{3.45}$$

donde $S = (s_{ij})$ es la matriz de covarianza muestral, $s^2 = p^{-1}\sum_{i=1}^{p} s_{ii}$, $r = \frac{2}{p(p-1)}\sum_{i<j}^{p} s_{ij}/s^2$. Para n grande se tiene que

$$Q_2 = -\left[N-1-\left\{p(p-1)^2(2p-3)/6(p-1)(2p+p-4)\right\}\right]\log\lambda_2,$$

se distribuye asintóticamente chi-cuadrado con $\frac{1}{2}p(p+1)-2$ grados de libertad. H_{02} se rechaza si $Q_2 > \chi^2_{\left(\frac{1}{2}p(p+1)-2,\alpha\right)}$. Otro supuesto que se debe verificar es la igualdad de varianzas para los diferentes grupos. La hipótesis $H_{03} : \Psi_1 = \Psi_2 = \cdots = \Psi_h$, verifica igualdad de covarianzas en los diferentes grupos de tratamientos. El estadístico de prueba propuesto por Bartlett (1937), para el caso es

$$\lambda_3 = \left(\frac{|S_1|}{|S|}\right)^{n_1-1/2}\left(\frac{|S_2|}{|S|}\right)^{n_2-1/2}\cdots\left(\frac{|S_h|}{|S|}\right)^{n_h-1/2} \tag{3.46}$$

donde S_i matriz de varianzas y covarianzas de cada uno de los grupos de tratamientos $i = 1, 2, \ldots, h$, y son estimadores insesgados de la varianza. Además, $S = fV_i$ con $f = \sum_i n_i - 1$, $V_i = (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$. El estadístico de prueba esta dado por $Q_3 = \frac{2}{N-1}m\log\lambda_3$ el cual se distribuye chi cuadrado con $\frac{1}{2}p(p+1)(h-1)$ grados de libertad con

$$\alpha^* = \left(\sum_{h=1}^{k} r_h^{-1} - 1\right)\frac{2p^2+3p-1}{12(p+1)(h-1)} \qquad m = N-1-2\alpha^* \text{ y } r_i = \frac{(n_i-1)}{n-1}$$

Se rechaza H_0 si $Q_3 > \chi^2_{\left(\frac{1}{2}p(p+1)(h-1),\alpha\right)}$.

Si se rechaza H_{02}, se debe hacer inferencia sobre una de las estructuras planteadas anteriormente o construir algunas estructuras que se adapten al comportamiento de los datos. La elección de la mejor estructura para el modelo de la matriz de varianzas y covarianzas $\sigma^2\Psi$, puede realizarse haciendo inferencia sobre ésta matriz. Pero en ocasiones no es tan sencillo; si se cuenta con muchos grupos de observaciones. Una alternativa es estimar varios modelos para $\sigma^2\Psi$ recurriendo a métodos iterativos implementados en paquetes estadísticos y mediante criterios de selección como: los criterios de información de Akaike (AIC) o Bayesiano (BIC), para hacer la mejor elección del modelo (Takamitsu 1978).

3.2.4 Predicción de un nuevo individuo

Suponga que un conjunto de variables explicativas mixtas $v_{n+1} = (v_{(n+1)0}, v_{(n+1)1}, \ldots, v_{(n+1)p})$ se miden para un nuevo individuo, $n+1$. Entonces, es posible calcular las distancias entre esta nueva observación y cada uno de los individuos considerados en el modelo (2.1), es decir,

$$\delta_{(n+1)i} = \delta(v_{n+1}, v_i), \quad v_i \in \Omega, \quad i = 1, \ldots, n$$

El cuadrado de las distancias, $d = (\delta^2_{(n+1)1}, \ldots, \delta^2_{(n+1)n})'$ y el vector $x_{n+1} = (x_{(n+1)1}, \ldots, x_{(n+1)p})'$ que contiene las coordenadas principales de este nuevo individuo están relacionadas por (Gower 1968)

$$\delta^2_{(n+1)i} = (x_{n+1} - x_i)'(x_{n+1} - x_i)$$

En particular, si X es la matriz que contiene las coordenadas principales, entonces

$$x_{n+1} = \frac{1}{2}\Lambda_x^{-1}X'(b - d)$$

donde $b = (b_{11}, \ldots, b_{nn})$ and $b_{ii} = x_i'x_i$, $i = 1, \ldots, n$.

La predicción, considerando el modelo DB en dimensión k, esta dado por

$$\widehat{y}_{(n+1)}(k) = (\boldsymbol{x}' \otimes I_m)\hat{\theta} \tag{3.47}$$

donde $y_{(n+1)}(k) = vec\left(Y'_{(n+1)}(k)\right)$, $\hat{\theta} = vec\left(\widehat{\boldsymbol{B}}'\right)$, $\boldsymbol{x}' = \begin{pmatrix}1 & x'_{n+1}(k)\end{pmatrix} = \left(1, x_{(n+1)1}, \ldots, x_{(n+1)k}\right)$ y $\widehat{\boldsymbol{B}} = \begin{pmatrix}\widehat{B}'_0 & \widehat{B}'_{(k)}\end{pmatrix}'$.

En la siguiente sección se presenta una simulación donde se compara el método basado en distancias (con variables mixtas usando la distancia de Gower (1971)) con respecto al método clásico para analizar datos longitudinales en forma univariante.

3.3 Simulación

3.3.1 Detalles de la simulación

Se consideraron los mismos valores y condiciones en los parámetros del modelo como en la Sección 2.3 para efectos de la simulación. Además, se uso el software estadístico R para el análisis versión 2.15 (R Development Core Team 2012).

Adicionalmente, bajo el supuesto que el mejor modelo es el que presente menor varianza generalizada de los errores de cada modelo. Se obtienen E_1 para el modelo clásico y E_2 para el modelo DB propuesto, a través de los cuales se obtiene una medida de eficiencia relativa, por medio del cociente entre E_1 y E_2 con $E_1 = \frac{det(\Xi'\Xi)}{t(n-\boldsymbol{v}-1)} \Big/ \frac{det(Y'Y)}{t(n-1)}$, siendo det el determinante y $\boldsymbol{v}$ el número de variables independientes más el intercepto. En forma similar se obtiene E_2 con los residuales del modelo DB ajustado, reemplazando $\boldsymbol{v}$ por el número de coordenadas principales que se dejen, es decir k. Si el cociente es mayor a uno sería más eficiente el método DB propuesto, para una eficiencia menor a uno resulta más eficiente el método clásico y para una eficiencia de uno los dos métodos son iguales, para cada escenario considerado.

Se obtienen también los AIC y BIC para cada modelo, siendo los modelos de mejor ajuste aquellos con valores más pequeños.

3.3.2 Resultados y discusión

Se simularon en total 126 escenarios con 4,7 y 10 tiempos, varianzas $\sigma^2 = 10, 50$, correlaciones $\rho = -0.5, 0, 0.5, 0.9$, dos estructuras de autocorrelación (AR(1)) y compuesta simétrica, y para tamaños de muestra $n = 50, 100, 200$. En esta sección se presenta un resumen de los resultados obtenidos.

La Tabla 3.1 muestra en los diferentes niveles de varianza, correlación ρ y estructuras de autocorrelación, que el método basado en distancias tiene valores más bajos de AIC y BIC para muestras pequeñas ($n = 50$), independiente del número de tiempos, la varianza y la estructura de autocorrelación que se tome. En muestras grandes resulta mejor el método clásico, ya que tiene valores más bajos de AIC y BIC. Además, el método DB presenta un mejor ajuste en muestras pequeñas, pues arroja una mejor eficiencia relativa de las varianzas generalizadas de los errores de cada modelo, ya que es mayor a uno en los escenarios considerados bajo una muestra pequeña de 50. También, resultados similares fueron obtenidos bajo los otros escenarios considerados, los cuales se pueden ver en las Tablas A.1 a A.5 presentadas en el anexo A del trabajo y cuya interpretación es similar.

La Tabla 3.2 muestra que si la varianza es grande, la muestra es grande

Parámetros			$m = 4$						
			Basado en distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	0	50	935.09	1052.88	6.14E-6	1004.22	1074.46	7.11E-6	1.19
		100	2071.39	2219.48	3.83E-5	2045.10	2131.78	1.56E-5	0.41
		200	4140.98	4317.25	1.56E-5	4120.92	4223.43	1.09E-5	0.70
	0.5	50	883.48	1001.27	5.13E-6	951.88	1022.13	6.43E-6	1.30
		100	1921.84	2069.93	2.11E-5	1934.33	2021.01	1.38E-5	0.66
		200	3877.34	4053.61	1.12E-5	3894.24	3996.74	9.78E-6	0.87
	0.9	50	702.55	820.35	3.53E-7	752.25	822.49	4.28E-7	1.27
		100	1523.83	1671.92	1.23E-6	1512.89	1599.57	8.00E-7	0.65
		200	3027.71	3203.98	6.03E-7	3029.17	3131.67	5.35E-7	0.89
50	0	50	1193.644	1311.44	1.240E-4	1293.92	1364.16	1.622E-4	1.35
		100	2590.64	2738.73	3.19E-4	2656.69	2743.37	2.64E-4	0.83
		200	5308.82	5485.09	2.10E-4	5376.28	5478.79	2.03E-4	0.97
	0.5	50	1144.99	1262.79	1.82E-4	1241.58	1311.83	2.42E-4	1.37
		100	2468.76	2616.85	4.08E-4	2545.92	2632.60	4.01E-4	0.99
		200	5076.71	5252.98	3.24E-4	5149.59	5252.10	3.31E-4	1.02
	0.9	50	963.33	1081.12	3.19E-5	1041.95	1112.19	4.18E-5	1.36
		100	2065.35	2213.44	6.93E-5	2124.47	2211.16	6.82E-5	0.99
		200	4229.00	4405.27	6.30E-5	4284.53	4387.03	6.45E-5	1.03

Tabla 3.1: Simulación con estructura de correlación compuesta simétrica

y la correlación alta, los valores de AIC son más bajos usando DB en comparación con el método clásico pero es pequeña la diferencia, independiente de la estructura de autocorrelación empleada y del número de tiempos. Adicionalmente, bajo estos mismos escenarios en correlaciones altas (0.5 y 0.9) y con más tiempos, resulta más eficiente el método DB. También, se puede notar que cuando la varianza es 10 y la correlación es 0.5, el AIC es más bajo para la aproximación basada en distancias cuando el tamaño de muestra es grande independiente del tiempo y de las estructuras de autocorrelación consideradas.

Los resultados que se presentan en las Tablas 3.1 y 3.2 representan parte de los resultados obtenidos en los diagramas de cajas de la Figura 3.1. Los diagramas de caja de la figura presenta la varianza generalizada de los errores E_1 y E_2 con los métodos clásico y DB (propuesto), respectivamente. En todos los diagramas de cajas, los números pares corresponden a la aproximación basada en distancias, mientras los números impares corresponden al método clásico. En la primera columna de gráficos de cajas, las dos primeras cajas de izquierda a derecha corresponden a muestras de tamaño 50, las siguientes dos a muestras de tamaño 100, y las últimas dos, a muestras de tamaño 200. Además en todos los gráficos de esta columna se tiene una estructura de correlación AR(1). En las otras tres columnas de gráficos de la Figura 3.1, las cuatro primeras cajas de izquierda a derecha corresponden a muestras de tamaño 50, las siguientes cuatro a muestras de tamaño 100, y las últimas cuatro, a muestras de tamaño 200. Las dos primeras cajas en cada bloque corresponden a estructuras de correlación AR(1) y las dos siguientes a la estructura de correlación compuesta simétrica.

Por ejemplo para los valores de $\rho = 0.5$, $m = 4$, y $\sigma^2 = 10$, en la Figura 3.1,

Parámetros			$m = 10$						
			Basado en Distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	-0.5	50	2252.97	2622.46	1.84E-9	2390.19	2603.87	2.32E-9	1.33
		100	5204.29	5647.12	2.36E-8	4870.68	5123.22	3.09E-9	0.13
		200	10135.41	10646.45	3.47E-8	9788.28	10078.21	9.80E-9	0.28
	0	50	2332.39	2701.88	1.39E-8	2506.61	2720.29	2.63E-8	2.07
		100	5165.40	5608.24	9.01E-8	5116.65	5369.19	2.92E-8	0.33
		200	10337.00	10848.05	1.74E-7	10293.3	10583.23	9.72E-8	0.56
	0.5	50	2218.11	2587.60	1.60E-8	2390.19	2603.87	3.56E-8	2.43
		100	4818.02	5260.86	4.01E-8	4870.68	5123.22	2.34E-8	0.59
		200	9728.51	10239.56	1.25E-7	9788.27	10078.20	1.04E-7	0.83
	0.9	50	1714.96	2084.44	4.4E-10	1834.23	2047.91	9.2E-10	2.22
		100	3697.30	4140.14	4.5E-10	3696.67	3949.21	2.7E-10	0.61
		200	7343.14	7854.19	3.28E-9	7378.33	7668.26	3.03E-9	0.93
50	-0.5	50	2879.85	3249.34	2.75E-7	3114.44	3328.12	6.03E-7	2.30
		100	6323.71	6766.55	1.35E-6	6399.65	6652.19	6.74E-7	0.51
		200	12838.55	13349.60	2.85E-6	12926.68	13216.61	2.113E-6	0.74
	0	50	2980.20	3349.69	1.74E-6	3230.86	3444.54	4.33E-6	2.63
		100	6466.99	6909.83	5.99E-6	6645.62	6898.16	5.25E-6	0.89
		200	13261.22	13772.26	1.42E-5	13431.70	13721.63	1.40E-5	0.99
	0.5	50	2872.42	3241.90	2.58E-6	3114.44	3328.12	6.54E-6	2.73
		100	6199.04	6641.88	5.17E-6	6399.65	6652.19	5.96E-6	1.17
		200	12738.78	13249.83	1.70E-5	12926.68	13216.61	1.91E-5	1.12
	0.9	50	2366.68	2736.17	1.47E-7	2558.47	2772.15	3.68E-7	2.66
		100	5071.96	5514.80	1.15E-7	5225.64	5478.17	1.36E-7	1.20
		200	10372.35	10883.39	1.12E-6	10516.73	10806.66	1.29E-6	1.16

Tabla 3.2: Simulación con estructura de autocorrelación AR(1)

se muestra que el método DB resulta ser más eficiente en muestras pequeñas (tamaño 50) independiente de la estructura de autocorrelación, la correlación y la varianza que se tome; aunque cuando el tamaño de muestra crece, resulta ser más eficiente el método clásico. Algo muy similar sucede para los valores de $\rho = 0$, $m = 4$ y $\sigma^2 = 50$.

Además, los diagramas de cajas con elecciones de $m = 10$, $\sigma^2 = 50$ y $\rho = 0.5, 0.9$, en los diferentes tamaños de muestra se puede apreciar que al aumentar el tamaño de muestra a 200 con varianza grande, correlación alta e independiente de la estructura de autocorrelación y aumentando el número de tiempos, se observa que el método basado en distancias resulta más eficiente en comparación con el clásico. Haciéndose en este caso mas grande la diferencia con el método clásico, pues se ve que se logra más eficiencia bajo este escenario.

En la siguiente sección se presenta una aplicación, en donde se puede visualizar la metodología propuesta y su comparación con el método clásico para análisis de datos longitudinales.

3.4 Aplicación

Se consideran para ilustrar los desarrollos teóricos de este capítulo los mismos datos de la National Youth Survey (Raudenbush & Chan 1992, Raudenbush

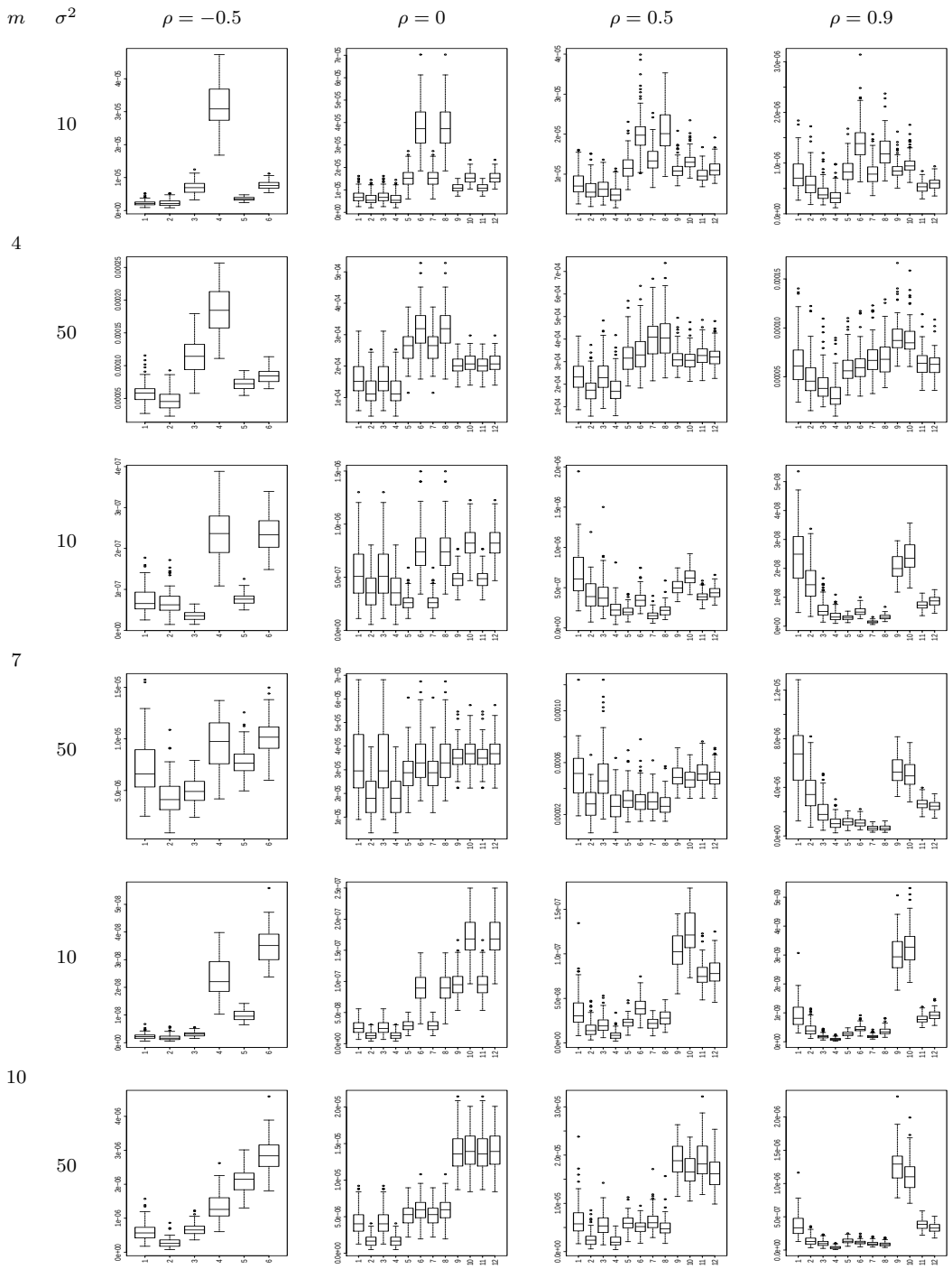


FIGURA 3.1: Varianza generalizada de los errores para DB (E_1) y análisis clásico (E_2) por tamaño de muestra, en estructuras de autocorrelación AR(1) y compuesta simétrica

& Chan 1993) que se presentaron en la Sección 2.4.

Para ver el funcionamiento del método propuesto se realizó un ajuste de un modelo lineal longitudinal a los datos usando la distancia de Gower en las variables explicativas (exposición y género) ya que son variables mixtas. La variable explicada fue tolerancia y se utilizó como criterio de ajuste del modelo los criterios de información; donde el criterio akaike (AIC) tuvo un valor de 8.89 y un criterio Bayesiano BIC de 29.51. Se determinaron las estructuras de correlación que mejor se adecuaban a los datos. Para este caso se encontró apropiado ajustar un modelo longitudinal autorregresivo de orden 1 (AR(1)), este coeficiente estimado usando distancias fue de 0.53.

Se realizó también el ajuste por medio de un modelo longitudinal clásico, donde además se encontró que la estructura mas apropiada para la matriz de correlación es un AR(1). Los valores de AIC y BIC para el método clásico fueron de 16.65 y 37.26 siendo mas altos que cuando se usó el método basado en distancias, con un coeficiente autorregresivo estimado de 0.55. Sin embargo, las predicciones resultan ser muy similares con algunas pequeñas diferencias en los últimos tiempos.

El modelo ajustado después de hacer validación de supuestos sin usar distancias es un modelo logarítmico en Y. Si I_{0j} representa los interceptos y ex_{2j} la variable exposición en los diferentes tiempos, el modelo obtenido es

$$\ln \hat{y}_{ij} = \hat{\beta}_{01} I_{01} + \hat{\beta}_{02} I_{02} + \hat{\beta}_{03} I_{03} + \hat{\beta}_{04} I_{04} + \hat{\beta}_{05} I_{05} + \hat{\beta}_{24} ex_{24} + \hat{\beta}_{25} ex_{25}$$

donde $i = 1, 2, \ldots, 16$ y $j = 1, 2, 3, 4, 5$. Así, el modelo ajustado usando estimación REML para los parámetros es

$$\ln \hat{y}_{ij} = 0.28 I_{01} + 0.34 I_{02} + 0.49 I_{03} + 0.52 I_{04} + 0.57 I_{05} + 0.13 ex_{24} + 0.17 ex_{25}$$

donde $i = 1, 2, \ldots, 16$ y $j = 1, 2, 3, 4, 5$. El modelo que se ajusta después de validar los supuestos de normalidad y homocedasticidad usando distancias fue también el logarítmico en Y, siendo I_{0j} como antes y X_{2j} corresponde a las segunda componente de la matriz de coordenadas principales, entonces el modelo encontrado es el siguiente

$$\ln \hat{y}_{ij} = \hat{\theta}_{01} I_{01} + \hat{\theta}_{02} I_{02} + \hat{\theta}_{03} I_{03} + \hat{\theta}_{04} I_{04} + \hat{\theta}_{05} I_{05} + \hat{\theta}_{24} X_{24} + \hat{\theta}_{25} X_{25}$$

donde $i = 1, 2, \ldots, 16$ y $j = 1, 2, 3, 4, 5$. Ahora, al considerar los valores obtenidos para la estimación de los parámetros por el método REML se obtiene

$$\ln \hat{y}_{ij} = 0.28 I_{01} + 0.34 I_{02} + 0.49 I_{03} + 0.52 I_{04} + 0.57 I_{05} - 0.59 X_{24} - 0.79 X_{25}$$

donde $i = 1, 2, \ldots, 16$ y $j = 1, 2, 3, 4, 5$. Se encontró al hallar el determinante de la matriz de residuos usando distancias un valor de 0.056 tomando la segunda componente de la matriz de coordenadas principales como variable

independiente ya que la primera no resulta ser significativa y para el modelo clásico sin usar distancias se obtuvo un valor del determinante de 0.065, lo cual nos muestra que sin hacer distancias la variabilidad es un poco más alta, pero son bastantes cercanos los valores. En este ejemplo, además se encontró que la variabilidad de la variable género esta recogida en la primera componente, pues al realizar una regresión simple entre la variable género y la primera componente se encuentra un coeficiente de determinación bastante alto de 0.98 lo cual nos muestra que al estar recogida la variabilidad en esta componente se puede decir que el género no tiene efecto significativo sobre la tolerancia. Además, gráficamente se puede ver que no hay diferencias significativas por género.

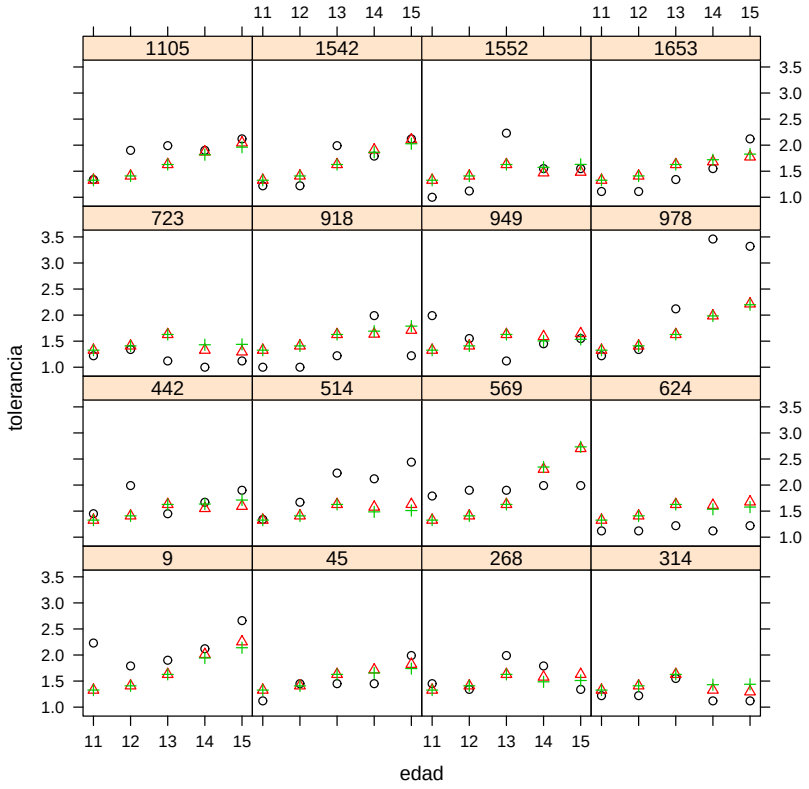


FIGURA 3.2: Tolerancia vs predicciones usando ambas aproximaciones por edad

La Figura 3.2 ilustra el ajuste de ambos modelos con respecto a los valores observados de tolerancia (los círculos). Las predicciones usando distancias (los triángulos) y el método clásico (las cruces) son casi iguales excepto para algunos individuos y en los últimos tiempos, pero difieren en muy poco. Con respecto al valor observado se aprecia un buen ajuste, excepto para el individuo 569, quien tuvo un comportamiento casi constante durante los cinco tiempos y en el ajuste se ve un comportamiento creciente.

Capítulo 4

Aproximación univariante a las curvas de crecimiento con distancias

Se consideran datos balanceados, en donde se toma el mismo número de mediciones en el tiempo para cada individuo y los tiempos son igualmente espaciados. En aquellas circunstancias donde los investigadores están principalmente interesados en hacer predicciones, la metodología propuesta resulta útil, además presenta un mejor ajuste en comparación con el modelo clásico ya que cuando se agregan componentes adicionales en el modelo se obtiene un mejor ajuste. Adicionalmente, en el análisis de curvas de crecimiento el investigador puede hacer predicciones en cada punto del tiempo, lo cual también resulta muy útil para estimar datos faltantes. Además, el uso de esta estrategia para modelar problemas de este tipo produce resultados igualmente de robustos que las estrategias de modelamiento tradicional y se utiliza en casos donde se tienen variables explicativas mixtas (categóricas, binarias y continuas).

Reinsel (1982) generalizó el modelo de Potthoff & Roy (1964) cuando varias respuestas se miden simultáneamente. Estudiaron el modelo de curva de crecimiento asumiendo efectos aleatorios. Después, Reinsel (1984) analizó el modelo de curvas de crecimiento el cual es analizado de forma multidimensional con efectos fijos y aleatorios. También, Laird & Ware (1982) discutieron la ventaja de trabajar con modelos de efectos aleatorios para datos longitudinales en dos casos, incluyendo los modelos de curvas de crecimiento y medidas repetidas como casos especiales. Luego, Verbyla & Venables (1988) trabajaron en el modelo de curvas de crecimiento como una suma de perfiles. De este modo, Lee (1991) propuso un criterio para la selección de la matriz de covarianza. Además, Faraway (1999) ajusto un modelo de regresión y uso funciones coeficiente para sugerir una forma paramétrica correspondiente. Después, Sri-

vastava (2001) estudió el modelo de curvas de crecimiento anidado propuesto por Srivastava & Katri (1979). Chiou et al. (2003) presentaron un modelo en el cual la curva de respuesta aleatoria esperada, se resume como funciones del tiempo y condicionado a una covariable, siendo productos de una función media suave del tiempo y una función suave de la covariable.

En este capítulo se presenta una metodología para analizar datos de curvas de crecimiento (Molenberghs & Verbeke 2005, Chaganty & Mav 2007) usando distancias entre pares de observaciones (Cuadras & Arenas 1990, Wessel & Schork 2006, Esteve et al. 2010) con respecto a las variables explicativas. El uso de esta estrategia es útil para predecir tanto datos faltantes como futuras observaciones, además tiene errores bajos de predicción en comparación con las curvas de crecimiento tradicional asumiendo normalidad. En este sentido, el método propuesto es mas robusto que las estrategias convencionales.

Además, en el capítulo se muestra el modelo propuesto el cual es usado para estimar los diferentes parámetros involucrados bajo el método de máxima verosimilitud restringida. La inferencia para muestras grandes y el proceso de validación del modelo propuesto se lleva a cabo. Con la aproximación sugerida, se presenta una aplicación donde se ajustan los datos al patrón de curvas de crecimiento usando la distancia valor absoluto en la variable explicativa tratamiento y también con la variable tiempo. Además, se obtienen los polinomios de Chebyshev para el ajuste de la curva de crecimiento.

El capítulo se desarrolla en siete secciones: la Sección 4.1 presenta cómo construir el modelo DB en curvas de crecimiento, la Sección 4.2 muestra el ajuste del modelo y estimación en el caso univariante, la Sección 4.3 presenta las hipótesis de interés, la Sección 4.4 muestra las distribuciones asociadas con formas cuadráticas, la Sección 4.5 muestra el análisis de varianza y pruebas estadísticas con la región de confianza, y finalmente, la Sección 4.6, muestra una aplicación de la metodología propuesta.

4.1 Construcción del modelo basado en distancias en curvas de crecimiento

El modelo de curvas de crecimiento supone que hay una sola variable de crecimiento y, la cual es medida en m puntos del tiempo $t_1, t_2, \ldots, t_m$ con n individuos elegidos en forma aleatoria. Una regresión polinomial de grado $q-1$ para y sobre la variable tiempo t es definida. Así,

$$\mathrm{E}(y_t) = \beta_{i0}t^0 + \beta_{i1}t^1 + \cdots + \beta_{i(q-1)}t^{q-1},$$

$t = t_1, \ldots, t_m, m > q-1,\ j = 1, 2, \ldots, n.$ Donde $\beta_i = [\beta_{i0}, \beta_{i1}, \ldots, \beta_{i(q-1)}]$ denota el vector de la regresión o coeficientes para la curva de crecimiento para

el i-th individuo. Las observaciones $y_{t_1}, \ldots, y_{t_m}$ sobre el mismo individuo están correlacionadas, y vienen de una distribución normal multivariante con matriz de varianzas y covarianzas desconocida Σ, igual para todos los individuos. Sea Y denota la matriz de observaciones de $n \times m$, entonces el modelo de curva de crecimiento esta dado por

$$Y = V\boldsymbol{\beta}\Upsilon + E \tag{4.1}$$

donde V es una matriz diseño de tamaño $n \times (p+1)$, $\boldsymbol{\beta}$ contiene los parámetros desconocidos de tamaño $(p+1)\times(q+1)$, Υ es la matriz de tamaño $(q+1)\times m$ que relaciona los parámetros de la curva con el correspondiente grado polinomial, y E es la matriz de errores de longitud $n\times m$ la cual tiene una distribución normal multivariante $NM(0_{n\times m}, I_n \otimes \Sigma)$. Además, $E(Y) = V\boldsymbol{\beta}\Upsilon$ y $Var(Y) = I_n \otimes \Sigma$.

Sea Ω y $\delta_{ii'}$ como en la Subsección 2.1.1. Entonces existe una configuración de puntos $v_1, \ldots, v_n \in \Re^p$, donde $v_i = (v_{i1}, \ldots, v_{ip})'$, $i = 1, \ldots, n$, como en la ecuación 2.2. Tal que la matriz V contiene las coordenadas de estos puntos, con $V = (v_{ij})$ de dimensión $n \times p$ (correspondiente a la matriz V en (4.1)) de tal forma que la distancia Euclidiana entre dos individuos i y i' es igual a ${\delta^2}_{ii'}$ (Cuadras 2008).

Por otro lado, en el modelo (4.1), la matriz V puede ser particionada como $V = (V_1 \quad V_2)$ donde V_1 es una submatriz con las variables continuas y V_2 contiene las variables cualitativas, y a partir de la ecuación (2.4) se puede obtener la similaridad. La distancia al cuadrado entre los individuos i y i' es dada en (2.5).

Si todas las variables independientes en el modelo (4.1) son cualitativas, un uso frecuente de la medida de similaridad para los individuos i y i' es $m_{ii'}$, definida en (2.6).

Por otro lado, si los puntos de la variable tiempo son equidistantes, la aproximación basada en distancias es equivalente a una regresión ordinaria sobre polinomios ortogonales; específicamente, estos son los polinomios de Chebyshev (Cuadras & Fortiana 1993). En este caso, la distancia valor absoluto esta dada por

$$d^2_{jj'} = |t_j - t_{j'}| \qquad \forall_{j\neq j'}, \quad j, j' = 1, \ldots, m$$

la cual también satisface las condiciones de una distancia Euclidiana. En el caso no equidistante también se relaciona a un conjunto de polinomios ortogonales definidos por una fórmula recurrente.

Una vez que una de las medidas de distancia ha sido seleccionada para las variables involucradas en la matriz V, se define $A_x = -\frac{1}{2}(\delta^2_{ii'})$ y $F_x = H_x A_x H_x$, donde $H_x = I_n - \frac{1}{n}J_n$ es la matriz centrada, con $J_n = \mathbf{1}_n\mathbf{1}'_n$ y $\mathbf{1}_n$ un vector de longitud n. Del mismo modo, para la variable tiempo, se define $A_g = -\frac{1}{2}(d^2_{jj'})$ y $F_g = H_g A_g H_g$, donde $H_g = I_m - \frac{1}{m}J_m$ es la matriz centrada, con $J_m = \mathbf{1}_m\mathbf{1}'_m$ y $\mathbf{1}_m$ un vector de longitud m. También, F_x y F_g son matrices definidas

semi-positivas (Mardia et al. 2002) de rango p y m, respectivamente. Por consiguiente, la descomposición espectral para cada caso es

$$\begin{aligned} F_x &= \left(I_n - \frac{1}{n}J_n\right) A_x \left(I_n - \frac{1}{n}J_n\right) \quad F_g = \left(I_m - \frac{1}{m}J_m\right) A_g \left(I_m - \frac{1}{m}J_m\right) \\ &= U_x \Lambda_x^2 U_x' = XX' \qquad\qquad = U_g \Lambda_g^2 U_g' = GG' \end{aligned}$$

donde $X = U_x\Lambda_x$ es una matriz $n \times p'$ de rango p', Λ_x es la matriz de valores propios positivos de F_x, $G = U_g\Lambda_g$ es una matriz $m \times q'$ de rango q', Λ_g es la matriz de valores propios positivos de F_g, y U_x y U_g contienen las coordenadas estándar de F_x y F_g, respectivamente. También, las filas $x_1', \ldots, x_n'$ de la matriz X son las coordenadas principales de F_x, y las filas $g_1', \ldots, g_m'$ de la matriz G son las coordenadas principales de F_g. Así, si un individuo i es similar a un individuo i' en (4.1) entonces $v_i \cong v_{i'}$ además $x_i \cong x_{i'}$; y similarmente, si un tiempo j es similar a un tiempo j' en (4.1) entonces $t_j \cong t_{j'}$ y además $g_j \cong g_{j'}$.

La relación entre $\Delta_x^{(2)} = (\delta_{ii'}^2)$ y F_x es $\Delta_x^{(2)} = \mathbf{1}f_x' + f_x\mathbf{1}' - 2F_x$, donde f_x es un vector de longitud n que contiene la diagonal de F_x. Se debe notar que si $S_x = (s_{ii'})$ es la matriz de similaridad y la función de distancia es seleccionada $\delta_{ii'}^2 = s_{ii} + s_{i'i'} - 2s_{ii'}$ entonces $F_x = H_xS_xH_x$. Lo mismo ocurre con la relación entre $\Delta_g^{(2)} = (d_{jj'}^2)$ y F_g, es decir, si $S_g = (s_{jj'})$ es la matriz de similaridad y la función distancia es seleccionada $d_{jj'}^2 = s_{jj} + s_{j'j'} - 2s_{jj'}$ entonces $F_g = H_gS_gH_g$. En ambos casos, este proceso corresponde a escalamiento multidimensional clásico o ACP (Cuadras 2010).

El modelo que se propone esta dado por

$$Y = \begin{pmatrix} \mathbf{1}_n & X \end{pmatrix} \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & B_{11} \end{pmatrix} \begin{pmatrix} \mathbf{1}_m' \\ G' \end{pmatrix} + \Xi \tag{4.2}$$

donde B_{00} es el intercepto desconocido asociado a tiempo y las variables explicativas, B_{01} es el vector de interceptos desconocidos asociado a la curva de crecimiento de tamaño $1 \times q'$, B_{10} es el vector de interceptos desconocidos asociado a las variables explicativas de tamaño $p' \times 1$, B_{11} es la matriz de parámetros desconocidos de tamaño $p' \times q'$, Ξ es la matriz de errores de dimensión $n \times m$, y las otras matrices fueron definidas arriba. Note que, como ambas $F_x\mathbf{1}_n = \mathbf{0}$ y $\mathbf{1}_n$ y las columnas $X_1, X_2, \ldots, X_{p'}$ de X son valores propios de F_x y ya que las dos $F_g\mathbf{1}_m = \mathbf{0}$, $\mathbf{1}_m$ y las columnas $G_1, G_2, \ldots, G_{m'}$ de G son valores propios de F_g.

De acuerdo a Cuadras (2007), algunas veces $p' = rang(F_x)$ crece con n (ni siquiera el caso $p' = n - 1$ se puede descartar). Entonces, el número de variables $X_1, X_2, \ldots, X_{p'}$ (columnas de X) puede ser excesivo, permitiendo un modelo arbitrariamente bien ajustado. Para evitar tales problemas, es conveniente particionar X en dos partes, $X = (X_{(k)} \quad L_1)$, donde $X_{(k)}$ contiene un suconjunto de k columnas de X y L_1 contiene el resto. Un procedimiento

similar se lleva a cabo para G', se particiona G' en dos partes, $G' = (G'_{(s)} \quad L_2)$, donde $G'_{(s)}$ contiene un subconjunto de s columnas de G' y L_2 contiene el resto. Así, un modelo basado en distancias ks dimensional se obtiene. Tal que el modelo reducido puede ser expresado como

$$Y = \boldsymbol{XBG} + \Xi_k \tag{4.3}$$

donde $\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_n & X_{(k)} \end{pmatrix} = (\mathbf{1}_n, X_1, \ldots, X_k)$ es una matriz de coordenadas principales, $\boldsymbol{G} = \begin{pmatrix} \mathbf{1}_m & G'_{(s)} \end{pmatrix} = (\mathbf{1}_m, G'_1, \ldots, G'_s)$ es la matriz que relaciona los parámetros de la curva con el correspondiente grado polinomial, Ξ_k es la matriz de errores y

$$\boldsymbol{B} = \begin{pmatrix} B_{00} & B_{01}^{(s)} \\ B_{10}^{(k)} & B_{11}^{(ks)} \end{pmatrix}$$

es una matriz de parámetros desconocidos. Por otra parte, las matrices $B_{01}^{(s)}$, $B_{10}^{(k)}$ y $B_{11}^{(ks)}$ son matrices reducidas correspondientes a la partición de X y G'.

El número de componentes principales asociadas con las variables explicativas pueden ser elegidas como k y con la variable tiempo pueden ser elegidas como s. Una buena aproximación para la selección de las columnas consiste en rankearlas de acuerdo a su coeficiente de correlación múltiple respecto a Y o $\bar{Y}_m = (\bar{Y}_{t_1}, \ldots, \bar{Y}_{t_m})$, respectivamente, es decir,

$$R^2(X_1, Y) > \cdots > R^2(X_k, Y) \qquad r^2(G'_1, \bar{Y}'_m) > \cdots > r^2(G'_s, \bar{Y}'_m)$$

donde $R^2(X_i, Y)$ es el coeficiente de determinación múltiple entre Y y X_i $(i = 1, \ldots, k)$ y $r^2(G'_j, \bar{Y}'_m)$ es el coeficiente de correlación entre G'_j y $\bar{Y}'_m$ $(j = 1, \ldots, s)$, pero en este caso, es mejor elegir un polinomio de grado dos o tres.

4.2 Ajuste del modelo y estimación en el caso univariante

Una aproximación al problema de datos de curvas de crecimiento esta basado en modelos univariantes. Porque las medidas repetidas, observaciones correspondientes al mismo individuo están correlacionadas, siendo la estructura de correlación generalmente desconocida. La estimación se lleva a cabo usualmente a través de mínimos cuadrados generalizados. El modelo reducido se presenta en (4.3) el cual puede ser expresado en forma univariante usando el operador *vec*. Entonces el modelo basado en distancias univariante esta dado por

$$y = (\boldsymbol{X} \otimes \boldsymbol{G'})\theta + \zeta \tag{4.4}$$

donde $y = vec(Y')$, $\theta = vec(\boldsymbol{B}')$ and $\zeta = vec(\Xi_k')$. Por otra parte, $Var(y) = Var(\zeta) = I_n \otimes \Sigma = \sigma^2\Psi = \sigma^2 I_n \otimes \Psi_0$ with $\Sigma = \sigma^2\Psi_0$.

Antes de estimar θ, σ^2 and Ψ_0, se requiere especificar la forma de los bloques diferentes de cero en $\sigma^2\Psi_0$ en la ecuación (4.3).

4.2.1 Estimación de parámetros

Estimación máxima verosimilitud bajo condiciones Gaussianas

Una estrategia para la estimación de parámetros en el modelo lineal general (MLG) consiste en considerar la estimación simultánea de los parámetros de interés, θ, y los parámetros de covarianza σ^2 y Ψ_0, usando la función de verosimilitud. Recuérdese que Ψ es una matriz de bloques diagonal con los bloques iguales diferentes de cero Ψ_0. Bajo la condición Gaussiana, $y \sim NM\left((\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta, \sigma^2\Psi\right)$, la logverosimilitud de los datos observados y es

$$\begin{aligned} L\left(\theta, \sigma^2, \Psi_0\right) = &- \frac{nm}{2}\log(\sigma^2) - \frac{n}{2}\log(|\Psi_0|) \\ &- \frac{1}{2\sigma^2}\left[y - (\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta\right]' \Psi^{-1}\left[y - (\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta\right] \end{aligned} \tag{4.5}$$

Dado Ψ_0, el EML para θ es el estimador MCP dado por

$$\hat{\theta}(\Psi_0) = \left[(\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1} y \tag{4.6}$$

Reemplazando en (4.5), se obtiene

$$L\left(\hat{\theta}(\Psi_0), \sigma^2, \Psi_0\right) = -\frac{1}{2}\left\{nm\log\sigma^2 + n\log(|\Psi_0|) + \sigma^{-2} SCR(\Psi_0)\right\} \tag{4.7}$$

donde

$$SCR(\Psi_0) = \left[y - (\boldsymbol{X} \otimes \boldsymbol{G'})\,\hat{\theta}(\Psi_0)\right]' \Psi^{-1}\left[y - (\boldsymbol{X} \otimes \boldsymbol{G'})\,\hat{\theta}(\Psi_0)\right]$$

Tomando derivadas en (4.7) con respecto a σ^2, manteniendo Ψ_0 fijo, el estimador EML para σ^2 se encuentra como

$$\hat{\sigma}^2(\Psi_0) = \frac{SCR(\Psi_0)}{nm} \tag{4.8}$$

De esta forma, reemplazando (4.6) y (4.8) en (4.5) la logverosimilitud reducida para Ψ_0, la cual se obtiene omitiendo el término constante es

$$L_r(\Psi_0) = L\left(\hat{\theta}(\Psi_0), \hat{\sigma}^2(\Psi_0), \Psi_0\right) = -\frac{1}{2} n\left[m\log SCR(\Psi_0) + \log(|\Psi_0|)\right] \tag{4.9}$$

Finalmente, la maximización de $L_r(\Psi_0)$ conduce a $\widehat{\Psi}_0$ y reemplazando en (4.9) y (4.14), los estimadores EML $\hat{\theta} \equiv \hat{\theta}\left(\widehat{\Psi}_0\right)$ y $\hat{\sigma}^2 \equiv \hat{\sigma}^2\left(\widehat{\Psi}_0\right)$ se obtienen.

Estimación máxima verosimilitud restringida

El estimador REML $\hat{\alpha}$ que maximiza la logverosimilitud es

$$L^*(\alpha) = -\frac{1}{2}\log|\Psi_1| - \frac{1}{2}\log\left|(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi_1^{-1}\right| - \frac{1}{2}\left[y-(\boldsymbol{X}\otimes\boldsymbol{G'})\hat{\theta}\right]'\Psi_1^{-1}\left[y-(\boldsymbol{X}\otimes\boldsymbol{G'})\hat{\theta}\right]$$

donde el EML $\hat{\alpha}$ maximiza

$$L(\alpha) = -\frac{1}{2}\log|\Psi_1| - \frac{1}{2}\left[y-(\boldsymbol{X}\otimes\boldsymbol{G'})\hat{\theta}\right]'\Psi_1^{-1}\left[y-(\boldsymbol{X}\otimes\boldsymbol{G'})\hat{\theta}\right]$$

De este último resultado se puede deducir que el algoritmo para implementar la estimación REML para el modelo $y \sim NM((\boldsymbol{X}\otimes\boldsymbol{G'})\theta, \sigma^2\Psi)$ es una simple modificación sobre el algoritmo EML. Por lo tanto, el estimador θ para un Ψ_0 dado es

$$\hat{\theta}(\Psi_0) = \left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}y \tag{4.10}$$

y también

$$SCR(\Psi_0) = \left\{y-(\boldsymbol{X}\otimes\boldsymbol{G'})\hat{\theta}(\Psi_0)\right\}'\Psi^{-1}\left\{y-(\boldsymbol{X}\otimes\boldsymbol{G'})\hat{\theta}(\Psi_0)\right\}$$

Ya que se consideraron n unidades con m medidas por unidad, donde $\sigma^2\Psi$ es una matriz de bloques diagonal cuyos bloques diferentes de cero son de tamaño $m \times m$. Además, el estimador REML para σ^2 es

$$\hat{\sigma}^2(\Psi_0) = SCR(\Psi_0)/(nm-k^*) \tag{4.11}$$

donde k^* es el número de elementos en θ. El estimador REML para Ψ_0 que maximiza la logverosimilitud reducida es

$$L^*(\Psi_0) = -\frac{1}{2}n\left\{m\log SCR(\Psi_0) + \log|\Psi_0|\right\} - \frac{1}{2}\log\left|(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right|$$

Reemplazando el estimador resultante $\widehat{\Psi}_0$ en (4.10) y (4.11), se obtienen los estimadores REML, es decir $\hat{\theta} = \hat{\theta}(\widehat{\Psi}_0)$ y $\hat{\sigma}^2 = \hat{\sigma}^2(\widehat{\Psi}_0)$.

4.2.2 Predicción de un nuevo individuo

Suponga que un conjunto de variables explicativas mixtas $v_{n+1} = (v_{(n+1)0}, v_{(n+1)1}, \ldots, v_{(n+1)p})'$ se miden para un nuevo individuo, $n+1$. Entonces es posible calcular las distancias entre esta nueva observación y cada uno de los individuos considerados en el modelo (4.1), es decir,

$$\delta_{(n+1)i} = \delta(v_{n+1}, v_i), \quad v_i \in \Omega, \quad i = 1, \ldots, n$$

El cuadrado de estas distancias, $d = (\delta^2_{(n+1)1}, \ldots, \delta^2_{(n+1)n})'$ y el vector $x_{n+1} = (x_{(n+1)1}, \ldots, x_{(n+1)p})'$ que contiene las coordenadas principales de este nuevo individuo están relacionadas por

$$\begin{aligned}\delta^2_{(n+1)i} =& (x_{n+1} - x_i)'(x_{n+1} - x_i) \\ =& x'_{n+1}x_{n+1} + x'_i x_i - 2x'_{n+1}x_i \end{aligned} \tag{4.12}$$

Este resultado es la base para hacer predicciones sobre este nuevo individuo. Agregando las expresiones anteriores de 1 a n y teniendo en cuenta que las columnas de la matriz de coordenadas X tienen suma cero,

$$\sum_{i+1}^{n} \delta^2_{(n+1)i} = nx'_{n+1}x_{n+1} + tr(\mathtt{B})$$

Reemplazando esta expresión en (4.12), se obtiene

$$2x'_{n+1}x_i = \frac{1}{n}\left(\sum_{i=1}^{n} \delta^2_{(n+1)i} - tr(\mathtt{B})\right) + b_{ii} - \delta^2_{(n+1)i}$$

lo cual puede ser expresado en forma matricial como

$$2Xx_{n+1} = \frac{1}{n}\left(\sum_{i=1}^{n} \delta^2_{(n+1)i} - tr(\mathtt{B})\right) 1_n + (b - d)$$

donde $b = (b_{11}, \ldots, b_{nn})$ es la diagonal de F_x con $b_{ii} = x'_i x_i$, $i = 1, ..., n$ y $d = (\delta^2_{(n+1)1}, \ldots, \delta^2_{(n+1)n})'$ son los cuadrados de las distancias.

Multiplicando por X' y dado que $X'1_n = 0$, se obtiene

$$\begin{aligned} 2X'Xx_{n+1} =& X'(b - d) \\ x_{n+1} =& \frac{1}{2}(X'X)^{-1}X'(b - d) \\ =& \frac{1}{2}\Lambda_x^{-1}X'(b - d) \end{aligned} \tag{4.13}$$

La predicción es entonces

$$\widehat{Y}_{(n+1)} = \begin{pmatrix} 1 & x'_{(k)} \end{pmatrix} \begin{pmatrix} \widehat{B}_{00} & \widehat{B}_{01} \\ \widehat{B}_{10} & \widehat{B}_{11} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_m \\ G' \end{pmatrix}$$

En forma reducida, la predicción esta dada por

$$\widehat{Y}_{(n+1)} = \begin{pmatrix} 1 & x'_{(k)} \end{pmatrix} \widehat{\boldsymbol{B}}\boldsymbol{G}$$

4.2.3 Hipótesis de interés y pruebas estadísticas

Adicionalmente, el vector de los errores en el modelo (4.4) se puede expresar como

$$\zeta = y - (\boldsymbol{X} \otimes \boldsymbol{G'})\theta$$

Ahora por mínimos cuadrados generalizados se obtiene

$$\begin{aligned}\zeta'\Psi^{-1}\zeta =& \left[y - (\boldsymbol{X} \otimes \boldsymbol{G'})\theta\right]' \Psi^{-1} \left[y - (\boldsymbol{X} \otimes \boldsymbol{G'})\theta\right] \\ =& y'\Psi^{-1}y - 2\left[(\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta\right]' \Psi^{-1}y + \left[(\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta\right]' \Psi^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta \\ =& y'\Psi^{-1}y - 2\theta' (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}y + \theta' (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})\,\theta\end{aligned}$$

Derivando parcialmente con respecto a θ e igualando a cero y como $(\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}\ (\boldsymbol{X} \otimes \boldsymbol{G'})$ es no singular se encuentra el estimador de los parámetros, que se presenta a continuación

$$\hat{\theta} = \left[(\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}y \tag{4.14}$$

El valor esperado del vector de parámetros estimado es θ, por consiguiente es insesgado, veamos esto.

$$\begin{aligned}E\left(\hat{\theta}\right) =& E\left(\left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}y\right) \\ =& \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}E\,(y) \\ =& \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\theta \\ =& \theta\end{aligned} \tag{4.15}$$

Los residuales se pueden expresar como

$$\hat{\zeta} = y - (\boldsymbol{X} \otimes \boldsymbol{G'}) \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}y$$

$$\hat{\zeta} = \left\{I - (\boldsymbol{X} \otimes \boldsymbol{G'}) \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}\right\} y$$

Su valor esperado es

$$\begin{aligned}E\left[\hat{\zeta}\right] =& E\left\{\left(I - (\boldsymbol{X} \otimes \boldsymbol{G'}) \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}\right) y\right\} \\ =& \left(I - (\boldsymbol{X} \otimes \boldsymbol{G'}) \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \Psi^{-1}\right) E(y) \\ =& (\boldsymbol{X} \otimes \boldsymbol{G'})'\theta - (\boldsymbol{X} \otimes \boldsymbol{G'}) \left[(\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X} \otimes \boldsymbol{G'})\right]^{-1} \\ & (\boldsymbol{X} \otimes \boldsymbol{G'})'\Psi^{-1} (\boldsymbol{X} \otimes \boldsymbol{G'})' \theta = 0\end{aligned}$$

Lo cual confirma que la media de los residuales es cero. Además, la varianza del vector de residuales es

$$\begin{aligned} Var\left[\hat{\zeta}\right] =& \boldsymbol{M} Var\left(y\right) \boldsymbol{M}' \\ =& \boldsymbol{M}\sigma^2\Psi \end{aligned}$$

donde $\boldsymbol{M} = I - \left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\left[\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)'\Psi^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\right]^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)'\Psi^{-1}$.

Luego se verifica que la distribución de $\hat{\zeta}$ es

$$\hat{\zeta} \sim N\left(0, \boldsymbol{M}\sigma^2\Psi\right)$$

Davis (2002) trabaja un estimador de la varianza insesgado dado por

$$\hat{\Sigma} = S = \frac{1}{nq' - sq'}\left(y - \left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\theta\right)\left(y - \left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\theta\right)' \tag{4.16}$$

La varianza de $\hat{\theta}$ es

$$\begin{aligned} Var\left[\hat{\theta}\right] =& Var\left[\left[\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)'\Psi^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\right]^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)'\Psi^{-1}y\right] \\ =& \left[\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)'\Psi^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\right]^{-1} \end{aligned} \tag{4.17}$$

Finalmente, el vector de parámetros $\hat{\theta}$ se distribuye así

$$\hat{\theta} \sim N\left(\theta; \left[\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)'\Psi^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G'}\right)\right]^{-1}\right) \tag{4.18}$$

4.3 Hipótesis de interés

Teniendo en cuenta el planteamiento dado por Crowder & Hand (1990), las hipótesis de interés están dadas para determinar; el efecto del tratamiento a través del tiempo, el efecto de los tratamientos sobre los efectos principales y sus interacciones en el tiempo.

En el análisis multivariante realizado por Ortiz et al. (2012), la matriz de parámetros en (4.6) en cada una de sus posiciones representa una combinación de la interacción de los parámetros de la curva y la superficie de respuesta. Sí se esta interesado en conocer el efecto de la superficie de respuesta a través del tiempo, entonces se centra el análisis en las filas de esta matriz, pero si por el contrario se desea conocer el efecto que tiene la curva de crecimiento (tiempo) a través de la superficie de respuesta (tratamientos) el análisis se centra en las columnas de esta matriz.

Una vez que se han estimado los parámetros, el interés ahora es plantear las hipótesis, para encontrar el tipo de relación de los tiempos y los tratamientos con la respuesta. En el modelo clásico se tiene una hipótesis de la forma

$$H_0 : A\boldsymbol{\beta}R = K \quad vs \quad H_1 : A\boldsymbol{\beta}R \neq K$$

En forma equivalente para el análisis univariante, después de aplicar distancias se obtiene la hipótesis $H_0 : C\boldsymbol{B}U = D \quad vs \quad H_1 : C\boldsymbol{B}U \neq D$ se le aplica el operador *vec*. El planteamiento general de las hipótesis corresponde a la forma planteada en (4.4) para el vector de parámetros, la matriz C es una matriz que contiene los coeficientes que permiten validar la hipótesis "dentro de tiempos" (es decir, la hipótesis en los elementos dentro de las columnas en $\boldsymbol{B}$). La matriz U permite probar hipótesis "entre tiempos" (es decir, en los elementos dentro de las filas en $\boldsymbol{B}$). Finalmente D es una matriz de constantes. Considerese ahora la siguiente hipótesis

$$H_0 : L\theta = d \tag{4.19}$$

donde $vec(C\boldsymbol{B}U)' = (C \otimes U')vec(\boldsymbol{B'})$ con $L = C \otimes U'$ y $d = vec(D)$ es un vector de constantes, donde (4.19) tiene la forma de la hipótesis lineal dada por Kshirsagar & Boyce (1995). Las hipótesis más comunes se hacen sobre tiempos y tratamientos.

Un estimador para contrastar H_0 esta dado por

$$L\hat{\theta} = d \tag{4.20}$$

El valor esperado de la anterior expresión es

$$E\left(L\hat{\theta} - d\right) = LE(\hat{\theta}) - d = L\theta - d$$

teniendo en cuenta (4.15), se tiene

$$L\theta = d$$

entonces este estimador es insesgado. Ahora la varianza de (4.20) es

$$\begin{aligned}
Var\left(L\hat{\theta} - d\right) =& Var\left\{L\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}y\right\} \\
=& \left\{L\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}\right\}Var\left[y\right] \\
& \left\{L\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}\right\}' \\
=& L\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}\Psi \\
& \Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}L' \\
=& L\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}L'
\end{aligned}$$

La distribución de esta forma es la siguiente

$$\left(L\hat{\theta} - d\right) \sim N\left(L\theta - d; \quad L\left[(\boldsymbol{X}\otimes\boldsymbol{G'})'\Psi^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})\right]^{-1}L'\right)$$

4.4 Distribuciones asociadas a las formas cuadráticas

Haciendo $\Psi = PP'$, premultiplicando (4.4) por P^{-1} y teniendo en cuenta que $\zeta \sim N(0, \Psi)$, se encuentra

$$P^{-1}y = P^{-1}(\boldsymbol{X} \otimes \boldsymbol{G}')\theta + P^{-1}\zeta$$

Ahora considérese el modelo

$$Y^{\Delta} = \boldsymbol{X}^{\Delta}\boldsymbol{B}^{\Delta} + \Xi_k^{\Delta} \tag{4.21}$$

donde $Y^{\Delta} = P^{-1}y$, $\boldsymbol{X}^{\Delta}\boldsymbol{B}^{\Delta} = P^{-1}(\boldsymbol{X} \otimes \boldsymbol{G}')\theta$ y $\Xi_k^{\Delta} = P^{-1}\zeta$. Los nuevos errores son

$$\Xi_k^{\Delta} = Y^{\Delta} - \boldsymbol{X}^{\Delta}\boldsymbol{B}^{\Delta}$$

Su valor esperado se obtiene mediante

$$E(\Xi_k^{\Delta}) = E\left[P^{-1}\zeta\right] = P^{-1}E\left[\zeta\right] = 0$$

y la respectiva varianza

$$\begin{aligned} Var(\Xi_k^{\Delta}) =& Var(P^{-1}\zeta) = P^{-1}Var(\zeta)(P^{-1})' \\ =& P^{-1}\Psi(P^{-1})' = P^{-1}PP'(P^{-1})' = I \end{aligned}$$

De tal forma que se distribuye $\Xi_k^{\Delta} \sim N(0, I)$. La expresión $\Psi = PP'$ se factoriza utilizando la descomposición de Schur, entonces en este caso $\boldsymbol{K} = \Psi$ es una matriz simétrica y se puede hacer uso de la expresión $S'\boldsymbol{K}S = \Lambda$ como se muestra a continuación

$$\Psi = S\Lambda^{1/2}\Lambda^{1/2}S,$$

luego $P = S\Lambda^{1/2}$ y $P' = \Lambda^{1/2}S$.

Para estimar el vector de parámetros $\boldsymbol{B}^{\Delta}$ y como se tiene un modelo clásico, se obtiene

$$\begin{aligned} \hat{\boldsymbol{B}}^{\Delta} &= \left(\boldsymbol{X}^{\Delta'}\boldsymbol{X}^{\Delta}\right)^{-1}\boldsymbol{X}^{\Delta'}Y^{\Delta} \\ &= \left[\left(P^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)\right)'P^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)\right]^{-1}\left[P^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)\right]'P^{-1}y \\ &= \left[\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)'\left(PP'\right)^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)\right]^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)'\left(PP'\right)^{-1}y \\ &= \left[\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)\right]^{-1}\left(\boldsymbol{X} \otimes \boldsymbol{G}'\right)'\Psi^{-1}y. \end{aligned}$$

Esta estimación coincide con la dada en la expresión (4.10).

Las sumas de cuadrados asociadas con el modelo transformado (4.21), corresponde a las planteadas en el modelo clásico. Estas se presentan a continuación:

La suma de cuadrados del modelo esta dada por la siguiente expresión

$$\begin{aligned} SCM =& (Y^{\Delta})' \boldsymbol{X}^{\Delta} \left(\boldsymbol{X}^{\Delta'} \boldsymbol{X}^{\Delta} \right)^{-1} \boldsymbol{X}^{\Delta'} Y^{\Delta} \\ =& y' \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \left(\left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \right)^{-1} \\ & \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} y \end{aligned} \tag{4.22}$$

haciendo $\boldsymbol{M}_1 = \left(\left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \right)^{-1}$, entonces (4.22) se puede expresar como

$$\begin{aligned} SCM =& y' \left[\Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \right] y \\ =& y' A_1 y \end{aligned}$$

donde $A_1 = \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1}$ Además, hay que verificar si la matriz A_1 es idempotente

$$\begin{aligned} A_1 \Psi A_1 \Psi =& \left[\Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \right] \Psi [\Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \boldsymbol{M}_1 \\ & \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1}] \Psi \\ =& \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \\ =& A_1 \Psi \end{aligned}$$

Puesto que $\boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) = I$, por lo tanto la matriz $A_1 \Psi$ es idempotente. Por otro lado la suma de cuadrados de los residuos es

$$\begin{aligned} SCE =& \left(Y^{\Delta} - \boldsymbol{X}^{\Delta} \hat{\boldsymbol{B}}^{\Delta} \right)' \left(Y^{\Delta} - \boldsymbol{X}^{\Delta} \hat{\boldsymbol{B}}^{\Delta} \right) \\ =& Y^{\Delta\prime} Y^{\Delta} - 2 Y^{\Delta\prime} \boldsymbol{X}^{\Delta} (\boldsymbol{X}^{\Delta\prime} \boldsymbol{X}^{\Delta})^{-1} \boldsymbol{X}^{\Delta\prime} Y^{\Delta} \\ & + Y^{\Delta\prime} \boldsymbol{X}^{\Delta} (\boldsymbol{X}^{\Delta\prime} \boldsymbol{X}^{\Delta})^{-1} \boldsymbol{X}^{\Delta\prime} \boldsymbol{X}^{\Delta} (\boldsymbol{X}^{\Delta\prime} \boldsymbol{X}^{\Delta})^{-1} \boldsymbol{X}^{\Delta\prime} Y^{\Delta} \\ =& Y^{\Delta\prime} Y^{\Delta} - Y^{\Delta\prime} X^{\Delta} (\boldsymbol{X}^{\Delta\prime} \boldsymbol{X}^{\Delta})^{-1} \boldsymbol{X}^{\Delta\prime} Y^{\Delta} \\ =& y' \Big\{ (P'P)^{-1} - (PP')^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \\ & \left[\left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' (PP')^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right) \right]^{-1} \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' (PP')^{-1} \Big\} y \\ =& y' \left\{ \Psi^{-1} - \Psi^{-1} \left(\boldsymbol{X} \otimes G' \right) \boldsymbol{M}_1 \left(\boldsymbol{X} \otimes \boldsymbol{G}' \right)' \Psi^{-1} \right\} y \\ =& y' A_2 y \end{aligned}$$

donde $A_2 = \Psi^{-1} - A_1$. Ahora se debe verificar que $A_2 \Psi$ es idempotente

$$\begin{aligned} A_2 \Psi A_2 \Psi =& (\Psi^{-1} - A_1) \Psi (\Psi^{-1} - A_1) \Psi \\ =& (I - A_1 \Psi)(I - A_1 \Psi) \\ =& I - 2 A_1 \Psi + A_1 \Psi A_1 \Psi \\ =& I - A_1 \Psi = (\Psi^{-1} - A_1) \Psi = A_2 \Psi \end{aligned}$$

Por lo tanto, $A_2\Psi$ es una matriz idempotente. A continuación se demuestra que la suma de cuadrados del modelo y del residuo son independientes, es decir $A_1\Psi A_2\Psi = 0$, entonces

$$\begin{aligned} A_1\Psi A_2\Psi =& A_1\Psi(\Psi^{-1} - A_1)\Psi \\ =& A_1\Psi - A_1\Psi A_1\Psi = A_1\Psi - A_1\Psi = 0 \end{aligned}$$

ya que $A_1\Psi$ es idempotente. Por lo tanto, la SCM y la SCE son independientes. Finalmente la suma de cuadrados total esta dada por

$$SCT = Y^{\Delta\prime}Y^{\Delta} = y'\Psi^{-1}y$$

La suma de cuadrados de la hipótesis general se plantea a partir de (4.19), obteniendo

$$\begin{aligned} SCH =& \left(L\hat{\theta} - d\right)' \left\{L\left(\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\right)^{-1}L'\right\}^{-1}\left(L\hat{\theta} - d\right) \\ =& \left(L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}y - d\right)'\left[L\boldsymbol{M}_1L'\right]^{-1}\left(L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}y - d\right) \\ =& \left\{y'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L' - d'\right\}\left[L\boldsymbol{M}_1L'\right]^{-1} \\ & \left\{L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}y - d\right\} \end{aligned}$$

Si $d = 0$, entonces,

$$\begin{aligned} SCH =& y'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}y \\ =& y'A_0y \end{aligned} \tag{4.23}$$

donde $A_0 = \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}$.

Ahora se verifica si $A_0\Psi$ es idempotente, por consiguiente

$$\begin{aligned} A_0\Psi A_0\Psi =& \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\Psi\Psi^{-1} \\ & \left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\Psi \\ =& \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1} \\ & \left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\Psi \\ =& \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\Psi \\ =& A_0\Psi \end{aligned}$$

donde $\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right) = I$. Comprobándose de esta manera que $A_0\Psi$ es idempotente.

A continuación, se verifica que la SCH y la SCE son independientes, es decir

$$\begin{aligned} A_0\Psi A_2\Psi =& \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\Psi \\ & \left[\Psi^{-1} - \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\right]\Psi \\ =& \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)' \\ & - \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1L'\left[L\boldsymbol{M}_1L'\right]^{-1}L\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)' = 0 \end{aligned}$$

De esta manera se demuestra que $A_0\Psi$ y $A_2\Psi$ son independientes.

4.4.1 Distribución de la suma de cuadrados del error y del modelo

Teniendo las sumas de cuadrados respectivas, habiendo verificado idempotencia y comprobada la independencia entre SCM y SCE, la distribución de la SCM y SCE son:

La $SCM \sim \chi^2_{(rang A_1, \lambda)}$, donde el rango de A_1 es

$$\begin{aligned} rang(A_1\Psi) =& rang\left\{\left[\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\right]\Psi\right\} \\ =& tr\left[\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\right] \\ =& tr\left[\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\right] \\ =& tr(I_{sk}) = sk \end{aligned}$$

y el parámetro de no centralidad

$$\begin{aligned} \lambda =& \frac{1}{2}\mu' A_1 \mu \\ =& \frac{1}{2}\zeta'\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\left[\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\right]\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\zeta \\ =& \frac{1}{2}\zeta'\left[\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\right]\zeta \end{aligned}$$

entonces $SCM \sim \chi^2_{\left(sk;\frac{1}{2}\zeta'\left[\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\right]\zeta,\alpha\right)}$. De la misma forma,

$$\begin{aligned} rang(A_2\Psi) =& rang\left\{\left[\Psi^{-1} - \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\right]\Psi\right\} \\ =& tr\left[I_{nm} - \Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\right] \\ =& tr\left(I_{nm}\right) - tr\left(I_{sk}\right) = nm - sk \end{aligned}$$

y el parámetro de no centralidad

$$\begin{aligned} \lambda =& \frac{1}{2}\mu' A_2 \mu \\ =& \frac{1}{2}\zeta'\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)' A_2 \left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\zeta \\ =& \frac{1}{2}\Big\{\zeta'\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\zeta - \zeta'\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\boldsymbol{M}_1 \\ & \left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G}'\right)\zeta\Big\} = 0 \end{aligned}$$

De esta forma, $SCE \sim \chi^2_{(nm-sk)}$.

4.5 Análisis de varianza y estadísticos de prueba

Antes de validar las hipótesis se debe tener en cuenta que el modelo se ajuste a la estructura de los datos. Por esta razón se hace necesario realizar el análisis de varianza sobre el modelo propuesto. En la Tabla 4.1 se presenta dicho análisis.

Tabla 4.1: Análisis de varianza para verificar el ajuste del modelo.

Fuentes de Variación	Suma de cuadrados	gl	Estadístico
Modelo	$y'\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G'}\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G'}\right)'\Psi^{-1}y$	sk	$F=\frac{CMM}{CME}$
Error	$y'\left[\Psi^{-1}-\Psi^{-1}\left(\boldsymbol{X}\otimes\boldsymbol{G'}\right)\boldsymbol{M}_1\left(\boldsymbol{X}\otimes\boldsymbol{G'}\right)'\Psi^{-1}\right]y$	$nm-sk$	
Total	$y'\Psi^{-1}y$	nm	

donde $CMM=\frac{SCM}{sk}$ y $CME=\frac{SCE}{nm-sk}$

Cuando ya se ha verificado el ajuste del modelo, se realiza el análisis sobre cada una de las hipótesis planteadas. A continuación se presenta el estadístico de prueba para verificar las hipótesis.

4.5.1 Estadístico de prueba para el ajuste del modelo

Una vez se estima las sumas de cuadrados del modelo, de los errores y de las hipótesis, el estadístico de prueba para la validación del modelo, esta dado por

$$F=\frac{SCM/sk}{CME}\sim F_{(sk,nm-sk,\alpha)}$$

Si $F>F_{(sk,np-sk,\alpha)}$, entonces se rechaza H_0 que todos los parámetros asociados a las coordenadas principales son ceros, y por lo tanto, el modelo sería significativo. La hipótesis (4.20), se verifica a partir del siguiente estadístico de prueba

$$F=\frac{SCH/gl_H}{CME}\sim F_{(rang(H),nm-sk,\alpha)}$$

Este estadístico tendrá una distribución F central bajo H_0 cierta. Además, como las estimaciones se han trabajado bajo el supuesto de que la estructura de Σ es conocida, estos estadísticos tienen una distribución F exacta. Si $F>F_{(rang(H),nm-sk,\alpha)}$, entonces se rechaza $H_0: Lvec(\boldsymbol{B'})=d$.

Adicionalmente, la región de confianza para juzgar que tan confiable es la hipótesis dada en (4.19), a continuación se presenta la región de confianza con

un nivel de confianza de $100(1-\alpha)\%$:

$$\frac{\left[(L\hat{\theta}-d)-(L\theta-d)\right]'(L^{-1}\boldsymbol{M}_1L')^{-1}\left[(L\hat{\theta}-d)-(L\theta-d)\right]}{rang(L)CME} \leq F_{(rangL,nm-sk,\alpha)}$$

$$\left[L\hat{\theta}-L\theta\right]'(L\boldsymbol{M}_1L')^{-1}\left[L\hat{\theta}-L\theta\right] \leq rang(L)F_{(rangL,nm-sk,\alpha)}CME \tag{4.24}$$

4.5.2 Algunas consideraciones del estadístico de prueba F

Los desarrollos realizados hasta el momento se han hecho bajo el supuesto que Σ es conocida; pero esto por lo general no se tiene. Al respecto Singer & Andrade (1994), trabajaron esta situación en el estudio de medidas repetidas para ajustar modelos superparametrizados bajo un análisis univariante, el cual facilita la especificación de las hipótesis, los términos del error y el estadístico de prueba. Al respecto plantearon, el estadístico de prueba para la hipótesis lineal, el cual esta dado por

$$F = u^*(n-m)SCH_1/uaSCE_1 \tag{4.25}$$

donde u^* corresponde a los grados de libertad de la matriz U^* y a a los grados de libertad de la matriz A_1. Además,

$$\begin{aligned} SCH_1 =& y'\left[U(U'U)^{-1}U'\otimes A_1\right]y \\ =& tr\left(y'A_1yU(U'U)^{-1}U'\right) \end{aligned} \tag{4.26}$$

y

$$\begin{aligned} SCE_1 =& y'\left\{U^*U^{*\prime}\otimes\left[I_n-(\boldsymbol{X}\otimes\boldsymbol{G'})[(\boldsymbol{X}\otimes\boldsymbol{G'})'(\boldsymbol{X}\otimes\boldsymbol{G'})]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\right]\right\}y \\ =& tr\left\{y'\left[I_n-(\boldsymbol{X}\otimes\boldsymbol{G'})[(\boldsymbol{X}\otimes\boldsymbol{G'})'(\boldsymbol{X}\otimes\boldsymbol{G'})]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'\right]yU^*U^{*\prime}\right\} \end{aligned} \tag{4.27}$$

donde $A_1 = (\boldsymbol{X}\otimes\boldsymbol{G'})[(\boldsymbol{X}\otimes\boldsymbol{G'})'(\boldsymbol{X}\otimes\boldsymbol{G'})]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'C'\{C[(\boldsymbol{X}\otimes\boldsymbol{G'})'(\boldsymbol{X}\otimes\boldsymbol{G'})]^{-1}C'\}^{-1}C[(\boldsymbol{X}\otimes\boldsymbol{G'})'(\boldsymbol{X}\otimes\boldsymbol{G'})]^{-1}(\boldsymbol{X}\otimes\boldsymbol{G'})'$ y U^* es una matriz de orden $m\times u^*$ elegida convenientemente, esta depende de la estructura de la hipótesis que se este considerando y A_1 es la matriz que mide el efecto a través de los grupos de tratamientos. Bajo las siguientes proposiciones, se dan algunas consideraciones de la distribución de las formas cuadráticas de acuerdo a la estructura que presente Σ.

Proposición 4.1. *Considerando el modelo dado por* $y = (\boldsymbol{X}\otimes\boldsymbol{G'})\theta+\zeta$ *con* $E(y) = (\boldsymbol{X}\otimes\boldsymbol{G'})\theta$ *y* $Var(y) = I_n\otimes\Sigma$ *y bajo la hipótesis* $H_0 : C\boldsymbol{B}U = D$*, donde* $\boldsymbol{B}$ *es la matriz de parámetros,* C *es una matriz cuyos coeficientes permiten probar la hipótesis dentro de tiempos y la matriz* U *cuyos coeficientes permiten validar la hipótesis entre tiempos:*

1. $E(SCH) = c\ tr(U'\Sigma U(U'U)^{-1})$.

2. $SCH \sim \sum_{l=1}^{c} \theta_l \chi_l^2(c)$, *donde* $\chi_l^2(c)$ *son variables aleatorias independientes que siguen una distribución chi-cuadrado central con c grados de libertad y* θ_l *son los valores propios de* $U'\Sigma U(U'U)^{-1}$.

Proposición 4.2. *Bajo los mismos supuestos de la Proposición 4.1 y si la hipótesis* $H_0 : C\boldsymbol{B}U = D$ *se conserva o no:*

1. $E(SCE) = (n-m)tr(U^{*\prime}\Sigma U^*)$.

2. $SCE \sim \sum_{l=1}^{u^*} \phi_l \chi_l^2(n-m)$, *donde* $\chi_l^2(n-m)$ *son funciones de variables aleatorias independientes que siguen una distribución chi-cuadrado con* $(n-m)$ *grados de libertad y* ϕ_l *son valores los propios reales de* $U^{*\prime}\Sigma U^*$ *diferentes de cero.*

3. SCE *es independiente de* SCH.

Proposición 4.3. *Bajo los mismos supuestos de la Proposición 4.1 y cuando la hipótesis nula* $H_0 : C\boldsymbol{B}U = D$, *es cierta y el estadístico* (4.25), *sigue aproximadamente una distribución F con* ϵuc *grados de libertad en el numerador y* $\epsilon u^*(n-m)$ *grados de libertad en el denominador, donde*

$$\epsilon = \left\{tr\left[U'\Sigma U(U'U)^{-1}\right]\right\}^2 \Big/ \left\{u\ tr\left[(U'\Sigma U(U'U)^{-1})\right]^2\right\}$$

si $U^* = U(U'U)^{-1}$. *Adicionalmente, si todos los valores propios de* $U'\Sigma U(U'U)^{-1}$ *son iguales, entonces* $\epsilon = 1$ *y se tiene una distribución exacta del estadístico de prueba.*

La demostración de las anteriores proposiciones se encuentran en Singer & Andrade (1994).

De acuerdo con la forma general dada en el modelo, $(\boldsymbol{X} \otimes \boldsymbol{G'})$ es la matriz de coordenadas principales entre variables explicativas y tiempos asociada al vector de parámetros θ; luego se define A_1 y se selecciona convenientemente U^*. En el análisis clásico univariante, donde $\Sigma = \sigma^2 I$ y $U = I_p$, en la ecuación (4.27) para toda hipótesis de la forma $H_0 : C\boldsymbol{B}U = D$, el estadístico (4.25) sigue una distribución F exacta con uc grados de libertad en el numerador y $k(n-m)$ grados de libertad en el denominador; pero si Σ presenta otra estructura, la selección de U^* depende de las hipótesis que se estén considerando y en algunos casos el estadístico de prueba se aproxima asintóticamente a una F.

La esfericidad de Σ con respecto a U, implica esfericidad de Σ con respecto a alguna base del vector espacio columna de U, el cual podría tomarse como $U^* = U(U'U)^{-1/2}$ en la ecuación (4.27); entonces, teniendo en cuenta la Proposición 4.3, el estadístico de prueba F tiene una distribución F exacta con uc grados de libertad en el numerador y $u(n-m)$ grados de libertad en el denominador, bajo la hipótesis $H_0 : C\boldsymbol{B}U = D$ (ver mayores detalles en Singer & Andrade (1994)).

Hay que resaltar que la estructura de Σ es siempre esférica con respecto a la matriz de hipótesis U, si ésta matriz presenta una estructura unidimensional, para esos casos el estadístico de prueba F es siempre exacto, y para U no unidimensional, el estadístico de prueba F será aproximado. Mientras, cuando Σ no es esférica con respecto a la comparación definida por U, se tienen estadísticos exactos bajo un modelo MANOVA, donde no se impone ninguna restricción sobre Σ (Mardia et al. 2002, Cuadras 2010).

4.6 Aplicación

Se aplica esta nueva aproximación al estudio reportado por Frey et al. (1992), en donde se experimenta con el efecto de Zeolita A de sodio (SZA) en el crecimiento y fisiología de 60 caballos. La dieta alimenticia se aplico en forma aleatoria a los 60 caballos de forma individual, el tratamiento consta de cuatro niveles (0, 0.66, 1.32 y 2%); se midió la concentración de silicio en el plasma mediante muestras de sangre tomadas a las 0, 1, 3, 6 y 9 horas después de la ingestión a los ochenta y cuatro días de haber ingresado a la dieta. En este trabajo se hace una adaptación de este estudio sin tener en cuenta el tiempo 1, con el objetivo de tener tiempos igualmente espaciados. Sin embargo, se puede llevar a cabo el análisis teniendo en cuenta los cinco tiempos, pero para efectos de este trabajo se consideran cuatro tiempos.

Este caso fue estudiado también por Kshirsagar & Boyce (1995), quienes realizaron un análisis por medio de curvas de crecimiento. El objetivo es presentar un análisis donde se estudia la curva de crecimiento, para conocer la concentración de Zeolita A de sodio (SZA) que maximiza la concentración de silicio en el plasma y permite constatar que los caballos asimilan debidamente los nutrientes presentando así una buena fisiología reflejada a través del tiempo.

El tratamiento consta de cuatro niveles del factor cuantitativo SZA casi igualmente espaciados, cada nivel del tratamiento es observado sobre un mismo individuo a través del tiempo. Para ilustrar el método propuesto en este capítulo inicialmente se realizó un análisis exploratorio con los datos, donde se puede observar el comportamiento de los diferentes niveles del factor SZA con respecto a la concentración de silicio de los caballos en los diferentes tiempos

estudiados. Se observa que las concentraciones de silicio mas altas en los caballos corresponden a los niveles altos de SZA (1.32%,2%) y a mayor números de horas después de la ingestión. La Figura 4.1 sugiere el ajuste de un polinomio de grado uno o dos en la curva de crecimiento.

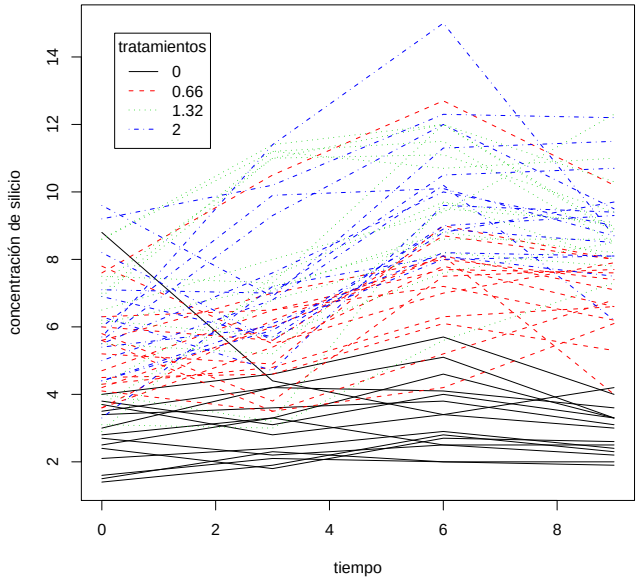


FIGURA 4.1: Concentración de silicio por tratamiento y tiempo

La Figura 4.2 muestra los perfiles medios de los tiempos. Se muestra la media de concentración de silicio para cada uno de los tiempos a través de los diferentes tratamientos. Es de esperarse que los perfiles medios en los diferentes tiempos no sean paralelos y por lo tanto, el crecimiento y cambio fisiológico que presentan los caballos se ve reflejado en el tiempo. Se puede ver que las respuestas medias más altas son para la concentración de SZA al 1.32% en cada uno de los tiempos.

La Figura 4.3 muestra que en el tiempo seis hay una mayor concentración media de silicio para cada uno de los tratamientos excepto con SZA de 0%.

Para mostrar la metodología propuesta se realizó un ajuste de un modelo de curvas de crecimiento a los datos usando la distancia de valor absoluto en la variable explicativa (tratamiento) y también en la variable tiempos, la variable explicada que se midió fue concentración de silicio en el plasma. Se utilizó como criterio de ajuste del modelo los criterios de información; Akaike (AIC) y Bayesiano (BIC), cuyos valores fueron de -551.88 y -517.42 respectivamente. Se determinaron las estructuras de correlación que mejor se adecuaban a los datos, para este caso se encontró apropiado ajustar un modelo de curva

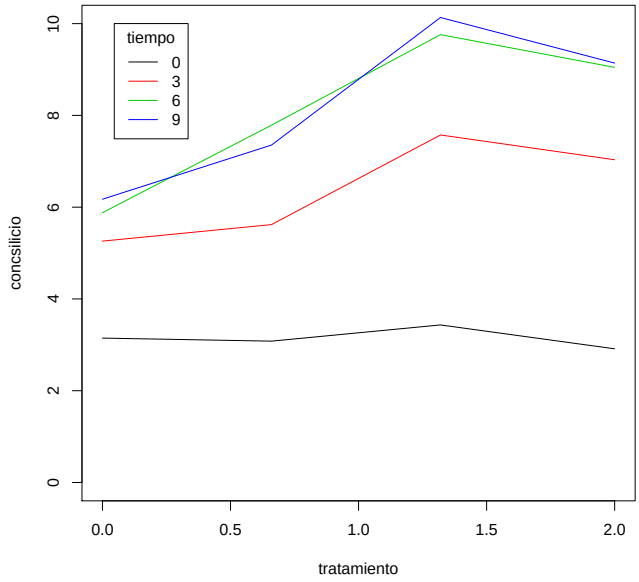


FIGURA 4.2: Perfiles medios de los tiempos a través de los tratamientos

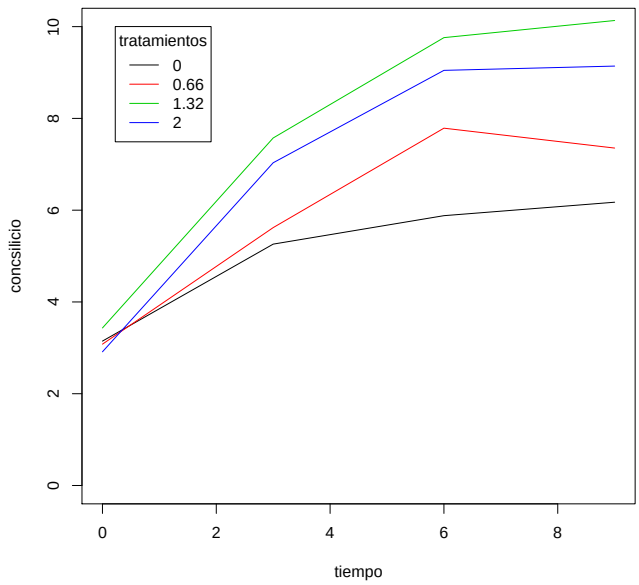


FIGURA 4.3: Perfiles medios de los tratamientos a través de los tiempos

de crecimiento con estructura compuesta simétrica (CompSymm), este coeficiente estimado usando distancias fue de 0.56. Se realizó también el ajuste del

modelo mediante la curva de crecimiento clásica, donde también se encontró que la estructura mas apropiada para la matriz de correlación es una Comp-Symm. Al utilizar el criterio de AIC para el ajuste del modelo se encontró un valor de -589.58 siendo mas bajo que cuando se usó el método de distancias y un valor de BIC$=-555.12$, con un coeficiente autorregresivo estimado de 0.55, se observa que los valores de AIC son muy cercanos. Por lo tanto, las predicciones resultan ser casi iguales con algunas pequeñas diferencias, por lo cual los resultados son prácticamente iguales por ambos métodos. Se encontró que el método basado en distancias genera predicciones similares comparado con la aproximación tradicional a las curvas de crecimiento.

El modelo ajustado después de hacer validación de supuestos sin usar distancias es un modelo con transformación en la variable respuesta, usando la potencia de 0.2 en Y, siendo β_0 el intercepto y t_j corresponde a los tiempos en la curva de crecimiento, v_j corresponde a la variable tratamientos con cuatro niveles y x_{jj} corresponde a las interacciones. Se utilizaron polinomios ortogonales para el análisis de los datos, la forma general de estos polinomios se encuentra definida en Hinkelmann & Kempthorne (1994), el polinomio de segundo orden ajustado esta dado por

$$\hat{y}_{ij}^{(0.2)} = \hat{\beta}_0 + \hat{\beta}_1 t_1 - \hat{\beta}_2 t_2^2 - \hat{\beta}_3 t_3^3 + \hat{\beta}_4 v_1 - \hat{\beta}_5 v_2^2 + \hat{\beta}_{14} x_{11} - \hat{\beta}_{15} x_{12}^2$$

donde $i = 1, 2, \ldots, 60$ y $j = 1, 2, 3, 4$. El modelo ajustado usando estimación REML para los parámetros es

$$\begin{aligned}\hat{y}_{ij}^{(0.2)} =& 1.4259 + 0.5769 t_1 - 0.2544 t_2^2 - 0.2420 t_3^3 + 1.5035 v_1 - 0.7338 v_2^2 \\ &+ 4.5150 x_{11} - 2.3829 x_{12}^2\end{aligned}$$

donde $i = 1, 2, \ldots, 60$ y $j = 1, 2, 3, 4$. El modelo ajustado después de validar los supuestos del modelo de normalidad y homocedasticidad usando distancias fue también con potencia de 0.2 en Y, siendo θ_0 el intercepto y g_j corresponde a las componentes de la matriz de coordenadas principales de F_g, X_j corresponde a las componentes de la matriz de coordenadas principales de F_x y X_{jj} son las componentes asociadas a las interacciones, de esta forma el modelo ajustado es el siguiente

$$\hat{y}_{ij}^{(0.2)} = \hat{\theta}_0 + \hat{\theta}_1 g_1 - \hat{\theta}_2 g_2 - \hat{\theta}_3 g_3 + \hat{\theta}_4 X_1 + \hat{\theta}_5 X_2 + \hat{\theta}_{14} X_{11} + \hat{\theta}_{24} X_{21}$$

donde $i = 1, 2, \ldots, 60$ y $j = 1, 2, 3, 4$. Ahora al considerar los valores obtenidos para la estimación de los parámetros por el método REML se obtiene

$$\begin{aligned}\hat{y}_{ij}^{(0.2)} =& 1.4259 + 0.0586 g_1 - 0.0464 g_2 - 0.0478 g_3 + 0.1561 X_1 + 0.1412 X_2 \\ &+ 0.0469 X_{11} + 0.0455 X_{21}\end{aligned}$$

donde $i = 1, 2, \ldots, 60$ y $j = 1, 2, 3, 4$

Por lo tanto, el polinomio ajustado haciendo distancias corresponde a un polinomio de segundo orden, el cual coincide con el ajustado haciendo el análisis por medio de curva de crecimiento clásica. La Figura 4.4 ilustra el ajuste de ambos modelos con respecto a los valores observados de concentración de silicio (los círculos). Las predicciones usando curvas de crecimiento con distancias (los triángulos) y el método clásico de curvas de crecimiento (las cruces) son casi iguales, difieren en muy poco. Con respecto al valor observado se aprecia un buen ajuste.

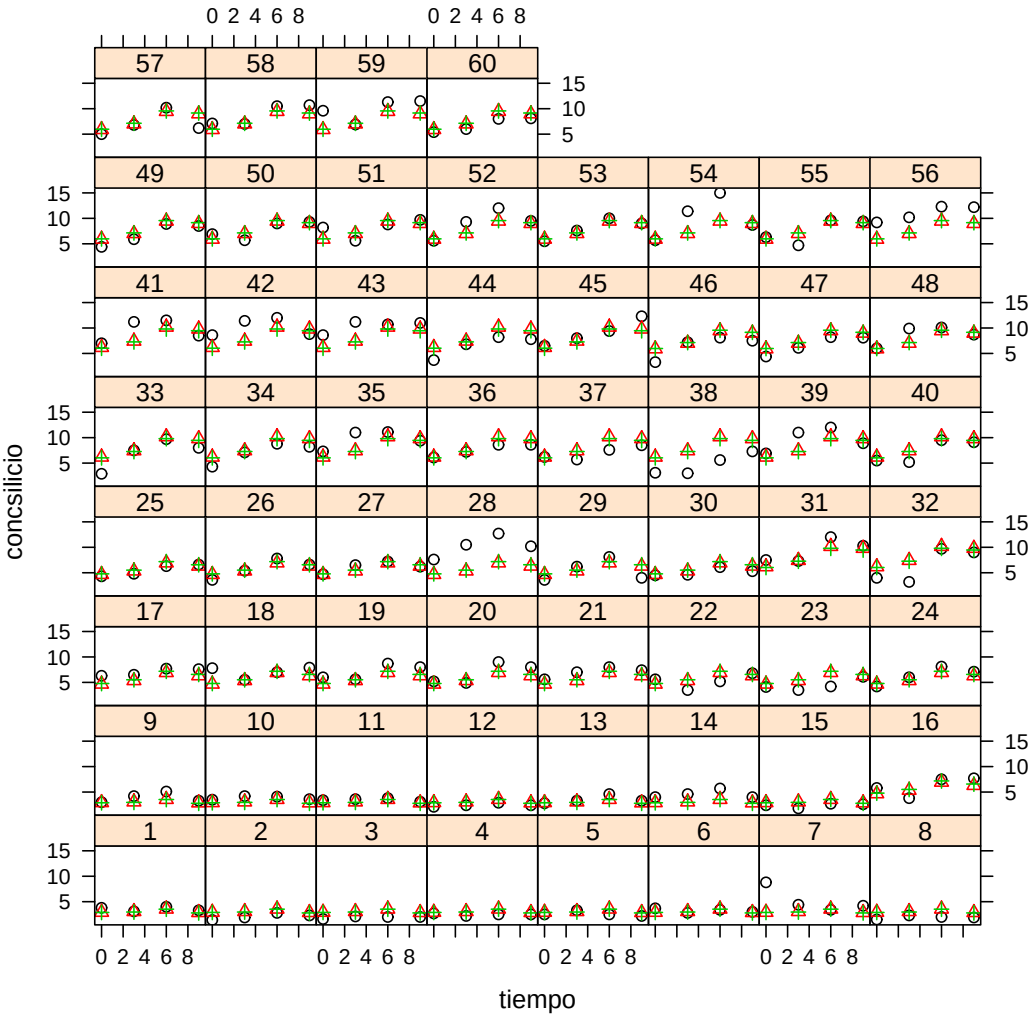


FIGURA 4.4: Concentración de silicio vs predicciones bajo ambas aproximaciones

Capítulo 5

Modelos lineales generalizados

Cuando no se asume la distribución normal para la variable respuesta, Liang & Zeger (1986a) proponen un conjunto de ecuaciones para la estimación de los parámetros del modelo, estas ecuaciones son conocidas en la literatura con el nombre de ecuaciones de estimación generalizada (EEG) y se presentan en la Sección 5.4. Liang et al. (1992) extienden el planteamiento de EEG, para permitir el modelamiento de los parámetros de correlación del modelo bajo un segundo sistema de ecuaciones, lo que se conoce como EEG2. Sutradhar (2003) presenta una completa revisión sobre los adelantos realizados a las EEG, así como una comparación entre los mismos.

Este capítulo describe el soporte teórico necesario para el desarrollo y comprensión de la metodología propuesta en el Capítulo 6, ésta teoría puede ser vista en su mayoría en un primer curso de modelos lineales generalizados y esta incluida en una gran cantidad de textos.

Algunos temas como la familia exponencial y las curvas de crecimiento, han sido fuertemente estudiados a lo largo de varias décadas y para su estudio, el lector especializado puede enfocarse en los libros que encuentra referenciados; sin embargo, otros temas como la quasiverosimilitud, los modelos lineales generalizados y las ecuaciones de estimación generalizada, han tenido un desarrollo acelerado durante las últimas 3 décadas, lo que obliga al lector interesado en el tema a la revisión de muchos de los artículos referenciados dentro del texto.

El capítulo se desarrolla en siete secciones: la Sección 5.1 presenta una revisión de la familia exponencial presentada por McCullagh & Nelder (1989) y Dobson (2002), la Sección 5.2 muestra los modelos lineales generalizados propuestos por Nelder (1972) y McCullagh & Nelder (1989), la Sección 5.3 presenta el método de quasioverosimiltud que fue presentado por Wedderburn (1974), y finalmente, la Sección 5.4 muestra las ecuaciones de estimación generalizadas propuestas por Liang & Zeger (1986a) y Liang et al. (1992).

5.1 Familia exponencial

Muchas distribuciones conocidas pueden ser reunidas en una familia de distribuciones llamada familia exponencial. Por ejemplo, pertenecen a esta familia las distribuciones normal, binomial, binomial negativa, gama, poisson, normal inversa, multinomial, beta, logarítmica, Rayleigh entre otras. Esta familia de distribuciones fue propuesta independientemente por Koopman, Pitman y Darmois a través del estudio de propiedades de suficiencia estadística. El concepto de familia exponencial fue introducido por Fisher, pero los modelos para distribuciones de la familia exponencial aparecen a finales del siglo XIX y fueron desarrollados por Maxwell, Boltzmann y Gibbs. La importancia de la familia exponencial cobra mayor relevancia en el área de los modelos de regresión, a partir del trabajo pionero de Nelder (1972) quienes definieron los modelos lineales generalizados. Estos modelos se popularizaron inicialmente en Inglaterra y posteriormente en Estados Unidos y Europa en la década de los 80.

Si una distribución pertenece a la familia exponencial uniparamétrica tiene una función de densidad que se puede escribir en la siguiente forma

$$f(y;\theta) = h(y)\exp\{\eta(\theta)t(y) - b(\theta)\} \tag{5.1}$$

donde las funciones $\eta(\theta)$, $b(\theta)$, $t(y)$ y $h(y)$ asumen valores en subconjuntos de la recta real. El soporte de la familia exponencial (5.1) es el conjunto $\{y : f(y;\theta) > 0\}$ y no puede depender de θ.

Dentro del campo de los modelos lineales es usual trabajar con una variación de la familia exponencial (5.1) en su forma canónica ($\eta(\theta) = \theta; t(y) = y$), la cual incluye un parámetro de dispersión $\phi > 0$, así

$$f(y;\theta,\phi) = \exp\left\{\frac{1}{\phi}[y\theta - b(\theta)] + c(y,\phi)\right\} \tag{5.2}$$

La Tabla 5.1 muestra como algunas de las distribuciones mas conocidas pueden ser escritas en la forma (5.2).

5.1.1 Momentos de la familia exponencial

La función generadora de momentos de la familia (5.2) esta dada por

$$M(t;\theta,\phi) = E(e^{ty}) = \exp\left\{\frac{1}{\phi}b(\phi t + \theta) - b(\theta)\right\}$$

La función generadora de cumulantes correspondiente a la familia (5.2) esta dada por

$$\varphi(t;\theta,\phi) = \log M(t;\theta,\phi) = \frac{1}{\phi}b(\phi t + \theta) - b(\theta) \tag{5.3}$$

Distribución	ϕ	θ	$b(\theta)$	$c(y,\phi)$	$V(\mu)$
Normal(μ,σ^2)	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+log(2\pi\sigma^2)\right]$	1
Poisson(μ)	1	$\log\mu$	e^θ	$-\log y!$	μ
Binomial(m,π)	1	$\log\left(\frac{\pi}{1-\pi}\right)$	$m\log(1+e^\theta)$	$\log\binom{m}{y}$	$\frac{1}{m}\mu(m-\mu)$
Bin. Neg.(μ,k)	1	$\log\left(\frac{\mu}{\mu+k}\right)$	$-k\log(1-e^\theta)$	$\log\left[\frac{\Gamma(k+y)}{\Gamma(k)y!}\right]$	$\mu\left(\frac{\mu}{k}+1\right)$
Gamma(μ,v)	v^{-1}	$-\frac{1}{\mu}$	$-\log(-\theta)$	$v\log(vy)-\log y-\log\Gamma(v)$	μ^2
Nor. Inv.(μ,σ^2)	σ^2	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{-1/2}$	$-\frac{1}{2}\left[\log(2\pi\sigma^2y^3)+\frac{1}{\sigma^2y}\right]$	μ^3

TABLA 5.1: Algunas distribuciones de la familia (5.2)

Derivando (5.3) sucesivamente con respecto a t se tiene

$$\varphi^{(r)}(t;\theta,\phi)=\phi^{(r-1)}b^{(r)}(\phi t+\theta)$$

en donde $b^{(r)}$ indica la r-ésima derivada de $b(.)$ con respecto a t. Para $t=0$ se obtiene el r-ésimo cumulante de la familia (5.2).

$$\kappa_r=\phi^{(r-1)}b^{(r)}(\theta)$$

De donde se observa que

$$\mu=E(Y)=\kappa_1=b'(\theta) \tag{5.4}$$

y

$$\sigma^2=Var(Y)=\kappa_2=\phi b''(\theta)=\phi V(\mu) \tag{5.5}$$

La expresión $\sigma^2=\phi V(\mu)$ muestra que para las distribuciones de la familia exponencial existe una relación entre la media y la varianza como se muestra en la Tabla 5.1, esta relación fue estudiada con mas detalle por Wedderburn (1974) dando origen al método de quasiverosimilitud.

5.2 Modelos lineales generalizados

La selección de modelos es una parte importante de toda investigación que considere el modelamiento estadístico como herramienta. Se requiere de un modelo que sea lo más simple posible y que describa bien los datos observados en áreas como agronomía, biología, economía, ingeniería, medicina, zootecnia, etc. Nelder (1972) mostraron que una serie de técnicas estadísticas comúnmente usadas por separado, pueden ser formuladas de una manera unificada como una clase de modelos de regresión. A esta teoría unificadora de modelamiento estadístico (una extensión de los modelos clásicos de regresión), le dieron el

nombre de modelos lineales generalizados (MLG). Estos modelos envuelven una variable respuesta univariada, variables explicativas y una muestra aleatoria de n observaciones independientes en donde

- La variable respuesta Y, *componente aleatorio* del modelo, tiene una distribución perteneciente a la familia (5.2) con valor esperado μ.
- Las variables explicativas $X = (X_1, X_2, \ldots, X_k)$ entran en forma de una estructura lineal acompañadas del vector de parámetros β, constituyendo el *componente sistemático* del modelo denotado por $\eta = X\beta$.
- Un enlace es hecho entre los componentes aleatorio y sistemático del modelo, a través de una función adecuada g llamada *función de enlace*, la cual conecta el predictor lineal con una función de la media μ mediante la expresión
$$g(\mu) = \eta.$$

En modelos lineales clásicos, la media y el predictor lineal son idénticos así que la función de enlace identidad es adecuada ya que η y μ pueden asumir cualquier valor en la recta real. Por ejemplo, cuando se trabaja con datos de conteo y la distribución de Poisson, se debe cumplir que $\mu > 0$, así que la identidad como función de enlace es menos atractiva ya que permite a η tomar valores negativos mientras que μ no puede hacerlo. En este caso, la función de enlace $\eta = g(\mu) = \log(\mu)$ con inversa $e^{\eta} = \mu$ resuelve el problema en mención.

En el caso de la distribución binomial, se debe cumplir que $0 < \mu < 1$ así que una función de enlace debe satisfacer la condición de asignar a cualquier número en el intervalo $(0, 1)$, un único valor de la recta real. Para este caso se consideran las tres principales funciones de enlace

1. Logit: $\eta = \log\left(\dfrac{\mu}{1-\mu}\right)$.
2. Probit: $\eta = \Phi^{-1}(\mu)$, donde $\Phi()$ es la función de distribución acumulativa normal estándar.
3. Complemento log-log: $\eta = \log\{-\log(1-\mu)\}$.

Aranda (1981) propuso una familia de funciones de enlace para analizar datos en forma de proporciones dada por

$$\eta = \log\left[\frac{(1-\pi)^{-\lambda} - 1}{\lambda}\right] \tag{5.6}$$

siendo λ una constante desconocida, que tiene como casos particulares el modelo logístico para $\lambda = 1$ y el complemento log-log para $\lambda \to 0$.

Otra familia importante de funciones de enlace propuesta por Box (1964) y utilizada principalmente para datos con media positiva, es la familia potencia que esta especificada por

$$\eta = \begin{cases} \dfrac{\mu^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \log \mu & \lambda = 0. \end{cases}$$

Cada una de las distribuciones de la Tabla 5.1 tiene una función de enlace especial para la cual existen estadísticos suficientes de la misma dimensión de β. Estos enlaces canónicos ocurren cuando $\theta = \eta$, donde θ es el parámetro canónico mostrado en la Tabla 5.1. Los enlaces canónicos para las distribuciones ya tratadas se muestran en la Tabla 5.2.

Distribución	Enlace (η)
Normal	μ
Poisson	$\log \mu$
Binomial	$\log\left(\dfrac{\pi}{1-\pi}\right)$
Gamma	μ^{-1}
Normal inversa	μ^{-2}

Tabla 5.2: Enlaces canónicos

5.2.1 Estimación de parámetros en un MLG

El método utilizado para la estimación de parámetros en un MLG es el de máxima verosimilitud. En el caso de la familia (5.2) la log-verosimilitud para una muestra de n observaciones esta dada por

$$\ell(\beta) = \sum_{i=1}^{n} \left[\frac{y_i - b(\theta_i)}{\phi} + c(y, \phi) \right] \tag{5.7}$$

Como es usual, el método busca maximizar la expresión anterior, al utilizar la regla de la cadena en el proceso de derivación se tiene

$$U_r = \frac{\partial \ell(\beta)}{\partial \beta_r} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\partial \ell(\beta)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r}$$

Lo que se reduce a

$$U_r = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ir} \tag{5.8}$$

El estimador de máxima verosimilitud $\hat{\beta}$ del vector de parámetros β se obtiene igualando U_r a cero para $r = 1, 2, \ldots, k$. En general las ecuaciones $U_r = 0$ son no lineales y tienen que ser resueltas por métodos numéricos como el de Fisher scoring, el cual suministra una expresión para la solución de $\hat{\beta}$ y utiliza dentro de su procedimiento la matriz de información de Fisher K y el vector U en una iteración m. En términos matemáticos la expresión es la siguiente

$$\beta^{(m+1)} = \beta^{(m)} + (K^{-1})^{(m)} U^{(m)} \tag{5.9}$$

Premultiplicando por $K^{(m)}$ se tiene

$$K^{(m)}\beta^{(m+1)} = K^{(m)}\beta^{(m)} + U^{(m)} \tag{5.10}$$

Los elementos de K se obtienen de la siguiente igualdad

$$\begin{aligned} k_{r,s} =& - E\left(\frac{\partial^2 \ell(\beta)}{\partial\beta_r \partial\beta_s}\right) = E\left(\frac{\partial \ell(\beta)}{\partial\beta_r}\frac{\partial \ell(\beta)}{\partial\beta_s}\right) = E(U_r U_s) \\ =& E\left(\frac{1}{\phi^2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{V(\mu_i)^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_{ir}x_{is}\right) \\ =& \frac{1}{\phi}\sum_{i=1}^{n} w_i x_{ir} x_{is}, \end{aligned}$$

con $w_i = \dfrac{1}{V(\mu_i)}\left(\dfrac{\partial\mu_i}{\partial\eta_i}\right)^2$.

Luego, la matriz de información de Fisher para β tiene la forma

$$K = \frac{1}{\phi} X^T W X, \tag{5.11}$$

donde $W = diag\{w_1, \ldots, w_n\}$. Nelder (1972) mostraron que la matriz de varianza-covarianza asintótica para $\hat{\beta}$ esta dada por

$$Var(\hat{\beta}) = K^{-1}.$$

Al expresar el vector U en forma matricial se tiene

$$U = \frac{1}{\phi} X^T W H (Y - \mu) \tag{5.12}$$

donde $H = diag\left\{\dfrac{\partial\eta_i}{\partial\mu_i}\right\}$, es una matriz diagonal formada por las primeras derivadas de la función de enlace. Reemplazando (5.11) y (5.12) en (5.10) se tiene

$$\begin{aligned} \frac{1}{\phi} X^T W^{(m)} X \beta^{(m+1)} &= \frac{1}{\phi} X^T W^{(m)} X \beta^{(m)} + \frac{1}{\phi} X^T W^{(m)} H^{(m)} (Y - \mu^{(m)}) \\ X^T W^{(m)} X \beta^{(m+1)} &= X^T W^{(m)} \left[\eta^{(m)} + H^{(m)} (Y - \mu^{(m)})\right] \\ X^T W^{(m)} X \beta^{(m+1)} &= X^T W^{(m)} Z^{(m)} \end{aligned}$$

donde se define una variable dependiente ajustada $z = \eta + H(y - \mu)$, con lo cual la solución para $\beta^{(m+1)}$ es

$$\beta^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} Z^{(m)} \tag{5.13}$$

La ecuación (5.13) es valida para cualquier MLG y muestra que la solución de las ecuaciones de MV equivale a calcular repetidamente una regresión lineal ponderada de una variable dependiente ajustada z sobre una matriz X con una matriz de peso W que se modifica en el proceso iterativo. Es importante enfatizar que la ecuación iterativa (5.13) no depende del parámetro de dispersión ϕ.

5.3 Quasiverosimilitud

El método de quasioverosimiltud (QV) fue presentado por Wedderburn (1974) y busca obtener conclusiones inferenciales, en experimentos donde no se especifica una distribución para la variable respuesta, se basa en el supuesto de que la varianza de Y se puede escribir como una función de su media μ. Tal supuesto permite la creación de una función de QV que permite la estimación de los parámetros presentes en el modelo, junto con sus errores estándar.

Para la explicación del método se considera la situación en la cual se tiene un conjunto de observaciones independientes y_i, $i = 1, \ldots, n$, provenientes de alguna distribución desconocida para la cual se presume que exista una relación entre la media y la varianza, es decir $E(Y_i) = \mu_i$ y $Var(Y_i) = \sigma^2 V(\mu_i)$. En este caso la función

$$U_i = \frac{y_i - \mu_i}{\sigma^2 V(\mu_i)}$$

cumple las propiedades características de una función score

$$E(U_i) = 0$$
$$Var(U_i) = -E\left(\frac{\partial U_i}{\partial \mu_i}\right) = \frac{1}{\sigma^2 V(\mu_i)}$$

por lo que la integral

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V(\mu_i)} dt$$

si existe, debe comportarse como una función de log-verosimilitud para μ bajo los supuestos establecidos en la varianza de las observaciones. Algunos ejemplos de funciones de QV son mostrados en la Tabla 5.3, varias de ellas corresponden a verdaderas funciones de log-verosimilitud para distribuciones conocidas, de hecho Wedderburn (1974) demostró que la función de QV coincide

con la función de log-verosimilitud cuando se especifica una distribución de la familia (5.2) para Y.

Se hace referencia a $Q(\mu_i, y_i)$ como la QV para μ basada en los datos y_i. Dado que las observaciones y_i se asumen independientes, la QV para una muestra de n observaciones es la suma de las contribuciones individuales

$$Q(\mu, Y) = \sum_{i=1}^{n} Q(\mu_i, y_i).$$

De manera análoga a los MLG se construye la función de desvíos en el caso de QV, para una sola observación esta función se define como

$$d_i = d(y_i; \mu_i) = -2\sigma^2 Q(\mu_i, y_i) = 2\int_{\mu_i}^{y_i} \frac{y_i - t}{V(\mu_i)} dt$$

y es utilizada como una medida de discrepancia entre el valor ajustado μ_i y el valor observado y_i.

Función de varianza $V(\mu)$	Quasiverosimilitud	Distribución	Restricciones
1	$-(y-\mu)^2/2$	Normal	–
μ	$y \log \mu - \mu$	Poisson	$\mu > 0$, $y \geq 0$
μ^2	$-y/\mu - \log \mu$	Gamma	$\mu > 0$, $y > 0$
μ^3	$-y/(2\mu^2) + 1/\mu$	Normal Inversa	$\mu > 0$, $y > 0$
μ^ι	$\mu^{-\iota}\left(\frac{uy}{1-\iota} - \frac{\mu^2}{2-\iota}\right)$	-	$\mu > 0$, $\iota \neq 0, 1, 2$
$\mu(1-\mu)$	$y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$	Binomial/m	$0 < \mu < 1$, $0 \leq y \leq 1$
$\mu^2(1-\mu)^2$	$(2y-1)\log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$	-	$0 < \mu < 1$, $0 < y < 1$
$\mu + \mu^2/k$	$y \log\left(\frac{\mu}{\mu+k}\right) + k \log\left(\frac{k}{\mu+k}\right)$	Bin.Negativa	$\mu > 0$, $y \geq 0$

Tabla 5.3: Algunas funciones de quasiverosimilitud.

5.3.1 Estimación de parámetros vía quasiverosimilitud

Las ecuaciones de estimación para los parámetros β derivadas por el método de QV, son obtenidas diferenciando la función $Q(\mu, y)$ y pueden ser escritas en la forma $U(\hat{\beta}) = 0$ donde

$$U(\beta) = D'\boldsymbol{V}^{-1}(Y - \mu)/\sigma^2$$

es llamada la función quasi-score. En esta expresión los elementos de $D_{n\times k}$ son $D_{rs} = \dfrac{\partial \mu_r}{\partial \beta_s}$ y corresponden a la derivada de μ con respecto a los parámetros β

y $\boldsymbol{V}$ es la matriz de varianza-covarianza de Y construida como

$$\boldsymbol{V} = diag\{\boldsymbol{V}(\mu_1), \ldots, \boldsymbol{V}(\mu_n)\}.$$

La matriz de varianza-covarianza de $U(\beta)$ la cual corresponde al valor esperado negativo de $\partial U(\beta)/\partial\beta$ es

$$i_\beta = D'\boldsymbol{V}^{-1}D/\sigma^2$$

esta matriz, juega el rol de la matriz de información de Fisher en la teoría de QV, por esta razón la matriz de varianza-covarianza asintótica de $\hat{\beta}$ es

$$Var(\hat{\beta}) \simeq i_\beta^{-1} = \sigma^2(D'\boldsymbol{V}^{-1}D)^{-1}$$

Así, al aplicar el método de Fisher scoring con la matriz i_β y el vector score se tiene la ecuación para la estimación de β en la iteración $m+1$.

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + (i_{\hat{\beta}}^{(m)})^{-1}U(\hat{\beta}^{(m)}) \tag{5.14}$$

La función de QV se construye buscando que se comporte como una función de log-verosimilitud para μ, sin embargo cuando se trata de σ^2 esto no ocurre, razón por la cual, el estimador para σ^2 se basa en el método de momentos y es

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\frac{y_i - \hat{\mu}_i}{\boldsymbol{V}(\hat{\mu}_i)} = \frac{X^2}{n-p}$$

donde X^2 es el estadístico de Pearson generalizado, para solucionar este inconveniente Nelder & Pregibon (1987) construyen la función de QV extendida (QVE), la cual permite comparar el ajuste de varias funciones de varianza a un mismo conjunto de datos; además de proporcionar estimaciones directas para σ^2. La función de desvíos generada por esta extensión es

$$D^+ = \sum_{i=1}^{n} d_i/\sigma^2 + \sum_{i=1}^{n}\log\{2\pi\sigma^2 V(y_i)\} \tag{5.15}$$

Otra extensión al concepto de QV, se conoce con el nombre de pseudoverosimilitud (SV) (Davidian & Carrol 1987, Breslow 1990) en la cual se utiliza como medida de discrepancia el estadístico de Pearson generalizado en lugar del estadístico de desvíos que utiliza la QVE. La función de desvíos generada por esta extensión es

$$D_p = \sum_{i=1}^{n} X_i^2/\sigma^2 + \sum_{i=1}^{n}\log\{2\pi\sigma^2\boldsymbol{V}(\mu_i)\} \tag{5.16}$$

Si se especifica la distribución normal, los tres métodos (QV,QVE y SV) tienen resultados equivalentes a los generados por el método de máxima verosimilitud, sin embargo cuando se especifica una distribución diferente los métodos proporcionan resultados diferentes, Davidian & Carrol (1988) encontraron resultados

asintóticos favorables a SV mientras que Nelder & Lee (1992) encontraron resultados favorables a QVE para muestras pequeñas via simulación.

Algunos trabajos han aportado grandes adelantos al tema de QV, entre ellos de destaca la QV con restricciones de Heyde & Morton (1993) y la QV con enfoque no paramétrico de Chiou & Müller (1998, 1999), en los cuales no se especifica la forma de la función de enlace, ni de la función de varianza para luego ser estimadas por métodos de suavizamiento no paramétrico.

Es posible extender el problema presentado en esta sección, cuando se dispone de un conjunto de datos longitudinales con distribución no normal. La extensión del método de QV para este caso, se conoce con el nombre de ecuaciones de estimación generalizada y se presenta en la siguiente sección.

5.4 Ecuaciones de estimación generalizada (EEG)

El análisis de datos longitudinales surge cuando una variable respuesta y un conjunto de covariables son medidas en varias ocasiones del tiempo sobre un conjunto de individuos.

Cuando la variable respuesta es de carácter continuo, muchas técnicas basadas principalmente en el uso de la distribución normal multivariante en el vector de respuestas, han sido desarrolladas para el análisis de los datos. Ware (1985) presenta una completa revisión de modelos lineales cuando se analiza datos longitudinales normales, sin embargo, cuando la variable respuesta es de carácter discreto, la falta de una distribución como la normal multivariante para el caso discreto dificulta el análisis de la información vía máxima verosimilitud.

Es por esta razón que Liang & Zeger (1986b) proponen una metodología derivada de la quasiverosimilitud, que permite obtener conclusiones inferenciales en análisis de datos longitudinales, sin necesidad de especificar la distribución conjunta del vector de respuestas, dicha metodología recibió el nombre de ecuaciones de estimación generalizada (EEG) y la aceptación que tuvo dentro de la comunidad académica, la llevo a tener un desarrollo bastante acelerado dentro de los 20 años siguientes a su creación. Zorn (2001) y Sutradhar (2003) publicaron dos trabajos bastante completos sobre el desarrollo de la metodología en el modelamiento de datos longitudinales discretos y continuos.

En lugar de especificar la distribución conjunta del vector de respuestas, se especifica la distribución marginal de Y en el tiempo t $(t = 1, 2, \ldots, m)$ como un miembro de la familia exponencial (5.2) y se modela la media de esta

distribución marginal bajo la expresión

$$g(\mu_{it}) = X_{it}\beta \tag{5.17}$$

La varianza de Y_i $(i = 1, 2, \ldots, n)$ se especifica como una función de la media, siguiendo el razonamiento descrito por la quasiverosimilitud y utilizando una matriz de correlación aproximada $R(\alpha)$, la cual es seleccionada por el investigador y esta completamente determinada por el vector α, que contiene los parámetros de correlación del modelo.

$$\Lambda_i = A_i^{1/2} R(\alpha) A_i^{1/2}/\phi \tag{5.18}$$

donde A_i es una matriz diagonal de tamaño $m \times m$ con $\boldsymbol{V}(\mu_{ij})$ como el $j-ésimo$ elemento de la diagonal.

La expresión (5.18) es exactamente igual a $Var(Y_i)$ para distribuciones de la familia exponencial solo en el caso que $R(\alpha)$ es igual a la matriz de correlación de Y_i. Sin embargo, los estimadores de los parámetros del modelo son consistentes aun cuando esta igualdad no se tenga, pero la función de enlace $g(.)$ este especificada correctamente. Diferentes tipos de estructuras para la matriz $R(\alpha)$ se presentan en la Subsección 5.4.1. Las ecuaciones de estimación generalizada tienen la forma

$$U(\beta) = \sum_{i=1}^{n} D_i^t \Lambda_i^{-1} (Y_i - \mu_i) = 0 \tag{5.19}$$

donde $(D_i)_{tp'} = \dfrac{\partial \mu_{it}}{\partial \beta'_p}$; $t = 1, \ldots, m$ y $p' = 1, \ldots, k$.

El término $U_i = D_i^t \Lambda_i^{-1}(Y_i - \mu_i)$ es similar a la función de quasiverosimilitud presentada por Wedderburn (1974) excepto porque la matriz Λ_i no solo depende de β sino también de α y de ϕ.

Liang & Zeger (1986b) combinaron el método iterativo de Fisher scoring para la estimación de β y el método de momentos para las estimaciones de α y ϕ, dando como resultado el estimador $\hat{\beta}_{\text{EEG1}}$ solución del sistema de ecuaciones (5.19). Además, encontraron que cuando el tamaño de la muestra aumenta, el estimador $\hat{\beta}_{\text{EEG1}}$ hallado como solución de la ecuación (5.19) tiene distribución normal multivariante de media β. Un estimador consistente para $Var(\hat{\beta}_{\text{EEG1}})$, válido aun cuando la matriz $R(\alpha)$ no se selecciona de forma correcta esta dado por

$$\hat{Var}(\hat{\beta}_{\text{EEG1}}) = n \left[H_1(\hat{\beta}_{\text{EEG1}})\right]^{-1} H_2(\hat{\beta}_{\text{EEG1}}) \left[H_1(\hat{\beta}_{\text{EEG1}})\right]^{-1} \tag{5.20}$$

con

$$H_1(\hat{\beta}_{\text{EEG1}}) = \left(\sum_{i=1}^{n} \hat{D}_i^t \hat{\Lambda}_i^{-1} \hat{D}_i \right)$$

$$H_2(\hat{\beta}_{\text{EEG1}}) = \left(\sum_{i=1}^{n} \hat{D}_i^t \hat{\Lambda}_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^t \hat{\Lambda}_i^{-1} \hat{D}_i \right)$$

El estimador (5.20) está construido bajo el mismo razonamiento utilizado por Crowder (2001), para modelos longitudinales con distribución normal multivariante y solo requiere que la función de enlace en (5.17) este bien especificada. Para el correcto uso de las ecuaciones de estimación generalizada, los datos faltantes existentes deben tener una estructura completamente aleatoria tal como lo define Rubin (1976), para determinar si este requerimiento se satisface se pueden utilizar las pruebas desarrolladas por Park & Lee (1997) y Little & Chen (1999) o utilizar la extensión para EEG cuando este supuesto no se cumple (Cho 1997).

5.4.1 Selección de la matriz de correlación

Una de las principales ventajas de utilizar EEG es la gran cantidad de opciones disponibles para especificar la estructura de correlación $R(\alpha)$. Fitzmarice et al. (1993) consideran cuatro estructuras básicas para la matriz $R(\alpha)$.

1. $R(\alpha) = I_m$. No se obtiene estimaciones para α ya que se asume que no hay correlación entre las observaciones para cada individuo.

2. $(R(\alpha))_{tt'} = \rho$, $(t, t' = 1, \ldots, m)$. Se asume que la correlación entre dos observaciones de un mismo individuo es igual sin importar los tiempos en que son tomadas. Bajo distribución normal, esta estructura es similar a la de un modelo de efectos aleatorios.

3. $(R(\alpha))_{tt'} = \rho^{|t-t'|}$,$(t, t' = 1, \ldots, m)$. Se asume que la correlación entre las observaciones de un mismo individuo es una función exponencial del rezago $|t-t'|$ tal como ocurre en los modelos autorregresivos de series de tiempo.

4. $(R(\alpha))_{tt'} = \alpha_{tt'}$,$(t, t' = 1, \ldots, m)$. Esta matriz de correlación no estructurada no asume ningún tipo de restricción entre sus elementos, razón por la cual se debe estimar $\frac{m(m-1)}{2}$ parámetros.

Esta matriz de correlación juega un papel importante dentro de la metodología de EEG. Sutradhar & Das(1999, 2000) analizaron el problema de pérdida de eficiencia de los estimadores cuando se selecciona una matriz equivocada.

Varios trabajos han sido dedicados a proponer estructuras para la matriz $R(\alpha)$ en situaciones especificas dependiendo de la variable respuesta. Lumley (1996) propone varias estructuras posibles para el modelamiento de datos ordinales longitudinales, mientras que Wang (1996) trata el problema de datos ordinales con sobredispersión en un contexto transversal. Para el manejo de datos longitudinales con sobredispersión, Thall & Vail (1990) proponen una

familia de estructuras para $R(\alpha)$ cuando se dispone de variables de tipo conteo medidas en forma longitudinal; esta idea es extendida por Sutradhar & Jowaheer (2002) quienes proponen el modelamiento conjunto de los parámetros de regresión del modelo y el parámetro de sobredispersión.

5.4.2 Modelamiento conjunto de media y varianza en EEG

Cuando se selecciona una estructura para la matriz de correlación $R(\alpha)$, se debe estimar una cantidad determinada de parámetros de correlación del modelo, por ejemplo, cuando no se asume alguna estructura sobre la matriz $R(\alpha)$, se debe estimar $\frac{m(m-1)}{2}$ parámetros asociados a la correlación del modelo y los contemplados en el vector α. Liang & Zeger (1986b) proponen estimar estos parámetros por el método de momentos utilizando los residuales de Pearson del modelo ajustado $r_{it} = (y_{it} - \mu_{it})/\phi V(\mu_{it})$, sin embargo, éste ha sido un tema de gran estudio por parte de muchos otros autores. Prentice (1988), Liang et al. (1992) y Prentice & Zhao (1991), han propuesto estimar estos parámetros bajo un segundo sistema de ecuaciones de estimación generalizada como el siguiente

$$U(\alpha) = \sum_{i=1}^{n} E_i^t \Omega_i^{-1} (Z_i - \sigma_i) = 0$$

donde $Z_i = (Z_{i12}, Z_{i13}, \ldots, Z_{i1m}, \ldots, Z_{im(m-1)})$ son las $\frac{m(m-1)}{2}$ correlaciones pareadas observadas en el $i - \acute{e}simo$ individuo y σ_i es un vector columna que contiene los valores esperados de estas correlaciones. Donde $(E_i)_{oq} = \dfrac{\partial \sigma_{io}}{\partial \alpha_q} \left(o, q = 1, \ldots, \frac{m(m-1)}{2}\right)$ y $\Omega_i = Var(Z_i)$ siguiendo el mismo razonamiento aplicado en (5.18).

Las ecuaciones (5.19) y (5.21) pueden ser parte de un sistema de ecuaciones mas amplio en el cual se asume que las estimaciones de α y β no están correlacionadas, esto conduce al sistema

$$U(\alpha, \beta) = \sum_{i=1}^{n} \begin{bmatrix} D_i^t & 0 \\ 0 & E_i^t \end{bmatrix} \begin{bmatrix} \Lambda_i & 0 \\ 0 & \Omega_i \end{bmatrix}^{-1} \begin{pmatrix} Y_i - \mu_i \\ Z_i - \sigma_i \end{pmatrix}$$

La anterior expresión, así como la ecuación (5.19) descrita por Liang & Zeger (1986b) son conocidas en la literatura como el método EEG1 y se caracteriza por el supuesto de no correlación entre Y_i y Z_i. En cambio Prentice & Zhao (1991) amplían este modelo para permitir la correlación entre Y_i y Z_i

lo cual lleva al sistema de ecuaciones

$$U(\alpha,\beta)=\sum_{i=1}^{n}\begin{bmatrix} D_i^t & 0 \\ F_i^t & E_i^t \end{bmatrix}\begin{bmatrix} \Lambda_i & Cov(Y_i,Z_i) \\ Cov(Y_i,Z_i) & \Omega_i \end{bmatrix}^{-1}\begin{pmatrix} Y_i-\mu_i \\ Z_i-\sigma_i \end{pmatrix} \quad (5.21)$$

donde $(F_i)_{oq}=\dfrac{\partial\alpha_{io}}{\partial\beta_q}$, $o=1,\ldots,\frac{m(m-1)}{2}$, $q=1,\ldots,m$. $(F_i)_{oq}$ se conoce como el método EEG2. Sutradhar (2003) presenta un análisis comparativo de los diferentes estimadores propuestos, analiza la existencia y la eficiencia de los estimadores generados por los métodos EGG1, EEG2 y algunas otras propuestas existentes en la literatura. Las conclusiones de este capítulo son las siguientes

- La formulación presentada por Liang & Zeger (1986b) (EEG1) presenta algunas falencias ocasionadas por la posible inexistencia de los estimadores de momentos para los parámetros de correlación del modelo. Además, que el estimador $\hat{\beta}_{\text{EEG1}}$ puede resultar menos eficiente que el estimador construido bajo el supuesto de independencia entre las observaciones de un mismo individuo $R(\alpha)=I$.

- La formulación del sistema (5.21) requiere el conocimiento de los momentos de orden superior (3, 4 y más) de la variable Y para la especificación del término $Cov(Y_i,Z_i)$. Algunos autores han encarado este problema imponiendo el supuesto de normalidad para la variable respuesta, lo que origina estimadores ineficientes.

Luego de esta comparación entre los métodos, Sutradhar (2003) propone utilizar el sistema (5.19) para estimar los parámetros de localización del modelo bajo la selección de una matriz de correlación por bandas como la siguiente

$$R(\alpha)=\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{m-1} \\ \rho_1 & 1 & \cdots & \rho_{m-2} \\ \vdots & \cdots & \ddots & \vdots \\ \rho_{m-1} & \rho_{m-2} & \cdots & 1 \end{bmatrix}$$

en la cual se tiene $m-1$ parámetros de correlación que se estiman por el método de momentos utilizando los residuales estandarizados del modelo.

$$\tilde{y}_{it}=\frac{y_{it}-\mu_{it}}{\{\phi V(\mu_{it})\}^{1/2}}$$

donde, $V(\mu_{it})$ es la función de varianza vista en (5.5)

5.4.3 Selección de modelos y bondad de ajuste en EEG

Cuando se utiliza el método de EEG, debe tenerse en cuenta que por construcción se trata de un método que no esta basado en el uso de la función de verosimilitud, esto genera que muchas herramientas diseñadas para la construcción de modelos dentro del campo de la verosimilitud, no puedan ser utilizadas en el contexto de EEG.

Uno de los criterios mas utilizados e implementado en diferentes paquetes de análisis de datos es el estadístico de Wald, con este criterio se puede hacer selección de variables dentro del predictor lineal. Pan (2001) presento un criterio que consiste en una modificación del conocido criterio de Akaike que lo hace aplicable a EEG. Este criterio puede ser utilizado para seleccionar la mejor estructura de la matriz $R(\alpha)$ de acuerdo a los datos obtenidos, o para seleccionar variables a tener en cuenta dentro del predictor lineal del modelo.

Capítulo 6

Análisis de datos longitudinales mediante distancias en modelos lineales generalizados

En este capítulo se presenta la metodología propuesta para el análisis de datos longitudinales con variables respuesta de distribución no normal utilizando distancias. Se inicia con la Sección 6.1, en donde se presenta el modelo propuesto basado en distancias, junto con las ideas principales que dan su origen. La Sección 6.2 esta dedicada a la estimación de parámetros y el contraste de hipótesis. Además, cuando se tiene una variable respuesta de tipo Poisson, se tiene relación entre la media y la varianza, por lo cual se presenta el problema de sobredispersión, el cual es estudiado en la Sección 6.3 y en la aplicación presentada en este capítulo. En la Sección 6.4 se presenta la metodología aplicada para la estimación de los parámetros involucrados en el modelo lineal generalizado basado en distancias, la cual se realiza aplicando la metodología de EEG; además, para juzgar las hipótesis se utiliza el estadístico de Wald. En la última Sección 6.5 se presenta una aplicación real con sobredispersión donde se pueden visualizar los resultados para la metodología propuesta.

6.1 Modelo propuesto

El modelo que se propone en este trabajo esta construido utilizando las ideas propuestas en la metodología de curvas de crecimiento y EEG. De esta forma se define el modelo bajo las siguientes características

1. Se considera una variable respuesta Y seguida a través del tiempo, cuya distribución marginal pertenezca a la familia (5.2).

2. Un predictor lineal que permita un modelamiento adecuado de los datos.

3. Una función de enlace que relacione la media de la distribución marginal de Y en el tiempo t con el predictor lineal anterior.

Estas tres características se unen en el siguiente modelo

$$g(\mu_{it}) = \boldsymbol{X_i}\xi\boldsymbol{G_t} \quad i = 1, \ldots, n \quad t = 1, \ldots, m \tag{6.1}$$

donde cada elemento de la matriz ξ corresponde a los parámetros asociados a las p' componentes de la matriz de coordenadas y a los q coeficientes de la curva de crecimiento en los tiempos. Este modelo utiliza un predictor lineal como el propuesto por Potthoff & Roy (1964) con el objetivo de incorporar polinomios en el tiempo dentro del modelamiento. Las matrices $\boldsymbol{X}$ y $\boldsymbol{G}$ están dadas en la ecuación (4.6), utilizadas dentro de la metodología de curvas de crecimiento del capítulo 4. Siendo $\boldsymbol{X_i}$ las coordenadas para el i ésimo individuo de la matriz de coordenadas $\boldsymbol{X}$ y $\boldsymbol{G_t}$ las coordenadas para el t ésimo tiempo en la matriz de coordenadas $\boldsymbol{G}$ asociadas a los tiempos.

Al agrupar la información para todos los individuos, el predictor lineal del modelo (6.1) puede ser escrito en forma matricial así

$$\eta = \boldsymbol{X}\xi\boldsymbol{G} \tag{6.2}$$

donde $\boldsymbol{X}$ es la matriz de coordenadas principales con p' componentes y $\boldsymbol{G}$ toma la forma descrita en (4.6) con q' componentes. El modelo (6.2) también puede ser expresado en la siguiente forma

$$g(\mu) = \eta = \boldsymbol{X}\xi\boldsymbol{G} = (\boldsymbol{X} \otimes \boldsymbol{G'})Vec(\xi') \tag{6.3}$$

Además, el modelo (6.1) también puede ser escrito para el i ésimo individuo como se presenta a continuación

$$g(\mu_{it}) = \boldsymbol{X_i}\xi\boldsymbol{G_t} = (\boldsymbol{X_i} \otimes \boldsymbol{G_t'})Vec(\xi') \tag{6.4}$$

esta expresión, facilita la estimación de los parámetros del modelo mediante la metodología de EEG y será discutida en la Sección 6.2.

6.2 Inferencia sobre el modelo propuesto

Las herramientas inferenciales utilizadas en este capítulo se enmarcan dentro del análisis de datos longitudinales.

Sea Y una variable cuya distribución puede ser escrita en la forma (5.2), con valor esperado μ ligado al predictor lineal $\eta = \boldsymbol{X}\beta$ mediante la expresión $g(\mu) = \eta$. La función de log-verosimilitud puede ser escrita como

$$\ell(\beta) = \sum_{i=1}^{n} \left[\frac{y_i - b(\theta_i)}{\phi} + c(y, \phi) \right] \tag{6.5}$$

luego, la función score utilizada para el proceso de estimación de parámetros tiene la forma

$$U_r = \frac{\partial \ell(\beta)}{\partial \beta_r} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\partial \ell(\beta)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r} \tag{6.6}$$

lo que se reduce a

$$U_r = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\boldsymbol{V}(\mu_i)} \frac{\partial \mu_i}{\partial \beta_r} \tag{6.7}$$

esta función puede ser escrita en forma matricial así

$$U(\beta) = \frac{1}{\phi} D' \boldsymbol{V}^{-1} (Y - \mu) \tag{6.8}$$

donde $D_{rp'} = \dfrac{\partial \mu_r}{\partial \beta_{p'}}$, $\boldsymbol{V}$ es una matriz diagonal con $\boldsymbol{V}(\mu_i)$ como el $i - ésimo$ elemento de la diagonal y $\boldsymbol{V}(.)$ es la función de varianza que se determina una vez se asume una distribución para Y según la Tabla 5.1.

Para encontrar el vector solución $\hat{\beta}$, basta con igualar $U(\beta)$ a cero y utilizar el método de Fisher-scoring tal como se describió en la Subsección 5.2.1.

Cuando la variable Y esta medida en los tiempos $t_1, t_2, \ldots, t_m$ y se especifica un modelo de la forma

$$g(\mu_{it}) = \eta_{it} = (\boldsymbol{X_i} \otimes \boldsymbol{G_t'}) Vec(\xi') \tag{6.9}$$

las ecuaciones de estimación (6.8) fueron extendidas por Liang & Zeger (1986b) con el fin de contemplar la correlación existente entre observaciones para un mismo individuo, esta estructura de correlación se incorpora al seleccionar una matriz de correlación $R(\alpha)$ en la expresión

$$\Lambda_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi \tag{6.10}$$

donde A_i tiene la siguiente forma

$$A_i = \begin{pmatrix} \boldsymbol{V}(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & \boldsymbol{V}(\mu_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{V}(\mu_{im}) \end{pmatrix} \tag{6.11}$$

Sutradhar (2003) recomienda seleccionar la matriz R bajo la siguiente estructura

$$R(\alpha) = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{m-1} \\ \rho_1 & 1 & \cdots & \rho_{m-2} \\ \vdots & \cdots & \ddots & \vdots \\ \rho_{m-1} & \rho_{m-2} & \cdots & 1 \end{pmatrix} \tag{6.12}$$

con estas consideraciones, las ecuaciones de estimación generalizada para los parámetros del vector $Vec(\xi')$ toman la forma

$$U(\xi') = \frac{1}{\phi} \sum_{i=1}^{n} D_i^t \Lambda_i^{-1} (Y_i - \mu_i) = 0 \tag{6.13}$$

donde

$$(D_i)_{tr} = \frac{\partial \mu_{it}}{\partial \xi_r}; i = 1, \ldots, n, \quad t = 1, \ldots, m \;\; y \quad r = 1, \ldots, p'q \tag{6.14}$$

Liang & Zeger (1986b) demostraron que el vector $Vec(\hat{\xi}')$, solución de las ecuaciones (6.13) sigue una distribución normal multivariante de media $Vec(\xi')$ y matriz de varianza-covarianza dada por

$$H_1^{-1}(\hat{\xi}') = \phi \left(\sum_{i=1}^{n} \hat{D}_i^t \hat{\Lambda}_i^{-1} \hat{D}_i \right)^{-1} \tag{6.15}$$

La consistencia de los estimadores dados en la expresión anterior depende de la correcta especificación de la función de enlace utilizada en (6.9), para remediar este problema, se suele utilizar como matriz de varianza-covarianza de $Vec(\hat{\xi}')$ la siguiente expresión

$$\hat{V}ar(Vec(\hat{\xi}')) = n \left[H_1(\hat{\xi}') \right]^{-1} H_2(\hat{\xi}') \left[H_1(\hat{\xi}') \right]^{-1} \tag{6.16}$$

con

$$H_2(\hat{\xi}') = \frac{1}{\phi} \left(\sum_{i=1}^{n} \hat{D}_i^t \hat{\Lambda}_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^t \hat{\Lambda}_i^{-1} \hat{D}_i \right) \tag{6.17}$$

Para la solución del sistema de ecuaciones (6.13) se utiliza el método de Fisher-scoring con la matriz $H_1^{-1}(\hat{\xi}')$ y el vector $U(\hat{\xi}')$. En la $m - ésima$ iteración del proceso la expresión es la siguiente

$$Vec(\hat{\xi}')^{(m+1)} = Vec(\hat{\xi}')^{(m)} + \left[H_1^{(m)}(\hat{\xi}') \right]^{-1} U^{(m)}(\hat{\xi}') \tag{6.18}$$

premultiplicando por la matriz $\left[H_1^{(m)}(\hat{\xi}') \right]$ se tiene

$$\left[H_1^{(m)}(\hat{\xi}') \right] Vec(\hat{\xi}')^{(m+1)} = \left[H_1^{(m)}(\hat{\xi}') \right] Vec(\hat{\xi}')^{(m)} + U^{(m)}(\hat{\xi}') \tag{6.19}$$

Al definir las matrices

$$D = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix}, \tilde{\Lambda} = diag\{\Lambda_1, \ldots, \Lambda_n\} \quad y \quad S = \begin{pmatrix} Y_1 - \mu_1 \\ Y_2 - \mu_2 \\ \vdots \\ Y_n - \mu_n \end{pmatrix} \tag{6.20}$$

se puede reescribir la matriz $H_1(\hat{\xi}')$ y el vector $U(\xi')$ así

$$H_1(\hat{\xi}') = \frac{1}{\phi} D' \tilde{\Lambda}^{-1} D \tag{6.21}$$

y

$$U(\xi') = \frac{1}{\phi} D' \tilde{\Lambda}^{-1} S \tag{6.22}$$

con lo cual, la expresión (6.19) toma la forma

$$\begin{aligned} &D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} D^{(m)} Vec(\hat{\xi}')^{(m+1)} = \\ &\qquad D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} D^{(m)} Vec(\hat{\xi}')^{(m)} + D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} S^{(m)} \end{aligned} \tag{6.23}$$

y la solución queda expresada como

$$\begin{aligned} &Vec(\hat{\xi}')^{(m+1)} = \\ &\qquad \left[D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} D^{(m)}\right]^{-1} D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} \left[D^{(m)} Vec(\hat{\xi}')^{(m)} + S^{(m)}\right] \end{aligned} \tag{6.24}$$

$$Vec(\hat{\xi}')^{(m+1)} = \left[D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} D^{(m)}\right]^{-1} D^{(m)'} \tilde{\Lambda}^{(m)^{-1}} Z^{*^{(m)}} \tag{6.25}$$

es decir, el vector solución puede ser calculado mediante una regresión iterativa entre $Z^* = [DVec(\xi') + S]$ y D con matriz de peso Λ^{-1}.

Otro punto de vital importancia dentro del proceso inferencial llevado a cabo sobre el modelo propuesto es el contraste de hipótesis, estas pueden ser planteadas en la forma de la hipótesis lineal general así

$$\mathbf{C}\xi\mathbf{U} = 0 \tag{6.26}$$

donde $\mathbf{C}_{c \times p'}$ y $\mathbf{U}_{q \times u}$ son matrices conocidas de rangos c $(\leq p')$ y u $(\leq q)$ respectivamente. Las matrices que definen las hipótesis principales, junto con su correspondiente interpretación, se muestran en la Tabla 6.1.

Para probar si las curvas de crecimiento generadas para dos tratamientos diferentes (A y B) difieren a lo largo del tiempo, las matrices $\mathbf{C}$ y $\mathbf{U}$ toman la siguiente forma

$$\mathbf{C} = \vec{X}_A - \vec{X}_B \qquad \mathbf{U} = I_q \tag{6.27}$$

donde $\vec{X}_A$ y $\vec{X}_B$ corresponden a dos filas de la matriz de coordenadas $\boldsymbol{X}$ cuya combinación de niveles de los K factores generan los tratamientos A y B respectivamente.

Todas estas hipótesis pueden ser contrastadas utilizando el estadístico de Wald. Para probar la hipótesis $LVec(\xi') = 0$ con $L = \mathbf{U}' \otimes \mathbf{C}$, este estadístico tiene la forma

$$W = (LVec(\hat{\xi}'))'(L\hat{V}ar(Vec(\hat{\xi}'))L')^{-1}(LVec(\hat{\xi}')) \tag{6.28}$$

donde $\hat{V}ar(Vec(\hat{\xi}'))$ se estima según la ecuación (6.16). Este estadístico sigue una distribución asintótica χ^2_l con $l = uc = rang(L)$.

H_0	Interpretación	$\mathbf{C}$	$\mathbf{U}$
$\xi = 0$	El modelo no es significativo (ajustando por los interceptos)	$\begin{pmatrix} 0 & 0_{1\times p'-1} \\ 0_{p'-1\times 1} & I_{p'-1} \end{pmatrix}$	$\begin{pmatrix} 0 & 0_{1\times q-1} \\ 0_{q-1\times 1} & I_{q-1} \end{pmatrix}$
$\xi^{(m)} = 0$	La $m-ésima$ columna de ξ es cero, indicando que el coeficiente de grado m no es importante en la curva de crecimiento.	$\mathbf{I}_{p'}$	$(0,\ldots,\underset{\substack{\downarrow \\ m-ésima}}{1},\ldots,0)'_{q\times 1}$
$\xi_l = 0$	La $l-ésima$ fila de ξ es cero, indicando que el parámetro (asociado a esta fila) no es significativo.	$(0,\ldots,\underset{\substack{\downarrow \\ l-ésima}}{1},\ldots,0)_{p'\times 1}$	$\mathbf{I}_q$
$\xi_l^{(m)} = 0$	El $l-ésima$ fila no ejerce influencia dentro del $m-ésimo$ grado de la curva.	$(0,\ldots,\underset{\substack{\downarrow \\ l-ésima}}{1},\ldots,0)_{p'\times 1}$	$(0,\ldots,\underset{\substack{\downarrow \\ m-ésima}}{1},\ldots,0)'_{q\times 1}$

Tabla 6.1: Resumen pruebas de hipótesis

6.3 Sobredispersión

Tradicionalmente, la distribución de Poisson ha sido usada para modelar datos relacionados con conteos y tiene la característica de suponer igualdad entre la media y la varianza; es decir, $E(Y) = Var(Y) = \mu > 0$, pero desafortunadamente este supuesto resulta ser muchas veces inconsistente en la práctica. Cuando ocurre esta situación se dice que existe sobredispersión.

Algunas de las causas de la sobredispersión son

1. Variabilidad del material experimental: Puede haber un componente de variabilidad de las unidades experimentales, que no haya sido considerado en el modelo.

2. Correlación entre respuestas individuales. Por ejemplo: En estudios del cáncer, con camadas de ratones se puede esperar que ratones de la misma camada tengan alguna correlación.

3. Muestreo de conglomerados.

4. Datos de nivel agregado: El proceso de agregación puede conducir a distribuciones compuestas.

5. Las condiciones experimentales pueden no estar perfectamente bajo control y por lo tanto, los parámetros desconocidos μ_i varían no solo con las covariables medidas sino también con factores "latentes" y no controlados.

En algunas circunstancias, las causas de la sobredispersión pueden surgir de la naturaleza del proceso de recolección de datos y tener como consecuencia, que los errores estándar obtenidos del modelo sean incorrectos o seriamente subestimados, conduciendo a una incorrecta evaluación de la significancia de los parámetros de regresión individual.

Una vez determinada la presencia de sobredispersión en un conjunto de datos, es necesario extender el modelo planteado, para tenerla en cuenta. Existen diferentes modelos específicos para la sobredispersión, los cuales se pueden categorizar en dos grandes grupos. Primero asumir alguna forma más general para la función de varianza, posiblemente incluyendo parámetros adicionales. Este tipo de modelos pueden no corresponder a una distribución específica de probabilidad para la respuesta, pero pueden ser vistos como extensiones útiles del modelo básico. Los parámetros de regresión pueden ser estimados usando métodos quasi-verosimilitud con algún procedimiento ad hoc para estimar cualquier parámetro adicional en la función de varianza. Un ejemplo

de esto es el uso de un factor de heterogeneidad en datos binomiales sobredispersos. También se puede asumir un modelo de dos etapas para la respuesta. Este tipo de modelos conducen a un modelo de probabilidad compuesta para la respuesta y, en principio, todos los parámetros pueden ser estimados usando máxima verosimilitud completa. Un ejemplo estándar es el uso de la distribución Binomial Negativa para datos "contables" sobredispersos.

Dentro de la segunda alternativa, la sobredispersión puede también ser explicada utilizando modelos de efectos aleatorios para cantidades, asumiendo una heterogeneidad natural entre las respuestas esperadas a través de las observaciones. Específicamente, en Diggle et al. (2002) se supone que

1. Condicional a μ_i, cada variable respuesta, Y_{ij} tiene distribución Poisson con media μ_i.
2. Las μ_i son variables aleatorias gamma independientes, con media μ y varianza $\phi\mu^2$.

Entonces, la distribución marginal de Y_{ij} es binomial negativa con

$$E(Y_{ij} = \mu) \quad y \quad Var(Y_{ij} = \mu + \phi\mu^2)$$

6.4 Metodología aplicada

En muchos estudios biomédicos, los datos contables longitudinales comprenden respuestas repetidas y un conjunto de covariables multidimensionales, para un gran número de individuos; pero cuando la variable respuesta en estos modelos está sujeta a sobredispersión, el parámetro de sobredispersión influye en las varianzas marginales y por lo tanto en la estimación eficiente de los parámetros de regresión. Es por esta razón que Jowaheer & Sutradhar (2002) desarrollaron una aproximación GEE, basados en una estructura de autocorrelación general para las observaciones repetidas, sobredispersas, discutiendo las propiedades asintóticas de los estimadores de los parámetros principales. Los desarrollos del artículo mencionado se presentan a continuación.

6.4.1 El modelo para los datos contables repetidos sobredispersos

Se supone que una respuesta escalar y_{it} y un vector p-dimensional de covariables x_{it} son observados para el conglomerado o unidad observacional $(i = 1, \ldots, n)$, en el tiempo t con $(t = 1, \ldots, m)$. También se supone que β es un vector $(p \times 1)$ de parámetros de regresión. Sea $y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{im})'$

el vector de respuestas $(m \times 1)$, para el i-ésimo individuo o conglomerado y sea $X_i = (x_{i1}, \ldots x_{it}, \ldots, x_{im})'$ la matriz $(m \times p)$ de covariables. En el análisis longitudinal, los componentes del vector y_i son respuestas repetidas, las cuales, probablemente están correlacionadas. En la práctica, esta estructura de correlación longitudinal es desconocida, lo que hace difícil la estimación de β. Bajo una distribución de familia exponencial para y_{it}, Liang & Zeger (1986b) usan una correlación de "trabajo" basados en la aproximación de ecuaciones de estimación generalizadas para estimar β. Para el caso de datos contables longitudinales con sobredispersión, en el artículo de Jowaheer & Sutradhar (2002) se supone que, condicional a un factor $\gamma_{it}^* = \gamma_i \psi_t$, los y_{it} tienen distribución Poisson dada por

$$f(y_{it}|\gamma_{it}^*) = \frac{1}{y_{it}!} \exp\left(y_{it}\eta_{it}^* - \exp(\eta_{it}^*)\right), \tag{6.29}$$

donde, γ_i son los efectos aleatorios propios del individuo o conglomerado y ψ_t son los efectos aleatorios propios del tiempo. Esta función tiene esperanza condicional, dada por $E(y_{it}|\gamma_{it}^*) = Var(y_{it}|\gamma_{it}^*) = \exp(\eta_{it}^*)$, donde $\eta_{it}^* = x_{it}'\beta + \log(\gamma_{it}^*)$.

Ahora, se supone que γ_{it}^* tiene distribución gamma con media 1 y varianza c, con densidad

$$g(\gamma_{it}^*) = \frac{c^{1/c}}{\Gamma(c^{-1})} \exp\left(-c^{-1}\gamma_{it}^*\right) {\gamma_{it}^*}^{c^{-1}-1} \tag{6.30}$$

Esto lleva a que, marginalmente, y_{it} tiene distribución binomial negativa dada por

$$f(y_{it}) = \frac{\Gamma(c^{-1} + y_{it})}{\Gamma(c^{-1})y_{it}!} \left(\frac{1}{1 + c\theta_{it}}\right)^{c^{-1}} \left(\frac{c\theta_{it}}{1 + c\theta_{it}}\right)^{y_{it}} \sim BinNeg((1/c), c\theta_{it}), \tag{6.31}$$

la cual acomoda la sobredispersión, indexada por c. Mas específicamente, bajo (6.31) se tiene

$$E(y_{it}) = \theta_{it} = \exp\left(x_{it}'\beta\right) \quad \text{y} \quad Var(y_{it}) = \theta_{it} + c\theta_{it}^2 \tag{6.32}$$

Como los $y_{it}'s$ son cantidades repetidas registradas sobre el tiempo $t = 1, \ldots, m$, es probable que, como en el caso de datos longitudinales con distribución Poisson, las m observaciones binomiales negativas repetidas, $y_{i1}, \ldots, y_{im}$, bajo el *i-ésimo* $(i = 1, \ldots, n)$ conglomerado están correlacionadas es por eso que el autor asume que las m respuestas binomiales negativas repeti-

das tienen estructura de correlación

$$C(\rho_1,\ldots,\rho_{m-1}) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{m-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m-1} & \rho_{m-2} & \rho_{m-3} & \cdots & 1 \end{bmatrix} \tag{6.33}$$

obteniendo la estructura de autocovarianza

$$\Sigma_i(\widetilde{\rho}) \equiv \Sigma_i(\rho_1,\ldots,\rho_{m-1}) = A_i^{1/2} C(\rho_1,\ldots,\rho_{m-1}) A_i^{1/2}, \tag{6.34}$$

con $A_i = diag\{Var(y_{it})\}$ y $Var(y_{it}) = \theta_{it} + c\theta_{it}^2$, como en (6.32).

6.4.2 Ecuaciones de estimación para los parámetros de regresión y sobredispersión

Sea $E(Y_i) = \theta_i$, el vector de medias del vector de respuestas *m-dimensional* y_i. Se sabe de (6.34) que $Var(Y_i) = \Sigma_i(\widetilde{\rho})$, es la matriz de covarianzas $(m \times m)$ dada por

$$\Sigma_i(\widetilde{\rho}) = A_i^{1/2} C(\widetilde{\rho}) A_i^{1/2}$$

donde $A_i = diag\{Var(Y_{it})\}$ y $Var(Y_{it}) = \theta_{it} + c\theta_{it}^2$. Obsérvese que la media θ_{it} es una función de β, donde $Var(Y_{it})$ es una función de ambos β y c.

Por consiguiente, las estimaciones de β y c se obtienen resolviendo las GEE conjuntas dadas por

$$n^{-1/2} \sum_{i=1}^{n} D_i' \widetilde{\Sigma}_i^{-1} (f_i - \mu_i) = 0, \tag{6.35}$$

donde $f_i = (f_{i1}', \ldots, f_{it}', \ldots, f_{im}')'$ y $\mu_i = (\mu_{i1}', \ldots, \mu_{it}', \ldots, \mu_{im}')'$ son vectores $(2m \times 1)$ con $f_{it} = (y_{it}, u_{it})'$, $u_{it} = y_{it}^2$, $\mu_{it} = (\theta_{it}, m_{it})'$, $\theta_{it} = E(y_{it})$ y $m_{it} = E(U_{it}) = \theta_{it} + (c+1)\theta_{it}^2$. En (6.35), Σ_i es la matriz de covarianzas $(2m \times 2m)$ de f_i y D_i es la matriz $(2m \times (p+1))$ de derivadas de μ_i con respecto a β y c asi

$$D_i = [\partial\mu_i/\partial\beta', \partial\mu_i/\partial c] = [D_{i1}', \ldots, D_{it}', \ldots, D_{im}']', \tag{6.36}$$

con

$$D_{it} = \begin{bmatrix} \partial\theta_{it}/\partial\beta' & 0 \\ \partial m_{it}/\partial\beta' & \partial m_{it}/\partial c \end{bmatrix}$$

donde $\partial\theta_{it}/\partial\beta' = \theta_{it}X'_{it}$, $\partial m_{it}/\partial\beta' = \theta_{it}X'_{it} + 2(c+1)\theta_{it}^2X'_{it}$ y $\partial m_{it}/\partial c = \theta_{it}^2$. Además la matriz de covarianzas de f_i se puede expresar como

$$\widetilde{\Sigma}_i = \begin{bmatrix} \widetilde{\Sigma}_{i1} & \widetilde{\Omega}_{i12} & \widetilde{\Omega}_{i13} & \cdots & \widetilde{\Omega}_{i1m} \\ & \widetilde{\Sigma}_{i2} & \widetilde{\Omega}_{i23} & \cdots & \widetilde{\Omega}_{i2m} \\ & & \widetilde{\Sigma}_{i3} & \cdots & \widetilde{\Omega}_{i3m} \\ & & & \ddots & \\ & & & & \widetilde{\Sigma}_{im} \end{bmatrix} \tag{6.37}$$

donde

$$\widetilde{\Sigma}_{it} = \begin{bmatrix} Var(Y_{it}) & Cov(Y_{it}, U_{it}) \\ & Var(U_{it}) \end{bmatrix}$$

es la matriz de covarianzas estructural (2×2) de f_{it} en un tiempo t dado, y para $t \neq w$, $t, w = a, \ldots, m$, $\widetilde{\Omega}_{itw}$ es la matriz de covarianzas longitudinal (2×2) de f_{it} y f_{iw}, donde f_{it} $(t = 1, \ldots, m)$ es el vector relacionado con la respuesta, definido como en (6.35).

Puesto que, para cada t, la respuesta y_{it} sigue una distribución Binomial Negativa, las matrices de la diagonal de la matriz $\widetilde{\Sigma}_i$ de (6.37), $\widetilde{\Sigma}_{it}$ se pueden obtener calculando los momentos de mayor orden necesarios de la ecuación (6.31). El cálculo exacto de las matrices fuera de la diagonal, sin embargo, no es posible como se requiere calcular los momentos de alto orden conjunto para las respuestas longitudinales cuya función de densidad de probabilidad conjunta es desconocida. Los tres elementos de la matriz $\widetilde{\Sigma}_{it}$, se pueden calcular directamente de la función generadora de momentos de y_{it}, dada por

$$M_{y_{it}}(s) = \left\{1 + c\theta_{it} - c\theta_{it}\exp(s)\right\}^{-1/c},$$

donde s es un parámetro real. De aquí se obtiene que

$$\begin{aligned} Var(Y_{it}) =& \theta_{it} + c\theta_{it}^2, \\ Cov(Y_{it}, U_{it}) =& \theta_{it}\left\{1 + (2+3c)\theta_{it} + 2c(1+c)\theta_{it}^2\right\}, \\ Var(U_{it}) =& \theta_{it} + (6+7c)\theta_{it}^2 + (4+16c+12c^2)\theta_{it}^3 + (4c+10c^2+6c^3)\theta_{it}^4. \end{aligned}$$

Ahora, el cálculo de los elementos fuera de la diagonal de $\widetilde{\Sigma}_i$ no puede obtenerse directamente de la función generadora de momentos, como en el caso anterior, dado que la función de densidad de probabilidad es desconocida. Sin embargo, como la elección de la matriz ponderada (o peso), en general, no afecta la consistencia de los estimadores de β y c, estas matrices

ponderadas, $\widetilde{\Omega}_{itw}$, denominadas *matrices de covarianza longitudinales de trabajo o empíricas*, serán calculadas pretendiendo que $Var(Y_i) = \Sigma_i$ en (6.34) es la matriz de covarianzas de un vector normal *m-dimensional* Y_i (Prentice & Zhao 1991). Por lo tanto, las fórmulas para las matrices de correlación longitudinal de trabajo, para $t \neq w$, son

$$\widetilde{\Omega}_{itw} = \begin{bmatrix} \widetilde{\Omega}_{itw,11} & \widetilde{\Omega}_{itw,12} \\ & \widetilde{\Omega}_{itw,22} \end{bmatrix} \tag{6.38}$$

asumiendo covariables independientes del tiempo se obtiene

$$\begin{aligned}
\widetilde{\Omega}_{itw,11} =& Cov(Y_{it}, Y_{iw}) = \rho_{|t-w|}(\theta_{i1} + c\theta_{i1}^2), \\
\widetilde{\Omega}_{itw,12} =& Cov(Y_{it}, Y_{iw}^2) = (2\rho_{|t-w|} + 2c\rho_{|t-w|}\theta_{i1})\theta_{i1}^2, \\
\widetilde{\Omega}_{itw,22} =& Cov(Y_{it}^2, Y_{iw}^2) = 2\rho_{|t-w|}(\theta_{i1} + c\theta_{i1}^2)\left\{\rho_{|t-w|}(\theta_{i1} + c\theta_{i1}^2) + \theta_{i1}^2\right\}.
\end{aligned}$$

6.4.3 Ecuaciones iterativas de β y c

La estimación de los parámetros se hace en dos pasos. Primero, para unos parámetros pre establecidos, de autocorrelación ρ_l $(l = 1, \ldots, m-1)$, el vector de parámetros de regresión β y el parámetro de sobredispersión c son estimados simultáneamente de las GEE propuestas aquí. Después, los valores estimados de β y c son usados en la ecuación de momentos, para estimar los parámetros de autocorrrelación. Estos pasos son iterados hasta obtener convergencia.

Resolviendo las GEE de (6.35), por métodos iterativos como el de "Newton-Raphson", se encuentra que dados los valores de $\widehat{\beta}_{GEE}(r)$ y $c_{GEE}(r)$ en la *i-ésima* iteración, las estimaciones en la $(r+1)$-*ésima* iteración, de β y c se obtienen como

$$\begin{bmatrix} \widehat{\beta}_{GEE}(r+1) \\ \widehat{c}_{GEE}(r+1) \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_{GEE}(r) \\ \widehat{c}_{GEE}(r) \end{bmatrix} + \left[\sum_{i=1}^{n} D_i' \widetilde{\Sigma}_i^{-1} D_i\right]_r^{-1} \left[\sum_{i=1}^{n} D_i' \widetilde{\Sigma}_i^{-1} (f_i - \mu_i)\right]_r$$

donde $[\quad]_r$ denota el hecho que la expresión dentro del paréntesis cuadrado es evaluada en $\beta = \widehat{\beta}_{GEE}(r)$ y $c = \widehat{c}_{GEE}(r)$. Bajo pocas condiciones de regularidad, se puede mostrar que $n^{1/2}[(\widehat{\beta}_{GEE} - \beta)', (\widehat{c}_{GEE} - c)]'$ tiene una distribución normal asintótica, cuando $n \to \infty$, con media cero y matriz de covarianzas V_{GEE}, la cual puede ser estimada consistentemente por

$$\widehat{V}_{GEE} = n \left(\sum_{i=1}^{n} D_i' \widetilde{\Sigma}_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i' \widetilde{\Sigma}_i^{-1} (f_i - \mu_i)(f_i - \mu_i)' \widetilde{\Sigma}_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i' \widetilde{\Sigma}_i^{-1} D_i\right)^{-1}$$

donde μ_i y f_i son como en (6.35).

6.4.4 Estimación de los parámetros de correlación longitudinal

Para $\beta = \widehat{\beta}_{GEE}$ y $c = \widehat{c}_{GEE}$ dados, el parámetro de correlación longitudinal $\widehat{\rho}_l$ puede ser estimado consistentemente por

$$\widehat{\rho}_l = \frac{\sum_{i=1}^{n}\sum_{t=1}^{m-1} y_{it}^* y_{i(t+l)}^*/(m-l)}{\sum_{i=1}^{n}\sum_{t=1}^{m} {y_{it}^*}^2/m} \qquad l = 1, \ldots, m-1 \tag{6.39}$$

donde $y_{it}^* = (y_{it} - \theta_{it})/(\theta_{it} + c\theta_{it}^2)^{1/2}$.

Acá el estimador de momentos $\widehat{\rho}_l$ dado en (6.39) es una modificación de la fórmula estimando

$$\widehat{\rho}_l^* = \sum_{i=1}^{n}\sum_{t=1}^{m-1} y_{it}^* y_{i(t+l)}^*/n(m-l) \quad (l = 1, \ldots, m-1) \tag{6.40}$$

Sugerido en Sutradhar & Das (1999) y Liang & Zeger (1986b). La principal razón para usar $\widehat{\rho}_l$ en (6.39) en vez de $\widehat{\rho}_l^*$ en (6.40) es que, aunque para β y c conocidos la autocovarianza $E(y_{it}^* y_{i(t+l)}^*)$ y autocorrelación $E(y_{it}^* y_{i(t+l)}^*)/E^2(y_{it}^*)$ de los residuales estandarizados y_{it}^* y $y_{i(t+l)}^*$ son los mismos, estos son en general diferentes para el caso cuando β y c son reemplazados por sus estimadores.

En la siguiente sección se presenta una aplicación real con sobredispersión donde se pueden visualizar los resultados para la metodología propuesta.

6.5 Aplicación

La información aquí analizada hace parte de un proyecto de conservación, cría y manejo de abejas silvestres, realizado en una zona rural de Acacías (Meta, Colombia) y fue tomada por Rodríguez (2005) con el fin de estudiar la actividad de forrajeo de polen (entiéndase forrajeo como el proceso de recolección, en este caso de polen, desarrollado por abejas de la especie Melipona Fasciata) en cinco nidos diferentes de abejas obreras de Melipona Fasciata y determinar así el efecto que podían tener la temperatura y la humedad relativa sobre esta actividad. Para cumplir con estos objetivos, el investigador registró, en tres días diferentes, el número de entradas de abejas con cargas de polen, desde las 6:00 a.m. hasta las 9:00 a.m., (horario de mayor actividad de forrajeo de polen) cada hora, durante 10 minutos. En cada uno de los nidos de estudio; paralelamente, se tomaron mediciones de temperatura y humedad relativa, cada hora de dichos nidos. Este proceso fue realizado durante los meses de enero y febrero de 2004, cuando el nivel de precipitaciones es menor (época

seca) y luego en los meses de abril y mayo del mismo año, considerado época lluviosa dado el aumento considerable de precipitaciones (lluvias).

Se colocaron cinco nidos en época lluviosa y cinco en época seca. Teniendo en cuenta que los nidos uno a tres, son aquellos ubicados en cajas racionales (grupo 1) y los nidos cuatro y cinco son los que se encuentran es estado natural (grupo 2); al hacer una comparación entre estos dos grupos, se observa un comportamiento muy atípico en el nido 3 (removido de su ambiente natural) frente a los otros nidos. El investigador atribuye esta situación a que, mientras los nidos uno, dos y tres del grupo 1 lograron acomodarse fácilmente a las cajas donde fueron trasladados, este nido en particular tuvo problemas de adaptación y fue atacado por moscas, ocasionando una fuerte disminución de su colonia afectando el proceso de recolección de polen. Al realizar un análisis descriptivo se encuentra que el nido de mayor recolección de polen en época lluviosa fue el nido dos, mientras que los nidos uno, cuatro y cinco, presentaron una actividad más baja, pero muy similar.

En época seca, se observa que aunque el nido dos continúa con un nivel alto de forrajeo de polen, comparada con los nidos uno y tres, éste es considerablemente menor a la del nido cuatro y cinco que presentan una mayor actividad. El promedio de temperatura de los nidos, en época lluviosa está alrededor de 23 grados, mientras en época seca es de 26 grados centígrados. Los promedios más bajos de temperatura se presentan en las dos primeras horas de la mañana (entre las 6:00 a.m. y 7:00 a.m.). Al comparar estas temperaturas con el número de entradas con polen, para estas horas de la mañana, se puede pensar en la temperatura como un factor influyente en el proceso de recolección de polen, siendo las horas del día con menor temperatura, las de mayor actividad.

En cuanto a la humedad relativa, el porcentaje promedio, en época lluviosa, está alrededor del 94 por ciento y 61 por ciento en época seca. Contrario a la temperatura, el mayor número de entradas con polen ocurren con altos porcentajes de humedad. De este modo se puede pensar en la humedad relativa como otro factor influyente en la actividad de forrajeo de polen.

El gráfico 6.1 permite ver la distribución de entradas con polen de las abejas a los nidos, se observan algunos valores atípicos en la cola derecha (80, 83, 100 y 141), cuyos valores corresponden al máximo número de entradas con polen. La mayoría de observaciones varían en el rango de 0 a 38, siendo la mayor frecuencia de entradas con polen en valores bajos, tal como lo muestra el gráfico y en las primeras horas de la mañana.

Continuando con el análisis, inicialmente se realiza un análisis basado en distancias para aplicar la metodología propuesta. Se busca que modelo lineal generalizado ajustar para la variable respuesta: número de entradas con polen y se realiza distancias entre las variables explicativas estudiadas. Ya que los

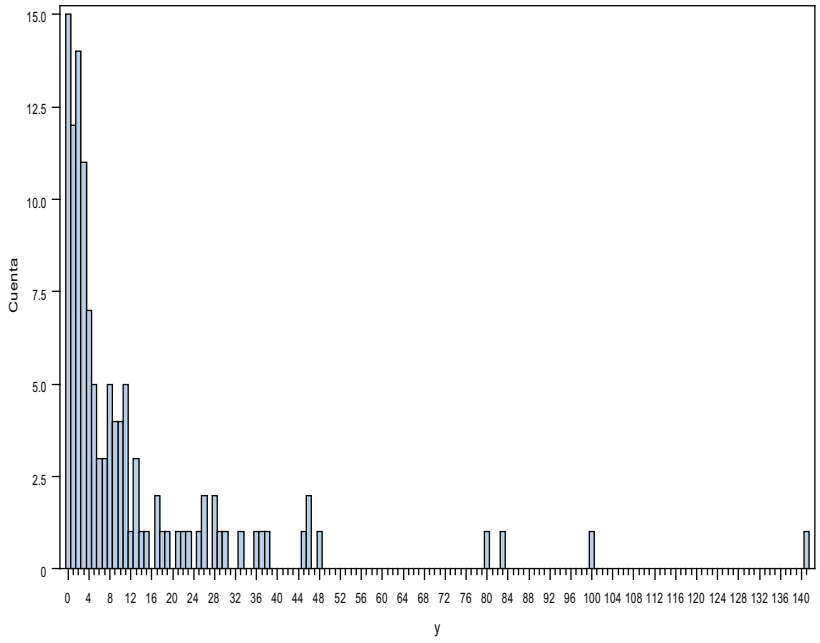


FIGURA 6.1: Distribución de entradas con polen

datos son mixtos se utilizó la distancia de Gower en las variables explicativas. Para la variable tiempos que fueron cuatro se utilizo la distancia valor absoluto y se procedió a construir la matriz de datos presentada en la ecuación (6.4), después de construir las matrices de coordenadas principales.

Se probaron varios modelos en el software estadístico R para el análisis versión 2.15 (R Development Core Team 2012), haciendo uso del paquete Geepack presentado en Halekoh & Yan (2006) que permite a través de GEE estimar los parámetros del modelo y modelar la estructura de autocorrelación. De acuerdo al tipo de variable respuesta se pensó inicialmente en un modelo con respuesta Poisson, pero después de realizar el análisis se encontró que no es bueno el ajuste por este modelo y que al haber sobredispersión en los datos se debía buscar una función de varianza que controlará esto. Entonces, se procedió a probar un modelo Binomial Negativo debido a la naturaleza de la variable respuesta.

Para el método clásico con los datos del experimento se procede a plantear un modelo que permita identificar si las variables independientes: época (lluviosa o seca), día, habitat de las abejas codificadas como grupos (cajas o árboles vivos), temperatura y humedad relativa, intervienen en el proceso de recolección de polen para los nidos estudiados. En las siguientes secciones se presentan ambos análisis y los modelos ajustados bajo la metodología propues-

ta usando DB y el método clásico de MLG.

6.5.1 Descripción de las variables y construcción del modelo

Dado que en este estudio se realizaron conteos en cuatro ocasiones, es decir las mediciones realizadas desde las 6:00 a.m. hasta las 9:00 a.m., es necesario tener en cuenta, en la construcción del modelo la correlación que pueda existir entre estas observaciones para cada uno de los nidos (unidades experimentales). De igual manera, como en este caso la variable de interés es discreta, es claro que un modelo lineal clásico no es el apropiado para resolver esta situación. Dentro del proceso de construcción del modelo para los datos se debe tener las siguientes consideraciones

1. Componente aleatorio: En este caso la variable respuesta está relacionada con conteos; por lo tanto se asume que su distribución es Poisson.

2. Componente sistemático: Las covariables en este caso son

$$x_{ij1} = \begin{cases} 1, & \text{en época lluviosa} \\ 0, & \text{en época seca} \end{cases} \qquad x_{ij2} = \begin{cases} 1, & \text{abejas en cajas racionales} \\ 0, & \text{abejas en árboles vivos} \end{cases}$$

   x_{ij3} = los tres días de observación, x_{ij4} = temperatura, and x_{ij5} = humedad relativa.

3. Función de enlace: La función que relaciona la media de la variable respuesta entradas con polen con las covariables es la función logaritmo.

4. Relación entre media y varianza: Para una variable distribuida Poisson, se sabe que la media es igual a la varianza; esto es $\mu = E(Y) = Var(Y)$, en este caso el parámetro de dispersión es $\phi = 1$.

5. Estructura de correlación: Para determinar la estructura de correlación se tuvo en cuenta lo explicado en la sección de las ecuaciones de estimación generalizadas (GEE) y las características de las diferentes estructuras de correlación que se pueden asumir, según las características propias de los datos. Dado que se tienen mediciones repetidas en el tiempo, es razonable pensar en una disminución en su asociación, a medida que la distancia entre dos puntos del tiempo incrementa; por lo tanto, un modelo autorregresivo de primer orden AR(1), es la estructura de correlación asumida para este modelo.

Teniendo en cuenta estas consideraciones y los objetivos del investigador, se decide emplear inicialmente un modelo marginal para datos longitudinales con distribución Poisson. La forma del modelo propuesto es

$Log(\mu_i) = Log(E(Y_{ij})) = x'_{ij}\beta = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_5 x_{ij5}$; para $i = 1, 2, \cdots, 30$ colonias diferentes de abejas y $j = 1, \cdots, 4$ tiempos de observación. Al realizar el ajuste del modelo se encuentra que no es muy apropiado pues el estadístico de deviance esta lejano de los grados de libertad indicando que no es bueno el ajuste de un modelo Poisson. Para encontrar el ajuste de un modelo apropiado tengamos en cuenta algunas consideraciones importantes en las siguientes secciones.

6.5.2 Bondad de ajuste del modelo propuesto

Cuando se emplean los modelos lineales generalizados con un parámetro de escala conocido, como es el caso de las distribuciones Binomial y Poisson (donde $\phi = 1$), una forma de evaluar el buen ajuste de un modelo es esperando que la varianza residual se aproxime a los grados de libertad de los residuales, pero cuando se emplean distribuciones como la Poisson para el modelamiento de datos puede suceder que la varianza residual se mayor o igual a los grados de libertad de los residuales. Indicando posibles problemas de sobredispersión, que deben tenerse en cuenta en el modelo definitivo. La sobredispersión es un fenómeno presente comúnmente cuando distribuciones como la Binomial, la multinomial o la Poisson son empleadas para el modelamiento de datos y significa que las varianzas reales de los datos exceden la varianza "nominal", asumida por la distribución empleada. La razón entre varianza y media (varianza/media) es alta, incumpliendo el supuesto de igualdad mencionado. Es evidente que esta sobredispersión se presenta en las primeras horas de la mañana.

Otra forma de evaluar la bondad de ajuste del modelo propuesto y detectar problemas de sobredispersión es por medio de los estadísticos: chi-cuadrado de Pearson y Deviance. En general, si el modelo seleccionado para describir el comportamiento de los datos es adecuado, estos dos estadísticos se distribuyen asintóticamente, como una χ^2 con $n - p$ grados de libertad, donde n es el número total de individuos o unidades experimentales y p es el número de parámetros ajustados. Entonces si el modelo ajustado es correcto los estadísticos, chi-cuadrado de Pearson y Deviance, serán iguales o cercanos a sus grados de libertad $n - p$, o lo que es lo mismo, el cociente de cualquiera de estos dos valores, con sus grados de libertad, será cercano a uno. Si esto no ocurre, se puede dudar acerca de la validez del modelo ajustado.

Para Davidian (2005), una forma de solucionar este problema de sobredispersión en un modelo Poisson es siendo un poco más flexible en el mode-

lamiento de la varianza, lo cual se puede lograr introduciendo un parámetro de dispersión ϕ en la relación entre la media y la varianza, es decir

$$Var(Y) = \phi$$

El procedimiento GENMOD de SAS, permite aplicar esta alternativa, estimando este parámetro de dispersión ϕ como una razón entre cualquiera de los valores de (Chi-cuadrado o Deviance) y sus grados de libertad, para luego estimar los parámetros normalmente. Haciendo uso del ajuste de un modelo Poisson con función de enlace log y estructura de autocorrelación AR(1), no se puede asegurar que este modelo describa bien el comportamiento de los datos, ya que, para los dos estadísticos (deviance y chicuadrado), se observan valores muy diferentes a los grados de libertad; de igual manera, el cociente entre cualquiera de estos dos estadísticos y sus grados de libertad es mucho mayor a uno, evidenciando la presencia de sobredispersión y confirmando lo observado al comparar la media y la varianza de los datos.

De este modo, el parámetro de sobredispersión con el método MLG con DB (considerando el intercepto y las tres primeras componentes g_1, g_2 y C_1 de la matriz de datos con distancias), estimado con el estadístico de deviance es

$$\sqrt{\phi} = \sqrt{\frac{Deviance}{D.F.}} = \sqrt{\frac{1290}{116}} = 3.3348$$

y con el método MLG clásico (considerando todas las variables explicativas de la matriz de datos) se tiene

$$\sqrt{\phi} = \sqrt{\frac{Deviance}{D.F.}} = \sqrt{\frac{1568}{113}} = 3.7251.$$

Además, bajo el método MLG clásico (considerando solo las variables significativas al 5% que fueron temperatura y humedad) se obtiene

$$\sqrt{\phi} = \sqrt{\frac{Deviance}{D.F.}} = \sqrt{\frac{1657}{117}} = 3.7633.$$

Se observa que hay sobredispersión en los modelos bajo ambos métodos, por lo tanto la distribución Poisson no resulta apropiada y se procede a probar con un modelo Binomial Negativa en la siguiente sección.

Algunas particularidades encontradas en el proceso de análisis y comparación de modelos bajo la Poisson fueron las siguientes

1. Comparando las estimaciones iniciales (donde se supone un modelo con independencia) y sus errores, para los modelos antes y después de tener en cuenta la sobredispersión, no se observa cambio en los parámetros estimados, sin embargo ahora los errores estándar están "inflados" por el valor del parámetro de escala.

2. Los intervalos de confianza son más amplios, los p-valores ya no son tan pequeños, es decir que las pruebas de significancia tienden a ser más conservativas, comparadas con un modelo sin ajuste por sobredispersión.

3. Contrario a lo sucedido con las estimaciones iniciales, obtenidas bajo el supuesto de independencia, en este caso, tanto las estimaciones de los parámetros como sus errores estándar, intervalos de confianza y p-valores, obtenidos por el método GEE, no se ven afectados por el parámetro de escala o sobredispersión estimado.

6.5.3 Análisis de los datos bajo una distribución Binomial Negativa

Prosiguiendo con el análisis bajo el método propuesto usando distancias en las variables explicativas, se encuentra apropiado dejar cuatro componentes de la matriz de coordenadas $\boldsymbol{X}$ y tres de la matriz de coordenadas $\boldsymbol{G}$, para el ajuste del modelo Binomial Negativa de la matriz de datos obtenida como se muestra en el modelo 6.3. Adicionalmente, se presentan las pruebas de bondad de ajuste con el intercepto y las componentes g_1, g_2 y C_1 de la matriz de datos con distancias, que corresponden al modelo escogido debido a ser este el de mejor ajuste. En este caso se observa que, contrario al modelo con distribución Poisson, las pruebas de bondad de ajuste arrojan valores muy cercanos a uno del cociente (Valor/gl), esto es un buen indicio de la calidad del modelo, con esta distribución, véase la Tabla 6.2.

Criterio	gl	Valor	Valor/gl
Deviance	116	133.7939	1.1534
Deviance escalada	116	133.7939	1.1534
Chi-cuadrado de Pearson	116	108.3396	0.9340
Pearson X2 escalada	116	108.3396	0.9340
Log Verosimilitud		3287.8206	
AIC		769.3191	

Tabla 6.2: Criterios para valorar la bondad de ajuste bajo el modelo Binomial Negativo con MLG en DB

Una vez estimado el parámetro de escala, se ajusta el nuevo modelo corregido, donde se observan los resultados presentados en la Tabla 6.3.

Con este análisis, se observa que el intercepto y la componente g_1 asociada a los tiempos son significativas y las componentes g_2 y C_1 no son significativas

Parámetro	Estimación	Error estándar	Límites 95% de confianza		Z	$Pr > \|Z\|$
Intercept	2.0854	0.1679	1.7563	2.4146	12.42	<0.0001
g_1	-1.4688	0.1353	-1.7340	-1.2036	-10.86	<0.0001
g_2	0.0304	0.2011	-0.3638	0.4247	0.15	0.8798
C_1	-0.1505	0.7762	-1.6717	1.3708	-0.19	0.8463

Tabla 6.3: Estimación GEE de parámetros utilizando el método DB

al 5% en la actividad de forrajeo de polen. En cuanto a las estimaciones de los parámetros por medio del método GEE, Tabla 6.3, y teniendo en cuenta la estructura de correlación asumida para los datos, se encontró como la de mejor ajuste a los datos una estructura autorregresiva de orden 1, bajo una función de enlace log. El p-valor=0.31903, por lo tanto con un nivel de significancia del 5% hay evidencia estadística para no rechazar la hipótesis nula de bondad de ajuste, por lo tanto el modelo ajusta los datos.

La forma de la ecuación del modelo para la regresión Binomial Negativa es la misma que para la regresión Poisson. El log de la variable respuesta es predicho como una combinación lineal de los predictores, entonces el modelo ajustado para el MLG con DB se puede escribir así

$$\log(\hat{Y}_{ij}) = \hat{\theta}_0 + \hat{\theta}_1 \cdot g_1 + \hat{\theta}_2 \cdot g_2 + \hat{\theta}_3 \cdot C_1$$

esto implica

$$\begin{aligned}\hat{Y}_{ij} &= \exp(\hat{\theta}_0 + \hat{\theta}_1 \cdot g_1 + \hat{\theta}_2 \cdot g_2 + \hat{\theta}_3 \cdot C_1) \\ &= \exp(2.0854 - 1.4688 \cdot g_1 + 0.0304 \cdot g_2 - 0.1505 \cdot C_1)\end{aligned}$$

para $i = 1, 2, \ldots, 30$ colonias diferentes de abejas, $j = 1, \ldots, 4$ tiempos de observación. Donde g_1, g_2 y C_1 son las primeras componentes que se obtienen de la matriz de datos al realizar el producto kronecker entre $\boldsymbol{X}$ y $\boldsymbol{G'}$ tal como se muestra en la ecuación 6.3.

El gráfico 6.2 corresponde a la validación de supuestos mediante MLG con DB, el cual muestra los residuales vs los ajustados donde se puede observar que hay aleatoriedad, por lo cual se comportan en forma adecuada.El gráfico de normalidad Q-Q muestra un ajuste apropiado a la distribución normal, sin embargo hay algunas observaciones atípicas en las colas y el gráfico de residuals vs leverage muestra que hay algunos valores extremos o atípicos en los residuales.

El gráfico 6.3 de validación de supuestos para el MLG clásico muestra los residuales vs los ajustados donde se puede observar que hay aleatoriedad,

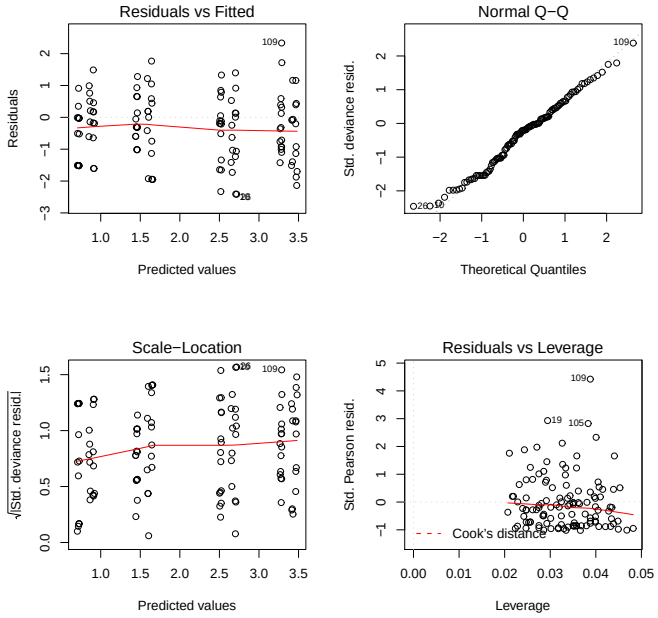


FIGURA 6.2: Gráfico de validación de supuestos MLG con DB

por lo cual se comportan en forma adecuada. El gráfico de normalidad Q-Q muestra un ajuste apropiado a la distribución normal, sin embargo hay algunas observaciones atípicas en las colas y el gráfico de residuals vs leverage muestra que hay algunos valores extremos o atípicos en los residuales, similar a los resultados obtenidos bajo DB.

Al realizar el análisis usando un MLG clásico, antes de revisar las estimaciones y sacar conclusiones acerca de la significancia de los factores para el modelo, se revisan los criterios de bondad de ajuste (ya discutidos ampliamente), proporcionados por el procedimiento GENMOD de SAS. Al modelar los datos bajo una distribución Binomial Negativa se obtienen los siguientes valores para los estadísticos.

Una vez estimado el parámetro de escala, se ajusta el nuevo modelo corregido y se modela la estructura de correlación, en este caso se encontró apropiado ajustar un modelo autorregresivo de orden 1 donde se observan los siguientes resultados

Al realizar el análisis se encuentra que las únicas variables significativas al 5% son el intercepto, la temperatura y la humedad relativa y que ninguna de las variables discretas es importante para el modelo. Aunque acá resulta ser significativa la época, al quitar las demás variables no significativas esta también resulta ser no significativa. Haciendo este nuevo análisis con el modelo Binomial Negativo se obtiene los resultados presentados en la Tabla 6.6. Dicha

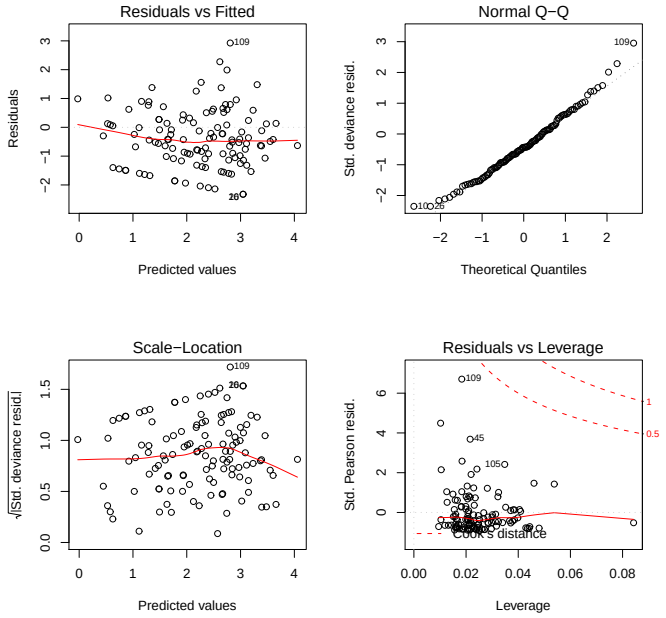


FIGURA 6.3: Gráfico de validación de supuestos MLG clásico

Criterio	gl	Valor	Valor/gl
Deviance	113	135.6273	1.2002
Deviance escalada	113	135.6273	1.2002
Chi-cuadrado de Pearson	113	136.8035	1.2107
Pearson X2 escalada	113	136.8035	1.2107
Log-verosimilitud		3277.1488	
AIC		796.6627	

TABLA 6.4: Criterios para valorar la bondad de ajuste bajo una Binomial Negativa utilizando el MLG clásico

tabla muestra las pruebas de bondad de ajuste, la cual arroja valores muy cercanos a uno del cociente (Valor/gl). Esto es un buen indicio de la calidad del modelo, con esta distribución.

Ahora se ajusta el nuevo modelo corregido y se modela la estructura de correlación, en este caso se encontró apropiado ajustar un modelo autorregresivo de orden 1 y se utilizo como función de enlace log. Finalmente, en cuanto a la estructura de correlación, SAS afirma que si las estimaciones de las covarianzas (empíricas y basadas en el modelo) son similares, entonces la estructura de correlación escogida para el análisis es adecuada. Antes de pasar a verificar

Parámetro	Estimación	Error estándar	Límites 95% de confianza		Z	$Pr > \|Z\|$
Intercept	24.5850	4.2544	16.2465	32.9236	5.78	<0.0001
grupo (campo)	0.2898	0.1717	-0.0468	0.6264	1.69	0.0915
época(lluviosa)	0.6851	0.2421	0.2106	1.1597	2.83	0.0047
día 2	-0.3670	0.3242	-1.0025	0.2684	-1.13	0.2576
día 3	-0.4808	0.2964	-1.0617	0.1002	-1.62	0.1048
temperatura	-0.6344	0.1062	-0.8426	-0.4262	-5.97	<0.0001
humedad rel	-0.0844	0.0217	-0.1269	-0.0419	-3.90	<0.0001

Tabla 6.5: Estimación GEE de parámetros en el modelo clásico

Criterio	gl	Valor	Valor/gl
Deviance	117	135.6753	1.1596
Deviance escalada	117	135.6753	1.1596
Chi-cuadrado de Pearson	117	149.7657	1.2800
Pearson X2 escalada	117	149.7657	1.2800
Log-verosimilitud		3273.2334	
AIC		796.4935	

Tabla 6.6: Criterios para valorar la bondad de ajuste bajo una Binomial Negativa utilizando el MLG clásico

esta propiedad con los datos de estudio, es necesario aclarar la diferencia entre estas dos covarianzas.

La estructura de covarianza basada en el modelo, es la estimada asumiendo que la media del modelo y la matriz de correlación de trabajo han sido especificadas correctamente. La matriz de covarianza empírica o robusta es la matriz obtenida, aún si la matriz de correlación de trabajo ha sido mal especificada.

Al comparar estas dos estructuras de covarianza, para los datos de estudio, se encontró que eran muy similares, permitiendo concluir que la estructura de correlación escogida AR(1), es la apropiada para los datos de estudio tanto en el análisis mediante DB y con el análisis clásico, donde se observan los siguientes resultados

El modelo para la regresión Binomial Negativa con el log de la variables respuesta es predicho como una combinación lineal de los predictores, entonces

Parámetro	Estimación	Error estándar	Límites 95% de confianza		Z	$Pr > \|Z\|$
Intercept	18.7269	2.3334	14.1536	23.3002	8.03	<0.0001
temperatura	-0.5239	0.0646	-0.6506	-0.3973	-8.11	<0.0001
humedad rel	-0.0456	0.0121	-0.0693	-0.0218	-3.76	0.0002

Tabla 6.7: Estimación GEE de parámetros

el modelo ajustado para el MLG clásico se puede escribir así

$$\log(\hat{Y}_{ij}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot temperatura + \hat{\beta}_2 \cdot humedad$$

esto implica

$$\begin{aligned} \hat{Y_{ij}} &= \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot temperatura + \hat{\beta}_2 \cdot humedad) \\ &= \exp(18.7269 - 0.5239 \cdot temperatura - 0.0456 \cdot humedad) \end{aligned}$$

para $i = 1, 2, \ldots, 30$ colonias diferentes de abejas y $j = 1, \ldots, 4$ tiempos de observación.

El p-valor=0.98, por lo tanto con un nivel de significancia del 5% hay evidencia estadística para no rechazar la hipótesis nula de bondad de ajuste, por lo tanto se puede decir que el modelo ajusta los datos.

Veamos el gráfico de predicciones para los modelos ajustados. La Figura 6.4 ilustra las predicciones de ambos modelos (MLG con DB y MLG clásico), con respecto a los valores observados de entradas con polen a los nidos (los círculos). Las predicciones usando un modelo Binomial Negativa mediante GEE con distancias (los triángulos) y el método clásico Binomial Negativa mediante GEE (las cruces), donde se aprecia que las predicciones son muy similares, difieren en muy poco. Con respecto al valor observado se aprecia un buen ajuste, pero resulta un poco mejor el ajuste del modelo usando distancias de acuerdo a los estadísticos de bondad de ajuste. En este gráfico el MLG clásico ajustado corresponde a aquel donde solo resulta ser significativa la temperatura y la humedad relativa al 5%.

Si se considera la Figura 6.5 de predicciones para el MLG con DB considerando una componente mas (intercepto, g_1, g_2, C_1 y X_{11}) y para el modelo clásico considerando el modelo ajustado en la Tabla 6.5. La Figura 6.5 ilustra el ajuste de ambos modelos (MLG con DB Y MLG clásico) con respecto a los valores observados de entradas con polen a los nidos (los círculos). Las predicciones usando un modelo Binomial Negativa mediante GEE con distancias (los triángulos) y el método clásico Binomial Negativa mediante GEE (las cruces), donde se aprecia que las predicciones son muy similares, difieren en

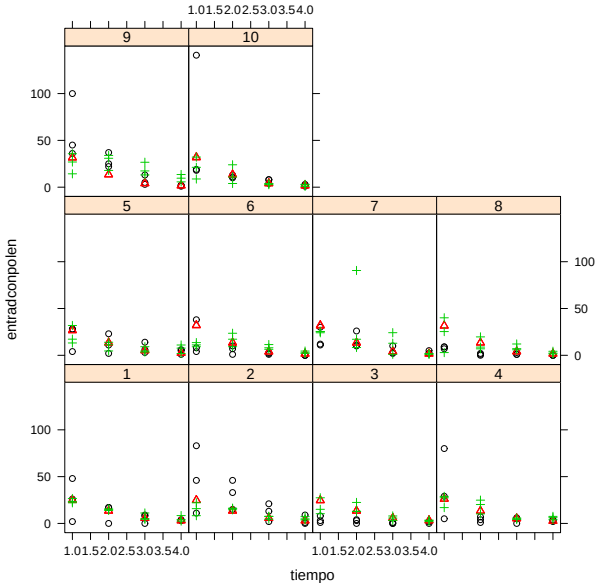


FIGURA 6.4: Entradas con polen vs predicciones usando MLG en DB y clásico en función del tiempo

muy poco. Con respecto al valor observado se aprecia un buen ajuste, pero resulta un poco mejor el ajuste del modelo usando distancias ya que al agregar mas componentes al modelo las predicciones resultan ser mejores.

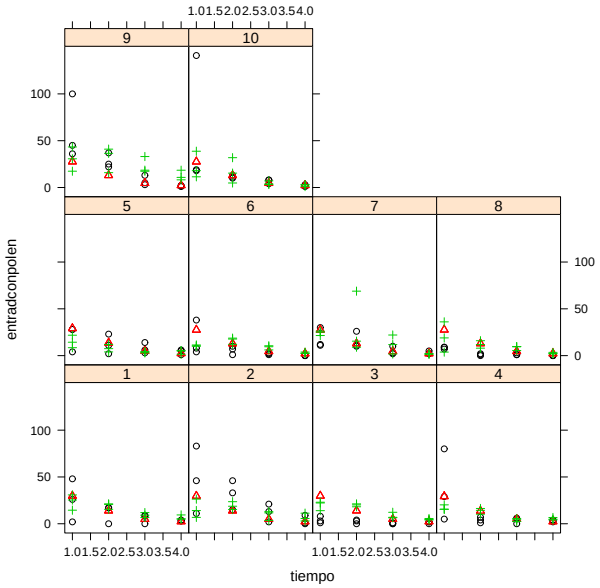


FIGURA 6.5: Entradas con polen vs predicciones usando MLG en DB y MLG clásico en función del tiempo y mas componentes

El análisis de las estimaciones obtenidas por la metodología GEE, permite encontrar que el intercepto y las variables que realmente aportan información al modelo son: temperatura y humedad relativa, mientras que ninguna de las variables discretas (época, día y grupo) son importantes para el modelo al 5% de significancia.

Capítulo 7

Conclusiones

Se propusieron varias metodologías para analizar datos longitudinales (en forma univariante, mediante MANOVA, en curvas de crecimiento y bajo respuesta no normal mediante modelos lineales generalizados) usando distancias entre observaciones (o individuos) con respecto a las variables explicativas con variables respuesta de tipo continuo. En todas las metodologías propuestas al agregar más componentes de la matriz de coordenadas principales se encuentra que se gana en las predicciones con respecto a los modelos clásicos. Por lo cual resulta ser una metodología alternativa frente a la clásica para realizar predicciones.

Se probó que el modelo MANOVA con DB y la aproximación univariante longitudinal con DB generan resultados tan robustos como la aproximación de MANOVA clásica y univariante clásica para datos longitudinales, haciendo uso en la aproximación clásica de máxima verosimilitud restringida y mínimos cuadrados ponderados bajo condiciones de normalidad. Los parámetros del modelo univariante con DB fueron estimados por el método de máxima verosimilitud restringida y por mínimos cuadrados generalizados. Para la aproximación MANOVA con DB se uso mínimos cuadrados bajo condiciones de normalidad. Además, se presentó como realizar inferencia sobre los parámetros involucrados en el modelo para muestras grandes.

Se explicó también una metodología para analizar datos longitudinales mediante modelos lineales generalizados con distancias entre observaciones con respecto a las variables explicativas, donde se encontraron resultados similares a la metodología clásica y la ventaja de poder modelar datos de respuesta continua no normal en el tiempo.

Por medio de una aplicación se ilustraron las metodologías propuestas. Se ajusto el modelo, se obtuvo la estimación de los diferentes parámetros involucrados, se realizó la inferencia estadística del modelo propuesto y la validación del modelo propuesto. Se encuentran pequeñas diferencias del método DB

con respecto al clásico en el caso de datos mixtos, especialmente en muestras pequeñas de tamaño 50, resultado de la simulación.

Mediante simulación para algunos tamaños de muestra se encontró que el modelo ajustado DB produce mejores predicciones en comparación con la metodología tradicional para el caso en que las variables explicativas sean mixtas utilizando la distancia de Gower. En tamaños de muestras pequeñas (50), independiente del valor de la correlación, las estructuras de autocorrelación, la varianza y el número de tiempos, usando los criterios de información Akaike y Bayesiano (AIC y BIC). Además, para muestras pequeñas de tamaño 50 también se encuentra más eficiente (eficiencia mayor a 1) el método DB en comparación con el método clásico, bajo los diferentes escenarios considerados. Otro resultado importante es que el método DB presenta mejor ajuste en muestras grandes (100 y 200), con correlaciones altas (0.5 y 0.9), varianza alta (50) y mayor número de mediciones en el tiempo (7 y 10).

Cuando las variables explicativas son solamente de tipo continuo o categórico o binario, se probó que las predicciones son las mismas con respecto al método clásico. Adicionalmente, se desarrollaron los programas en el software R para el análisis de este tipo de datos mediante la metodología clásica y por distancias DB para las diferentes propuestas en cada uno de los capítulos de la tesis, los cuales se anexan en un CD dentro de la tesis. Se esta trabajando en la creación de una librería en R con lo ya programado, para que todos los usuarios tengan acceso a este tipo de análisis.

Los métodos propuestos tienen la ventaja de poder hacer predicciones en el tiempo, se puede modelar la estructura de autocorrelación, se pueden modelar datos con variables explicativas mixtas, binarias, categóricas o continuas, y se puede garantizar independencia en las componentes de la matriz de coordenadas principales mientras que con las variables originales no se puede garantizar siempre independencia. Por último, el método propuesto produce buenas predicciones para estimar datos faltantes, ya que al agregar una o más componentes en el modelo con respecto a las variables explicativas originales de los datos, se puede mejorar el ajuste sin alterar la información original y por consiguiente resulta ser una buena alternativa para el análisis de datos longitudinales y de gran utilidad para investigadores cuyo interés se centra en obtener buenas predicciones.

Finalmente, en futuros trabajos en el área se puede abordar el modelo multinivel con distancias o series temporales mediante distancias que también resultarían de gran interés para tratar varios problemas de tipo práctico en áreas tales como la biología, agronomía y economía en donde son muy frecuentes este tipo de datos y donde es necesario la obtención de buenas predicciones. También, se pueden hacer estudios donde se comparen las metodologías propuestas con respecto a otras clásicas y además se pueden desarrollar propuestas

para datos missing que son muy frecuentes en la práctica.

Bibliografía

Anderson, T. W. (2003), *An introduction to multivariate analysis*, Jhon Wiley & Sons, New York.

Andreoni, S. (1989), *Modelos de efeitos aleatórios para análise de dados longitudinais não balanceados em relaçao ao tempo*, Disertación Maestría, Instituto de Matemática e Estatística, Universidad de São Paulo.

Aranda, O. F. (1981), 'On two families of transformations to additivity for binary response data', *Biometrika* **68**, 357–363.

Arenas, C. & Cuadras, C. (2002), 'Recent statistical methods based on distances', *Constributions to Science, Institut d'Estudis Catalans Barcelona* **2**(2), 183–191.

Arenas, C. & Cuadras, C. (2004), 'Comparing two methods for joint representation of multivariate data', *Communications in Statistics - Simulation and Computation* **33**, 415–430.

Bartkowiak, A. & Jakimiec, M. (1994), 'Distance-based regression in prediction of solar flare activicty', *Qüestió* **18**, 7–38.

Bartlett, M. (1937), 'Properties of sufficiency and statistical tests.', *Proceedings of the Royal Statistical Society Series* **A 160**, 268–282.

Bloomfield, P. & Watson, G. S. (1975), 'The inefficiency of least squares', *Biometrika* **62**(1), 121–128.

Boik, R. J. (1991), 'The mixed model for multivariate repeated measures: validity conditions and an approximate test', *Psychometryka* **53**, 469–486.

Boj, E., Claramunt, M. & Fortiana, J. (2007), 'Selection of predictors in distance-based regression', *Communications in Statistics - Simulation and Computation* **36**, 87–98.

Boj, E., Delicado, P. & Fortiana, J. (2010), 'Distance-based local linear regression for functional predictors', *Computational Statistics and Data Analysis* **54**, 429–437.

Boj, E., Grané, A., Fortiana, J. & Claramunt, M. (2007), 'Implementing PLS for distance-based regression: computational issues', *Computational Statistics* **22**, 237–248.

Box, G. and. Cox, D. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society, Series B* **26**, 211–246.

Breslow, N. (1990), 'Test of hypotheses in overdispersered poisson regression and othe quasilikelihood models', *Journal of the American Statistical Association* **85**, 565–571.

Chaganty, N. R. (1997), 'An alternative approach to the analysis of longitudinal data via generalized estimating equations', *Journal of Statistical Planning and Inference* **63**, 39–54.

Chaganty, N. R. & Mav, D. (2007), 'Estimation methods for analyzing longitudinal data ocurring in biomedical research', *Computational Methods in Biomedical Research* **12**, 371–400.

Chaganty, N. R. & Naik, D. N. (2002), 'Analysis of multivariate longitudinal data using quasi-least squares', *Journal of Statistical Planning and Inference* **103**, 421–436.

Chaganty, N. R. & Shults, J. (1999), 'On eliminating the asymptotic bias in the quasi- least squares estimate of the correlation parameter', *Journal of Statistical Planning and Inference* **76**, 145–161.

Chiou, J. & Müller, H. (1998), 'Quasi-likelihood regression with unknown link and variance functions', *Journal of the American Statistical Association* **93**, 1376–1387.

Chiou, J. & Müller, H. (1999), 'Nonparametric quasi-likelihood', *The Annals of Statistics* **27**, 36–64.

Chiou, J., Müller, H., Wang, J. & Carey, J. (2003), 'A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies', *Statistica Sinica* **13**, 1119–1133.

Cho, P. (1997), 'The generalized estimating equation approach when data are not missing completely at random', *Journal of the American Statistical Association* **92**, 1320–1329.

Cox, T. V. & Cox, M. A. A. (2001), *Multidimensional scaling*, Chapman and Hall/CRC, Boca Raton, Florida.

Cramer, E. M. & Nicewander, W. (1979), 'Some symmetric, invariant measures of multivariate association', *Psychometrika* **44**, 43–54.

Crowder, M. (1995), 'On the use of a working correlation matrix in using generalized linear models for repeated measures', *Biometrika* **82**, 407–410.

Crowder, M. (2001), 'On repeated measures analysis with misspecified covariance structure', *Journal of the Royal Statistical Society* **63**, 55–62.

Crowder, M. J. & Hand, D. J. (1990), *Analysis of repeated measures*, Chapman and Hall, London.

Cuadras, C. & Arenas, C. (1990), 'A distance based regression model for prediction with mixed data', *Communications in Statistics A - Theory and Methods* **19**, 2261–2279.

Cuadras, C., Arenas, C. & Fortiana, J. (1996), 'Some computational aspects of a distance-based model for prediction', *Communications in Statistics - Simulation and Computation* **25**(3), 593–609.

Cuadras, C. & Fortiana, J. (1993), 'Aplicaciones de las distancias en estadística', *Qüestió* **17**, 39–74.

Cuadras, C. M. (1989), *Distance Analysis in discrimination and classification using both continuous and categorical variables*, In: Recent Developments in Statistical Data Analysis and Inference. (Y. Dodge ed.). Elsevier Sciencie Publisher, North-Holland, Amsterdam, 459-474.

Cuadras, C. M. (2007), *Métodos multivariados basados en distancias*, Curso de doctorado, Universidad de Barcelona, Barcelona.

Cuadras, C. M. (2008), *Distance based association multi-sample tests for general multivariate data*, In: Advances in Mathematical and Statistical Modeling, (B.C. Arnold, N. Balakrishnan, J. M. Sarabia, R. Mínguez, Eds), Birkhauser, Boston, pp. 61-71.

Cuadras, C. M. (2010), *Nuevos métodos de análisis multivariante*, CMC Editions, Barcelona.

Cuadras, C. M. (2011), *Distance-Based Approach in Multivariate Association. In: New Perspectives in Statistical Modeling and Data Analysis*, Springer, Berlin, pp. 535-542.

Davidian, M. (2005), *Applied Longitudinal Data Analysis*, Chapman and Hall, North Carolina State University, North Carolina.

Davidian, M. & Carrol, R. (1987), 'Variance function estimation', *Journal of the American Statistical Association* **82**, 1079–1091.

Davidian, M. & Carrol, R. (1988), 'A note on extended quasilikelihood', *Journal Royal of the Statistical Society, Series B* **50**, 74–82.

Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer-Verlag, New York.

Diggle, P., Heagerty, P., Liang, K. Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford University Press, New York.

Diggle, P., Liang, K. Y. & Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford University Press, New York.

Dobson, A. J. (2002), *An Introduction to Generalized Linear Models*, 2nd ed. Chapman Hall, Boca Raton, Florida.

Escoufier, Y. (1973), 'Le trataiment des variables vectorielles', *Biometrics* **29**, 751–760.

Esteve, A., Boj, E. & Fortiana, J. (2010), 'Interaction terms in distance-based regression', *Communications in Statistics A - Theory and Methods* **38**(19), 3498–3509.

Faraway, J. J. (1999), 'A graphical method of exploring the mean structure in longitudinal data analysis', *Journal of Computational and Graphical Statistics* **8**, 60–68.

Fitzmarice, G., Laird, N. & Rotnizky, A. (1993), 'Regression models for discrete longitudinal responses', *Statistical Science* **8**, 284–309.

Frey, K. S., Potter, G. D., Odom, T. W., Senor, M. A., Reagan, V. D., Weir, V. H., Elsslander, R. V. T., Webb, M. S., Morris, E. L., Smith, W. B. & Weigand, K. E. (1992), 'Plasma silicon and radiographic bone density on weanling quarter horses fed sodium zelolite a', *Journal of Equine Veterinary Science* **12**, 292–296.

Genolini, C. & Falissard, B. (2011), 'KML: A package to cluster longitudinal data', *Computation Methods Programs Biomedical* **104**(3), 112–121.

Geraci, M. & Bottai, M. (2007), 'Quantile regression for longitudinal data using the asymmetric laplace distribution', *Biostatistics* **8**(1), 140–154.

Gower, J. (1968), 'Adding a point to vector diagrams in multivariate analysis', *Biometrika* **55**, 582–585.

Gower, J. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* **27**, 857–871.

Gower, J. C. & Krzanowski, W. (1999), 'Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance', *Applied Statistics* **48**(4), 505–519.

Gray, S. M. and, B. R. (2000), 'Multidimensional longitudinal data: estimating a treatment effect from continuous, discrete, or time to event response variables', *Journal of the American Statistical Association* **95**, 396–406.

Grizzle, J. E. & Allen, D. M. (1969), 'Analysis of growth and dose response curves', *Biometrics* **25**, 357–381.

Halekoh, U. and. Højsgaard, S. & Yan, J. (2006), 'The R package geepack for generalized estimating equations', *Journal of Statistical Software* **15**(2), 1–11.

Harville, D. (1974), 'Bayesian inference for variance components using only error contrasts', *Biometrika* **61**, 383–385.

Heyde, C. & Morton, R. (1993), 'On constrained quasi-likelihood estimation', *Biometrika* **80**, 755–761.

Hinkelmann, K. & Kempthorne, O. (1994), *Design and Analysis of Experiments*, John Wiley & Sons. Inc, New York.

Hwang, H., Montreal, H. & Takane, Y. (2004), 'A multivariate reduced rank growth curve model with unbalanced data', *Psychometrika* **69**(1), 65–79.

Jennrich, R. I. & Schluchter, M. D. (1986), 'Unbalanced repeated measures models with structured covariance matrices', *Biometrics* **42**, 805–820.

Jones, R. M. (1993), *Longitudinal Data with Serial Correlation: A State Space Approach*, Chapman and Hall, London.

Jowaheer, V. & Sutradhar, B. (2002), 'Analysing longitudinal count data with overdispersion', *Biometrika* **89**(2), 389–399.

Jørgensen, B., Lundbye, S., Song, P. & Sun, L. (1996), 'State-space models for multivariate longitudinal data of mixed types', *The Canadian Journal of Statistics* **24**(3), 385–402.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. & Wilcox, R. (2000), 'Testing treatment effects in repeated measures designs: trimmed means and bootstrapping', *British Journal of Mathematical and Statistical Psychology* **53**, 175–191.

Keselman, H. J. & Lix, L. M. (1997), 'Analysing multivariate repeated measures designs when covariance matrices are heterogeneous', *British Journal of Mathematical and Statistical Psychology* **49**, 275–298.

Khatri, C. G. (1966), 'A note on a MANOVA model applied to problems in growth curves', *Annals of the Institute of Statistical Mathematics* **18**, 75–86.

Kowalchuk, R., Keselman, H. & Algina, J. (2003), 'Repeated measures interaction test with aligned ranks', *Multivariate Behavioral Research* **38**(4), 433–461.

Kshirsagar, A. M. & Boyce, S. (1995), *Growth Curves*, Marcel Dekker, New York.

Laird, N. M., Lange, N. & Stram, D. (1987), 'Maximum likelihood computations with repeated measures: application of the EM algorithm', *Journal of the American Statistical Association* **82**, 97–105.

Laird, N. & Ware, J. (1982), 'Random effects models for longitudinal data', *Biometrics* **38**, 963–974.

Lee, J. C. (1991), *Growth Curve Models and Technological Forecasting*, Statiscal Theory and Data Analysis II, Matusita, K. (Ed). Elsevier Science, New York.

Li, K. C. (1991), 'Sliced inverse regression for dimension reduction with discussions', *Joournal of the American Statistical Association* **86**, 316–342.

Li, K. C., Aragon, Y., Shedden, K. & Agnan, C. (2003), 'Dimension reduction for multivariate response data', *Journal of the American Statistical Association* **98**, 99–109.

Liang, K. Y. & Zeger, S. (1986a), 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics* **42**, 121–130.

Liang, K. Y. & Zeger, S. (1986b), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Liang, K. Y., Zeger, S. & Qaqish, B. (1992), 'Multivariate regression analyses for categorical data', *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

Lindsey, J. (1993), *Models for Repeated Measurements*, Oxford University Press, New York.

Little, R. & Chen, H. (1999), 'A test of missing completely at random for generalized estimating equations with missing data', *Biometrika* **86**, 1–13.

Liugen, X. & Lixing, Z. (2007), 'Empirical likelihood for a varying coefficient model with longitudinal data', *Journal of the American Statistical Association* **102**, 642–654.

Lix, L. & Algina, J. (2003), 'Analyzing multivariate repeated measures designs: a comparison of two approximate degrees of freedom procedures', *Multivariate Behavioral Research* **38**(4), 403–431.

Lumley, T. (1996), 'Generalized estimating equations for ordinal data a note on working correlation structures', *Biometrics* **52**, 354–361.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (2002), *Multivariate Analysis*, Academic Press, London.

McArdle, B. H. & Anderson, M. J. (2001), 'Fitting multivariate models to community data: a comment on distance based redundancy analysis', *Ecology* **82**, 290–297.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman Hall, London.

Molenberghs, G. & Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

Müller, H. G. (2009), *Functional modeling of longitudinal data*, In: Longitudinal data analysis (Handbooks of modern statistical methods), New York.

Müller, H. G. & Yao, F. (2010), 'Empiricial dynamics for longitudinal data', *Annals of Statistics* **38**, 3458–3486.

Nelder, J. and. Wedderburn, R. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society* **Series A 135**, 370–384.

Nelder, J. & Lee, Y. (1992), 'Likelihood, quasilikelihood and pseudolikelihood: some comparisons', *Journal of the Royal Statistical Society, Series B* **54**, 273–284.

Nelder, J. & Pregibon, D. (1987), 'And extended quasi-likelihood function', *Biometrika* **74**, 221–232.

Ortiz, A. F., Rivera, J. C. & Melo, O. O. (2012), Response surfaces optimization in growth curves through multivariate analysis. Sometido: Revista Colombiana de Estadística.

Pan, W. (2001), 'Akaike´s information criterion in generalized estimating equations', *Biometrics* **57**, 120–125.

Park, T. & Lee, S. (1997), 'A test of missing completely at random for longitudinal data with missing observations', *Statistics in Medicine* **16**, 1859–1871.

Patterson, H. D. & Thompson, R. (1971), 'Recovery of inter-block information when block sizes are unequal', *Biometrika* **58**, 545–554.

Peña, D. (2002), *Análisis de Datos Multivariantes*, McGraw Hill, Madrid.

Peng, J. & Müller, H. G. (2008), 'Distance based clustering of sparsely observed stochastic processes, with application to online auctions', *Annals of Applied Statistics* **2**(3), 1056–1077.

Potthoff, R. & Roy, S. (1964), 'A generalized multivariate analysis of variance model useful especially for growth curve problems', *Biometrika* **51**, 313–326.

Prentice, R. (1988), 'Correlated binary regression with covariates specific to each binary observation', *Biometrics* **44**, 1033–1048.

Prentice, R. & Zhao, L. (1991), 'Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses', *Biometrics* **47**, 825–839.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Rao, C. (1965), 'The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves', *Biometrika* **52**, 447–458.

Rao, C. (1973), *Linear Statistical Inference and its Applications*, John Wiley & Sons, New York.

Raudenbush, S. W. & Chan, W. S. (1992), 'Growth curve analysis in accelerated longitudinal designs', *Journal of Research in Crime and Delincuence* **29**(4), 387–411.

Raudenbush, S. W. & Chan, W. S. (1993), 'Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design', *Journal of Consulting and Clinical Psychology* **61**(6), 941–951.

Reinsel, G. (1982), 'Multivariate repeated measurement or growth curve models with multivariate random effects covariance structure', *Journal of the American Statistical Association* **77**, 190–195.

Reinsel, G. (1984), 'Estimation and prediction in a multivariate random effects generalized linear model', *Journal of the American Statistical Association* **77**, 190–210.

Rencher, A. (2002), *Methods of Multivariate Analysis*, John Wiley and Sons, New York.

Rodríguez, A. (2005), *Forrajeo de polen por obreras de melipona fasciata (hymenoptera, apidae, meliponini) en una zona rural del 30 piedemonte llanero, (Acacías-Meta-Colombia).*, Trabajo de grado, Universidad Nacional de Colombia, facultad de ciencias, departamento de Biología, Bogotá D.C.

Rubin, D. (1976), 'Inference and missing data', *Biometrika* **63**, 81–92.

Sabo & Chaganty (2009), 'Adaptation of quasi-least squares to estimate correlations within a nuclear family', *Communications in Statistics - Theory and Methods* **38**, 3059–3076.

Shults, J. & Chaganty, N. (1998), 'Analysis of serially correlated data using quasi-least squares', *Biometrics* **54**, 1622–1630.

Singer, J. D. & Willett, J. B. (2003), *Applied Longitudinal Data Analysis-Modeling Change and Event Occurrence*, Oxford University Press, New York.

Singer, J. M. & Andrade, D. (1994), 'On the choice of appropiate error terms in profile analysis', *The Statistician, Royal of Statistical Society* **43**(2), 259–266.

Srivastava, M. (2001), *Nested Growth Curves Models*, Special Isue.

Srivastava, M. & Katri, C. A. (1979), *An Introduction to Multivariate Statistics*, North-Holland, New York.

Sutradhar, B. (2003), 'An overview on regression models for discrete longitudinal responses', *Statistical Science* **18**, 377–393.

Sutradhar, B. & Das, K. (1999), 'On the efficiency of regression estimators in generalized linear models for longitudinal data', *Biometrika* **86**, 459–465.

Sutradhar, B. & Das, K. (2000), 'On the accuracy of efficiency of estimating equation approach', *Biometrics* **56**, 622–625.

Sutradhar, B. & Jowaheer, V. (2002), 'Analysing longitudinal count data with overdispersion', *Biometrika* **89**, 389–399.

Takamitsu, S. (1978), 'Information criteria for discriminating among alternative regression models.', *Econometrica* **46**, 1273–1282.

Thall, P. & Vail, S. (1990), 'Some covariance models for longitudinal count data with overdispersion', *Biometrics* **46**, 657–671.

Verbyla, A. P. & Venables, W. N. (1988), 'An extension of the growth curve models', *Biometrica* **75**, 129–138.

Wang, Y. (1996), 'A quasi-likelihood approach for ordered categorical data with overdispersion', *Biometrics* **52**, 1252–1258.

Ware, J. (1985), 'Linear models for the analysis of longitudinal studies', *The American Statistician* **39**, 95–101.

Wedderburn, R. (1974), 'Quasi-likelihood functions generalized linear models and the gauss newton method', *Biometrika* **61**, 439–447.

Welch, B. (1951), 'On the comparison of several mean values: an alternative approach', *Biometrika* **38**, 330–336.

Wessel, J. & Schork, N. J. (2006), 'Generalized genomic distance-based regression methodology for multilocus association analysis', *The American Journal of Human Genetic* **79**, 792–806.

Yanai, H., Takane, Y. & Ishii, H. (2006), 'Nonnegative determinant of a rectangular matrix: its definition and applications to multivariate analysis', *Linear Algebra and Its Applications* **417**, 259–274.

Yao, F., Müller, H. & Wang, J. (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association* **100**, 577–590.

Zorn, C. (2001), 'Generalized estimating equation models for correlated data: A review with applications', *American Journal of political science* **45**, 470–490.

Anexos A

Tablas de la simulación en datos longitudinales mixtos

Parámetros			$m = 4$						
			Basado en distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	0	50	935.09	1052.88	6.14E-6	1004.22	1074.46	7.11E-6	1.19
		100	2071.39	2219.48	3.83E-5	2045.10	2131.78	1.56E-5	0.41
		200	4140.98	4317.25	1.56E-5	4120.92	4223.43	1.09E-5	0.70
	0.5	50	883.48	1001.27	5.13E-6	951.88	1022.13	6.43E-6	1.30
		100	1921.84	2069.93	2.11E-5	1934.33	2021.01	1.38E-5	0.66
		200	3877.34	4053.61	1.12E-5	3894.24	3996.74	9.78E-6	0.87
	0.9	50	702.55	820.35	3.53E-7	752.25	822.49	4.28E-7	1.27
		100	1523.83	1671.92	1.23E-6	1512.89	1599.57	8.00E-7	0.65
		200	3027.71	3203.98	6.03E-7	3029.17	3131.67	5.35E-7	0.89
50	0	50	1193.644	1311.44	1.240E-4	1293.92	1364.16	1.622E-4	1.35
		100	2590.64	2738.73	3.19E-4	2656.69	2743.37	2.64E-4	0.83
		200	5308.82	5485.09	2.10E-4	5376.28	5478.79	2.03E-4	0.97
	0.5	50	1144.99	1262.79	1.82E-4	1241.58	1311.83	2.42E-4	1.37
		100	2468.76	2616.85	4.08E-4	2545.92	2632.60	4.01E-4	0.99
		200	5076.71	5252.98	3.24E-4	5149.59	5252.10	3.31E-4	1.02
	0.9	50	963.33	1081.12	3.19E-5	1041.95	1112.19	4.18E-5	1.36
		100	2065.35	2213.44	6.93E-5	2124.47	2211.16	6.82E-5	0.99
		200	4229.00	4405.27	6.30E-5	4284.53	4387.03	6.45E-5	1.03

Tabla A.1: Simulación con estructura de correlación compuesta simétrica, m=4

Parámetros			$m = 10$						
			Basado en distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	0	50	2331.17	2700.66	1.39E-08	2506.61	2720.29	2.63E-08	2.07
		100	5091.60	5534.43	9.01E-08	5116.44	5368.97	2.92E-08	0.33
		200	10295.08	10806.12	1.74E-07	10293.44	10583.37	9.72E-08	0.56
	0.5	50	2135.92	2505.40	9.08E-09	2302.74	2516.42	2.14E-08	2.56
		100	4599.30	5042.14	2.84E-08	4685.83	4938.36	2.20E-08	0.79
		200	9321.51	9832.56	7.95E-08	9408.90	9698.83	7.73E-08	0.98
	0.9	50	1569.02	1938.51	9.01E-11	1673.58	1887.26	1.85E-10	2.21
		100	3401.83	3844.67	3.73E-10	3357.59	3610.13	1.85E-10	0.51
		200	6680.29	7191.34	9.28E-10	6682.53	6972.46	7.89E-10	0.85
50	0	50	2980.15	3349.64	1.74E-06	3230.86	3444.54	4.33E-06	2.63
		100	6457.81	6900.65	5.99E-06	6645.40	6897.94	5.25E-06	0.89
		200	13258.36	13769.41	1.42E-05	13431.84	13721.77	1.40E-05	0.99
	0.5	50	2792.21	3161.70	2.23E-06	3026.99	3240.67	5.76E-06	2.76
		100	6010.96	6453.79	5.06E-06	6214.79	6467.33	6.42E-06	1.29
		200	12359.55	12870.59	1.65E-05	12547.31	12837.24	1.92E-05	1.17
	0.9	50	2220.18	2589.67	4.13E-08	2397.82	2611.50	1.05E-07	2.69
		100	4755.21	5198.04	8.79E-08	4886.56	5139.10	9.66E-08	1.12
		200	9693.70	10204.75	3.38E-07	9820.931	10110.86	3.841E-07	1.14

TABLA A.2: Simulación con estructura de correlación compuesta simétrica, m=10

Parámetros			$m = 7$						
			Basado en distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	0	50	1632.93	1870.80	3.78E-07	1755.03	1893.88	5.58E-07	1.55
		100	3587.08	3876.77	7.62E-07	3582.52	3749.02	2.69E-07	0.36
		200	7230.05	7567.93	8.32E-07	7209.43	7402.53	4.92E-07	0.59
	0.5	50	1512.93	1750.79	2.41E-07	1630.15	1768.99	4.17E-07	1.79
		100	3272.58	3562.27	2.25E-07	3318.9	3485.39	1.57E-07	0.71
		200	6621.22	6959.10	4.37E-07	6668.70	6861.80	3.91E-07	0.90
	0.9	50	1138.12	1375.98	3.61E-09	1216.75	1355.60	5.97E-09	1.68
		100	2496.96	2786.65	3.17E-09	2446.17	2612.66	1.55E-09	0.49
		200	4878.58	5216.46	8.71E-09	4877.31	5070.41	7.37E-09	0.85
50	0	50	2086.87	2324.73	1.91E-05	2262.01	2400.85	3.40E-05	1.87
		100	4524.58	4814.27	3.39E-05	4652.80	4819.29	2.91E-05	0.87
		200	9288.91	9626.79	3.71E-05	9406.31	9599.41	3.56E-05	0.96
	0.5	50	1971.94	2209.80	2.70E-05	2137.13	2275.97	4.93E-05	1.91
		100	4247.09	4536.78	2.77E-05	4389.18	4555.67	3.11E-05	1.14
		200	8737.19	9075.07	4.89E-05	8865.59	9058.69	5.26E-05	1.08
	0.9	50	1595.48	1833.35	1.09E-06	1723.73	1862.57	2.02E-06	1.90
		100	3424.45	3714.14	6.54E-07	3516.45	3682.94	6.56E-07	1.01
		200	6984.62	7322.51	2.46E-06	7074.19	7267.29	2.62E-06	1.07

Tabla A.3: Simulación con estructura de correlación compuesta simétrica, m=7

Parámetros			$m = 4$						
			Basado en Distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	−0.5	50	908.50	1026.29	2.34E-06	965.49	1035.73	2.26E-06	0.99
		100	2107.47	2255.56	3.18E-05	1962.93	2049.61	7.02E-06	0.22
		200	4087.99	4264.25	7.66E-06	3952.74	4055.24	3.57E-06	0.47
	0	50	934.82	1052.62	6.14E-06	1004.10	1074.35	7.11E-06	1.19
		100	2085.48	2233.57	3.83E-05	2045.33	2132.01	1.56E-05	0.41
		200	4146.26	4322.53	1.56E-05	4120.98	4223.48	1.09E-05	0.70
	0.5	50	896.43	1014.23	6.17E-06	965.49	1035.73	7.65E-06	1.28
		100	1964.15	2112.24	2.02E-05	1962.93	2049.61	1.18E-05	0.59
		200	3943.41	4119.68	1.33E-05	3952.74	4055.24	1.12E-05	0.84
	0.9	50	727.24	845.04	6.24E-07	780.10	850.35	7.78E-07	1.29
		100	1592.09	1740.18	1.43E-06	1571.47	1658.15	8.56E-07	0.60
		200	3145.53	3321.79	9.78E-07	3149.51	3252.01	8.68E-07	0.89
50	−0.5	50	1160.07	1277.86	4.67E-05	1255.19	1325.43	5.89E-05	1.30
		100	2548.66	2696.75	1.85E-04	2574.52	2661.20	1.14E-04	0.62
		200	5171.29	5347.56	8.51E-05	5208.10	5310.61	7.25E-05	0.85
	0	50	1193.41	1311.21	1.24E-04	1293.80	1364.05	1.62E-04	1.35
		100	2592.03	2740.12	3.19E-04	2656.91	2743.60	2.64E-04	0.83
		200	5309.16	5485.42	2.10E-04	5376.34	5478.84	2.03E-04	0.97
	0.5	50	1157.51	1275.31	1.76E-04	1255.19	1325.43	2.36E-04	1.37
		100	2499.90	2647.99	3.36E-04	2574.52	2661.20	3.19E-04	0.96
		200	5136.12	5312.38	3.11E-04	5208.10	5310.61	3.14E-04	1.01
	0.9	50	988.50	1106.30	4.91E-05	1069.80	1140.05	6.56E-05	1.37
		100	2124.32	2272.41	6.11E-05	2183.06	2269.74	5.86E-05	0.97
		200	4346.80	4523.07	8.86E-05	4404.87	4507.37	9.07E-05	1.03

Tabla A.4: Simulación con estructura de correlación AR(1), m=4

Parámetros			$m = 7$						
			Basado en Distancias			Modelo clásico			$ER =$
σ^2	ρ	n	AIC	BIC	E_2	AIC	BIC	E_1	E_1/E_2
10	−0.5	50	1581.04	1818.91	6.83E-08	1677.30	1816.14	7.35E-08	1.13
		100	3662.34	3952.03	2.37E-07	3418.43	3584.92	3.59E-08	0.15
		200	7129.88	7467.76	2.35E-07	6872.66	7065.76	7.68E-08	0.33
	0	50	1633.45	1871.32	3.78E-07	1754.87	1893.72	5.58E-07	1.55
		100	3623.90	3913.59	7.62E-07	3582.44	3748.93	2.69E-07	0.36
		200	7252.73	7590.61	8.32E-07	7209.34	7402.44	4.92E-07	0.59
	0.5	50	1558.27	1796.13	4.17E-07	1677.30	1816.14	6.79E-07	1.70
		100	3388.26	3677.95	3.55E-07	3418.43	3584.92	2.09E-07	0.60
		200	6846.44	7184.32	6.29E-07	6872.66	7065.76	5.05E-07	0.80
	0.9	50	1221.03	1458.90	1.54E-08	1306.41	1445.26	2.51E-08	1.68
		100	2647.32	2937.01	5.20E-09	2635.71	2802.20	3.05E-09	0.59
		200	5257.39	5595.27	2.41E-08	5266.28	5459.38	2.07E-08	0.86
50	−0.5	50	2020.15	2258.02	4.35E-06	2184.27	2323.12	7.18E-06	1.73
		100	4440.7	4730.39	9.50E-06	4488.70	4655.19	4.99E-06	0.53
		200	9012.45	9350.34	1.02E-05	9069.55	9262.65	7.81E-06	0.77
	0	50	2086.72	2324.59	1.91E-05	2355.58	2400.69	3.40E-05	1.87
		100	4528.43	4818.12	3.39E-05	4652.72	4819.21	2.91E-05	0.87
		200	9290.24	9628.12	3.71E-05	9406.22	9599.32	3.56E-05	0.96
	0.5	50	2015.44	2253.31	2.80E-05	2184.27	2323.12	5.06E-05	1.89
		100	4348.24	4637.93	3.10E-05	4488.70	4655.19	3.28E-05	1.07
		200	8942.97	9280.85	4.74E-05	9069.55	9262.65	4.96E-05	1.05
	0.9	50	1677.52	1915.39	3.62E-06	1813.39	1952.23	6.60E-06	1.89
		100	3598.10	3887.79	1.11E-06	3705.99	3872.48	1.18E-06	1.08
		200	7365.37	7703.25	5.08E-06	7463.17	7656.27	5.42E-06	1.07

Tabla A.5: Simulación con estructura de correlación AR(1), m=7

Parámetros			$m = 7$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	-0.5	50	1754.87	1793.45	1765.43	1788.58
		100	3778.78	3824.29	3572.62	3599.92
		200	7322.05	7374.49	7174.46	7205.93
	0	50	1757.65	1796.23	1769.13	1792.28
		100	3777.65	3823.16	3574.70	3602.00
		200	7323.99	7376.44	7175.62	7207.08
	0.5	50	1755.87	1794.45	1765.43	1788.58
		100	3773.12	3818.63	3572.62	3599.92
		200	7325.77	7378.21	7174.46	7205.93
	0.9	50	1750.96	1789.54	1760.92	1784.07
		100	3764.79	3810.30	3567.25	3594.55
		200	7326.97	7379.41	7176.27	7207.73
50	-0.5	50	2306.33	2344.91	2328.73	2351.88
		100	4727.86	4773.37	4699.22	4726.53
		200	9441.99	9494.43	9427.67	9459.14
	0	50	2309.98	2348.56	2332.44	2355.58
		100	4728.89	4774.40	4701.30	4728.61
		200	9443.66	9496.10	9428.83	9460.29
	0.5	50	2306.83	2345.41	2328.73	2351.88
		100	4724.86	4770.37	4699.22	4726.53
		200	9443.96	9496.40	9427.67	9459.14
	0.9	50	2301.57	2340.15	2324.22	2347.37
		100	4717.53	4763.04	4693.85	4721.16
		200	9444.65	9497.09	9429.48	9460.95

Tabla A.6: Simulación con estructura de correlación AR(1), m=7

Parámetros			$m = 4$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	-0.5	50	1013.99	1046.98	1017.29	1037.084
		100	2182.35	2222.26	2047.37	2071.317
		200	4193.24	4240.09	4104.74	4132.85
	0	50	1013.46	1046.44	1016.22	1036.012
		100	2182.45	2222.36	2044.42	2068.37
		200	4193.50	4240.34	4105.64	4133.751
	0.5	50	1014.56	1047.54	1017.29	1037.084
		100	2184.92	2224.84	2047.37	2071.317
		200	4196.35	4243.20	4104.74	4132.85
	0.9	50	1013.76	1046.74	1016.33	1036.121
		100	2189.37	2229.29	2051.93	2075.88
		200	4196.77	4243.62	4104.81	4132.91
50	-0.5	50	1329.24	1362.23	1339.18	1358.97
		100	2715.79	2755.71	2691.14	2715.09
		200	5403.18	5450.02	5392.29	5420.4
	0	50	1328.42	1361.41	1338.11	1357.90
		100	2714.41	2754.32	2688.20	2712.15
		200	5404	5450.85	5393.19	5421.30
	0.5	50	1329.55	1362.53	1339.18	1358.97
		100	2717.44	2757.35	2691.14	2715.09
		200	5404.81	5451.65	5392.29	5420.4
	0.9	50	1328.71	1361.69	1338.22	1358.01
		100	2722.52	2762.43	2695.71	2719.66
		200	5404.95	5451.80	5392.36	5420.46

Tabla A.7: Simulación con estructura de correlación AR(1), m=4

Parámetros			$m = 10$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	−0.5	50	2506.39	2548.53	2526.26	2551.54
		100	5366.31	5415.39	5097.88	5127.32
		200	10431.99	10488	10247.25	10280.86
	0	50	2502.51	2544.66	2523.18	2548.47
		100	5370.69	5419.77	5101.64	5131.09
		200	10429.75	10485.76	10241.09	10274.7
	0.5	50	2505.45	2547.59	2526.26	2551.54
		100	5366.32	5415.39	5097.88	5127.32
		200	10424.3	10480.31	10247.25	10280.86
	0.9	50	2505.41	2547.55	2527.93	2553.22
		100	5358.95	5408.03	5089.22	5118.67
		200	10421.91	10477.91	10250.85	10284.45
50	−0.5	50	3295.28	3337.42	3330.98	3356.26
		100	6740.72	6789.8	6707.32	6736.76
		200	13477.06	13533.07	13466.13	13499.73
	0	50	3291.77	3333.91	3327.90	3353.19
		100	6745.77	6794.84	6711.08	6740.52
		200	13472.72	13528.73	13459.97	13493.57
	0.5	50	3294.90	3337.04	3330.98	3356.26
		100	6740.92	6790.00	6707.32	6736.76
		200	13473.44	13529.45	13466.13	13499.73
	0.9	50	3295.07	3337.22	3332.65	3357.94
		100	6732.27	6781.35	6698.66	6728.11
		200	13473.7	13529.71	13469.72	13503.33

TABLA A.8: Simulación con estructura de correlación AR(1), m=10

Parámetros			$m = 4$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	0	50	1013.46	1046.44	1016.22	1036.01
		100	2182.45	2222.36	2044.42	2068.37
		200	4193.50	4240.34	4105.64	4133.75
	0.5	50	1014.14	1047.12	1016.70	1036.49
		100	2186.85	2226.77	2049.78	2073.73
		200	4197.44	4244.28	4106.03	4134.14
	0.9	50	1013.39	1046.38	1015.89	1035.68
		100	2190.05	2229.96	2052.68	2076.63
		200	4197.13	4243.98	4105.19	4133.29
50	0	50	1328.42	1361.41	1338.11	1357.90
		100	2714.41	2754.32	2688.20	2712.15
		200	5404	5450.85	5393.19	5421.30
	0.5	50	1329.15	1362.13	1338.59	1358.38
		100	2719.89	2759.81	2693.56	2717.51
		200	5406.01	5452.86	5393.58	5421.69
	0.9	50	1328.32	1361.30	1337.77	1357.56
		100	2723.32	2763.23	2696.45	2720.40
		200	5405.36	5452.20	5392.74	5420.85

Tabla A.9: Simulación con estructura de correlación compuesta simétrica, m=4

Parámetros			$m = 7$			
			Basado en distancias		Modelo clásico	
σ^2	ρ	n	AIC	BIC	AIC	BIC
10	0	50	1757.65	1796.23	1769.13	1792.28
		100	3777.65	3823.16	3574.70	3602.0
		200	7323.99	7376.44	7175.62	7207.081
	0.5	50	1755.63	1794.21	1765.5	1788.648
		100	3769.86	3815.37	3570.74	3598.05
		200	7328.91	7381.36	7178.04	7209.509
	0.9	50	1750.01	1788.59	1760.16	1783.311
		100	3763.04	3808.55	3565.18	3592.487
		200	7328.42	7380.86	7177.68	7209.146
50	0	50	2309.98	2348.56	2332.44	2355.584
		100	4728.89	4774.40	4701.30	4728.609
		200	9443.66	9496.10	9428.83	9460.294
	0.5	50	2306.73	2345.31	2328.80	2351.951
		100	4722.51	4768.03	4697.35	4724.656
		200	9446.63	9499.07	9431.26	9462.722
	0.9	50	2300.58	2339.16	2323.47	2346.615
		100	4715.81	4761.33	4691.79	4719.094
		200	9445.65	9498.10	9430.89	9462.359

TABLA A.10: Simulación con estructura de correlación compuesta simétrica, m=7