UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD IZTAPALAPA

MAESTRÍA EN CIENCIAS

(MATEMÁTICAS APLICADAS E INDUSTRIALES)

ANÁLISIS DE MODELOS DE DATOS LONGITUDINALES

T E S I S

QUE PARA OBTENER EL GRADO DE:

M A E S T R O E N C I E N C I A S

P R E S E N T A:

MIGUEL ANGEL POLO VUELVAS

ASESOR: DR. GABRIEL ESCARELA PÉREZ.

COASESOR EXTERNO: DR. RAMSÉS MENA CHÁVEZ.

Mayo de 2007.

ANÁLISIS DE MODELOS DE DATOS LONGITUDINALES

T E S I S

QUE PARA OBTENER EL GRADO DE:

M A E S T R O EN C I E N C I A S

(MATEMÁTICAS APLICADAS E INDUSTRIALES)

P R E S E N T A:

MIGUEL ANGEL POLO VUELVAS

No puede existir un lenguaje más universal y simple, más carente de errores y oscuridades y por lo tanto más apto para expresar las relaciones invariables de las cosas naturales [...]. [Las matemáticas] parecen constituir una facultad de la mente humana destinada a compensar la brevedad de la vida y la imperfección de los sentidos.

Joseph Fourier,
Théorie analytique de la chaleur:
Discurso preliminar (1822)

Agradecimientos

Desde el inicio hasta el final de este posgrado me has apoyado, me contagiaste de tu fuerza y tu tenacidad, permitiendo que a la vez, al tomarte de la mano y mirarte a los ojos, sintiera tu ternura y tu infinito amor. Gaby, TE AMO.

Doy las gracias a mis padres, porque siempre me han impulsado en la consecución de mis proyectos, y a mis hermanos, con los que he aprendido tantas cosas y con los que disfruto cada momento que podemos compartir.

Agradezco muy especialmente al Dr. Gabriel Escarela Pérez por la enorme paciencia que ha tenido para apoyarme en la consecución de esta meta.

A todos los profesores con los que compartí esos ratos de aprendizaje, en especial, al Dr. Carlos Ibarra y al Dr. Joaquín Delgado, quienes siempre estuvieron con la disposición para incrementar mis conocimientos.

No puedo dejar de lado los divertidos momentos compartidos con mis amigos; los que fueron de ocio, los que fueron de tensión, los que fueron de estudio, los que fueron de desvelo y los que fueron de cooperación. A todos los compañeros que conocí a lo largo de este posgrado, gracias por estar aquí.

Prefacio

La Estadística se ha convertido en una necesidad para un amplio sector de la sociedad. Problemas económicos, biológicos, físicos, químicos, incluso sociales, no podrían ser resueltos o manipulados, si no existiera una herramienta adecuada para tal efecto.

La búsqueda de tal herramienta, ha llevado a muchos científicos e ingenieros a desarrollar instrumentos, métodos u objetos matemáticos que ayuden a ampliar los conocimientos sobre el tema y a resolver problemas que antes eran irresolubles.

Así mismo, la comunicación, divulgación y exposición de estos temas, y de la ciencia en general, ha contribuido enormemente para que cada vez más personas estén interesadas por descubrir, aprender, experimentar e inventar.

El presente trabajo, es una investigación realizada a merced del interés que despierta en mí la Estadística Inferencial; ha sido realizado con la finalidad de presentar y exponer lo más claramente posible los modelos principales usados en el tratamiento de datos longitudinales, considerando de manera muy importante su respectivo apoyo computacional, que facilita y agiliza el análisis y tratamiento de dichos datos.

Tomando en consideración que esta Maestría en Ciencias es de Matemáticas Aplicadas y dado el enorme grado de aplicabilidad de la Estadística Inferencial en el mundo global, decidí trabajar en un tema que a mi parecer, está siendo estudiado y aplicado muy ampliamente en todas las ramas de la ciencia.

El objetivo de esta tesis es servir como referencia en el estudio de datos longitudinales considerando los enfoques más usados en la literatura estadística. La gran mayoría de los textos dedicados al estudio de datos longitudinales sólo se basan en algún tipo de modelo particular y restringe dicho estudio a un tipo de datos muy específico adecuado al modelo que presentan.

Por ello en este trabajo se exponen, además de los distintos modelos, los

casos en los cuales es más adecuado trabajar con uno u otro modelo.

La estructura de esta tesis es como sigue:

Capítulo 1. Se exponen las características de los datos longitudinales y sus diferencias con otros tipos de arreglos de datos, así como los ejemplos a utilizar en la ilustración de las metodologías.

Capítulo 2. En este capítulo se exponen los principales modelos usados en el tratamiento de datos longitudinales.

Capítulo 3. Aquí se presentan los métodos de estimación de parámetros que se usan en el ajuste de modelos de datos longitudinales.

Capítulo 4. Algunos ejemplos son analizados usando los enfoques de datos longitudinales mediante ajustes de modelos. Se hace uso del paquete computacional estadístico R.

Capítulo 5. Presenta las conclusiones de este trabajo y algunos ejemplos de otros enfoques que recientemente han sido aplicados al análisis de datos longitudinales.

Índice general

Capítulo 1

Introducción

En este capítulo se definen las características de los datos longitudinales incluyendo algunos de sus enfoques, los casos donde se utilizan, así como las diferencias con otros tipos de arreglos de datos como los de corte transversal. También se incluye la descripción de los datos que servirán para ilustrar algunas de las diferentes metodologías existentes en la literatura y otras contribuciones presentadas aquí.

1.1. Respuestas independientes y con dependencia en serie

En muchas ramas de la ciencia se realizan experimentos para medir distintas características de cierto fenómeno. Tales experimentos pueden consistir en mediciones repetidas a lo largo de un periodo de tiempo sobre un mismo individuo, obteniendo un historial que muestra el desarrollo o evolución de las características que se miden. Como ejemplos, se pueden mencionar los movimientos de los precios de ciertas acciones financieras, las mediciones diarias de glucosa en la sangre de un diabético (para controlar adecuadamente sus problemas de hipo o hiperglucemia), o las observaciones de algún otro padecimiento crónico. Las observaciones de este tipo de experimentos constituyen una *serie de tiempo*. Entonces se puede definir una serie de tiempo como el conjunto de datos resultado de las observaciones repetidas durante un periodo de tiempo de alguna o algunas características de interés sobre una sola unidad experimental (individuo).

Otro tipo de estudios, los cuales actualmente son los más comunes de

encontrar en la literatura estadística son los estudios de *corte transversal*. En este tipo de datos se obtiene una sola observación para cada uno de los individuos de una muestra; cada una de estas observaciones está constituida por una *variable respuesta* (o simplemente respuesta), que es el foco de atención del estudio, y por un conjunto de *variables explicativas* que son datos de las características del individuo estudiado o de las condiciones ambientales dentro de las cuales se realiza el estudio, las cuales sirven para explicar el comportamiento de la respuesta; dicho de otra forma, el comportamiento de la respuesta depende de los distintos valores que tomen las variables explicativas. Por ejemplo: el número de horas que un niño pasa viendo la televisión a la semana (variable explicativa) puede ser determinante en el riesgo de padecer problemas de lento aprendizaje (respuesta); o las mediciones de peso, nivel de triglicéridos y colesterol en la sangre, así como la presión sanguínea (variables explicativas), pueden constituir factores muy significativos en el desarrollo de problemas cardíacos (respuesta), como puede ser una obstrucción de la arteria coronaria (lo que puede desembocar en un infarto al miocardio) debida a altos porcentajes de colesterol y triglicéridos.

Los datos de corte transversal permiten modelar la respuesta de la muestra, caracterizada por los valores de las variables explicativas a fin de estimar los parámetros de población que relacionan a las variables explicativas con la respuesta.

Los datos longitudinales pueden verse como la fusión del enfoque de series de tiempo y el enfoque de corte transversal. Son arreglos en los cuales se consideran varias unidades experimentales (personas, empresas, ciudades, animales, etc.) de las cuales se registran repetidamente a lo largo de un periodo de estudio las observaciones de las respuestas de interés conjuntamente con sus variables explicativas. A continuación se describe a los datos longitudinales con mayor detalle.

1.2. Características de los datos longitudinales

La característica principal que define a los datos longitudinales o a un estudio longitudinal, a diferencia de un estudio de corte transversal, es que los individuos son observados repetidamente a lo largo del tiempo, registrando los valores que toman las variables explicativas que intervienen en el estudio

para entender el comportamiento de la respuesta, la cual también se registra a lo largo del periodo de estudio. En este sentido, es posible estudiar los efectos del tiempo en cada individuo, así como las características globales de la población estudiada.

Estos datos pueden ser coleccionados por medio de un seguimiento, ya sea partiendo de una fecha base hacia adelante en el tiempo (estudio prospectivo), o bien, recavando información histórica de los individuos a estudiar (estudio retrospectivo).

Los datos longitudinales requieren métodos estadísticos especiales, debido a que el conjunto de respuestas en un sujeto comúnmente está intercorrelacionado, es decir, los resultados de cierta medición pueden depender (están correlacionados) de los resultados de alguna medición pasada; por ejemplo, el valor de la medición actual del nivel de glucosa en la sangre de un sujeto depende del valor registrado en la medición anterior. También es posible que las respuestas entre individuos se consideren correlacionadas. Esta correlación necesita tomarse en cuenta para obtener inferencias válidas. Las estructuras de dependencia entre individuos pueden darse de dos formas: asumir independencia entre los individuos, o bien, definir una forma de relación entre ellos: por ejemplo, los elementos de una familia podrían estar relacionados por su grupo de sangre o su color de piel. Esta es una gran diferencia que tienen los estudios longitudinales con los de corte transversal, pues estos últimos generalmente suponen independencia entre respuestas.

La principal ventaja de un análisis longitudinal es su efectividad para estudiar los cambios en las respuestas. Respecto a un estudio de corte transversal, los estudios longitudinales tienen la ventaja de que predicen más acertadamente el comportamiento futuro, pues al tener observaciones repetidas para los individuos se puede obtener fácilmente la razón instantánea de cambio de los datos, además de que se toma en cuenta la correlación entre observaciones como parte de la información que explica el comportamiento de las respuestas.

Otro mérito del estudio longitudinal es su habilidad para distinguir el grado de variación en la respuesta a lo largo del tiempo para un individuo en particular, de la variación de la respuesta poblacional.

Cabe mencionar que la elección del modelo estadístico depende del tipo de respuesta y del tipo de inferencia que se desea hacer. Existen varias metodologías usadas para el modelado de datos longitudinales: los modelos marginales, dirigidos a modelar la media y varianza de los datos considerando ciertas estructuras de correlación; los modelos de efectos aleatorios, donde

un componente aleatorio "agregado" explica la correlación entre los individuos, así como la heterogeneidad no observada que está implícita en esta correlación; y los modelos de transición, enfocados en la distribución condicional de una respuesta, dada la historia pasada, tanto de las respuestas anteriores como de las variables explicativas. Los ejemplos que a continuación se describen permitirán ilustrar los enfoques mencionados antes, los cuales están diseñados para el estudio de datos longitudinales.

1.3. Presentación de los datos

Los ejemplos siguientes pertenecen a distintas ramas de la ciencia tales como: biología, medicina y economía. Fueron elegidos porque facilitan el entendimiento de los modelos y porque existe una fuente abierta para todo público de donde se pueden obtener fácilmente.

Ejemplo 1.1. *Número de células CD4+*

El VIH ataca a una célula inmune llamada la célula CD4+, la cual organiza la respuesta inmunológica a agentes infecciosos. El número de células CD4+ de una persona infectada puede ser usado para monitorear el avance de la enfermedad.

Esta base de datos consiste de 2376 observaciones del número de células CD4+ correspondientes a 369 hombres infectados registrados en el *Multicenter AIDS Cohort Study* [39].

Uno de los objetivos del estudio es el de caracterizar el decremento de las células CD4+ a lo largo del tiempo. La meta de un análisis longitudinal de estos datos incluye modelar el comportamiento del decremento de células CD4+, así como el comportamiento individual de cada sujeto tomando en cuenta sus características particulares e identificar los factores que predicen los cambios en las células CD4+. De esta manera, se asume que la variable respuesta es el conteo (número) de células CD4+, mientras que las demás variables, descritas a continuación, se consideran variables explicativas de los cambios en este conteo.

El arreglo de los datos consta de siete variables. La primera representa el tiempo, en años, desde la seroconversión, esto es, desde que el individuo se contagió o se detectó la presencia del virus. En este caso, algunos datos tienen signo negativo, pues fueron observados desde antes del contagio. La segunda corresponde a los conteos de células CD4+ por mililitro de sangre;

un individuo no infectado tiene alrededor de 1100 células CD4+ por mililitro. La tercera variable es la edad relativa a un origen arbitrario. La cuarta corresponde al número de cajetillas de cigarros fumados por día. Para la quinta variable se tienen respuestas dicotómicas que indican si el paciente usa drogas o no. La sexta es el número de parejas sexuales del paciente. El cesd que es una escala de medición para enfermedades mentales, se registra como la séptima variable. Finalmente, se identifica al paciente con un número de registro.

Ejemplo 1.2. *Ataques epilépticos*

Estos datos fueron analizados inicialmente por Thall y Vail [69] y por Breslow y Clayton [5]. Para cada uno de 59 pacientes con epilepsia, primeramente se registró el número de ataques durante un periodo base de 8 semanas. Después fueron seleccionados al azar para formar uno de dos grupos; uno de esos grupos se asignó para ser tratados con *progabide* (un medicamento antiepiléptico) y el otro con un placebo. El número de ataques fue registrado en cuatro intervalos bisemanales consecutivos. El interés médico principal es conocer la efectividad del progabide para reducir la cantidad de ataques epilépticos en comparación con la efectividad de un placebo.

Ejemplo 1.3. *Monitoreo de transplantes de corazón*

Este estudio corresponde una serie de exámenes angiográficos aproximadamente anuales de receptores de transplante de corazón [65]. En cada examen para cada paciente se mide el grado de vasculopatía de la arteria coronaria (CAV por sus siglas en inglés), que es un deterioro de las paredes arteriales y en este estudio está clasificado en uno de cuatro estados: 1 (no hay CAV), 2 (CAV moderada), 3 (CAV severa) y 4 (muerte). El conjunto de datos contiene 2846 observaciones correspondientes a 622 pacientes. Las variables explicativas son: la edad en años del receptor en cada examen médico, la fecha en que se realiza el examen (en años después del transplante), la edad del donante, el sexo del paciente, la razón del transplante, que puede ser por Isquemia ó por Cardiopatía Dilatada Idiopática, el número de episodios de rechazo agudo; el estado de CAV, explicado antes, es en este caso la respuesta. El objetivo es estudiar el curso de vida de personas con transplante de corazón (en términos del grado de CAV) y el periodo de estancia en cada estado, así como las probabilidades de mejora o deterioro.

Ejemplo 1.4. *Programa de seguro* Medicare

Un estudio realizado anualmente de 1990 a 1995 por la *Health Care Financing Administration* fue presentado por Frees, Young y Luo [23] en el cual se observaron los cargos médicos de los pacientes de hospital que fueron cobrados por el programa de seguro médico *Medicare*. Se consideran 54 Estados que incluyen los 50 Estados de la Unión Americana, el Distrito de Columbia, las Islas Virginia, Puerto Rico y otro Estado no especificado. La variable respuesta es el monto de indemnizaciones cubiertas (en dólares) por cada paciente dado de alta. Esta respuesta es de sumo interés para los actuarios en Estados Unidos de América porque el programa *Medicare* hace los reembolsos correspondientes a los hospitales en base a los días que el paciente estuvo hospitalizado. Las variables explicativas de este estudio son: Tiempo, medido en años; el número de pacientes dados de alta en el año correspondiente y la estancia promedio de hospitalización, medida en días por paciente. El objetivo es averiguar cuáles de las variables explicativas, son significativas en el modelo de regresión para estos datos, y a partir de su determinación hacer el mejor pronóstico de los comportamientos futuros de las indemnizaciones de este programa de seguros.

1.4. Notación

Para agilizar y hacer entendibles los modelos de datos longitudinales a exponer, se presenta una descripción de la notación usada.

La variable respuesta es representada por Y_{ij} y $\mathbf{x}_{ij}$ denota un vector de longitud p de variables explicativas observadas al tiempo t_{ij} para la observación $j = 1, \ldots, n_i$ en el sujeto $i = 1, \ldots, m$. La media y la varianza de Y_{ij} son representadas por $\mathrm{E}(Y_{ij}) = \mu_{ij}$ y $\mathrm{var}(Y_{ij}) = v_{ij}$. El conjunto de observaciones para el sujeto i se escribe como $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, con media $\mathrm{E}(\mathbf{Y}_i) = \mu_i$ y matriz de covarianza $\mathrm{var}(\mathbf{Y}_i) = V_i$ donde la entrada j, k de V_i es la covarianza entre Y_{ij} y Y_{ik}. R_i es la matriz de correlación de $\mathbf{Y}_i$. El vector de respuestas para todos los individuos es

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}.$$

Recuérdese que la unidad experimental no es la medición individual Y_{ij}, sino la secuencia, $\mathbf{Y}_i$, de medidas en un sujeto.

1.5. Organización de los capítulos subsecuentes

Una vez que se ha dado una breve descripción de las ideas que se encierran en el análisis de datos longitudinales, el siguiente capítulo presenta los diferentes enfoques para modelar datos longitudinales, como son: los modelos lineales generalizados (como punto de partida para los demás desarrollos), los modelos marginales, los modelos de efectos aleatorios y los modelos de transición. El capítulo 3 trata sobre los métodos de estimación, con énfasis en su aplicación a los modelos antes expuestos. En el capítulo 4 se ilustran las metodologías desarrolladas mediante aplicaciones en diferentes ramas de la ciencia. Finalmente, se presentan las conclusiones del estudio, subrayando las metas alcanzadas.

Capítulo 2

Modelos para datos longitudinales

En este capítulo se describen en forma general las principales metodologías usadas para el análisis de datos longitudinales. Con le fin de poner estas metodologías en un marco general, se da una breve descripción de los modelos lineales generalizados pues son la base del desarrollo de los esquemas de análisis de datos longitudinales.

2.1. Modelos lineales generalizados

El término de *modelo lineal generalizado* o MLG se debe a Nelder y Wedderburn [54] quienes describieron cómo la linealidad podría ser explotada para unificar varios modelos de regresión aparentemente distintos (e.g. modelos para respuestas discretas y modelos para respuestas continuas) mediante alguna transformación.

Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes con medias $\mu_1, \ldots, \mu_n$ respectivamente, y $y_1, \ldots, y_n$ un conjunto de observaciones donde y_i es una realización de la variable aleatoria Y_i. Dichas observaciones representan las respuestas de una muestra. Sean $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vectores de dimensión p que son los valores de las distintas variables explicativas para cada una de las respuestas de la muestra. Sea $\beta = (\beta_1, \ldots, \beta_p)^T$ un vector de parámetros desconocidos que relaciona a las variables explicativas con la respuesta. En el modelo lineal normal la relación entre las variables explicativas y la respuesta

se puede expresar como

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i = \mathbf{x}_i^T\beta + \epsilon_i, \quad i = 1, \ldots, n$$

En notación matricial, se tiene que

$$\mathbf{Y} = \mathbf{x}\beta + \epsilon$$

con $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ y $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$.

Existen tres especificaciones en un modelo lineal generalizado. Primero, el predictor lineal de un MLG, denotado como η_i, es de la forma

$$\eta_i = \mathbf{x}_i^T\beta. \tag{2.1}$$

Segundo, se especifica una función $g(\cdot)$ monótona creciente (por lo tanto invertible) conocida como la **función liga** la cual relaciona el valor esperado μ_i de la respuesta en el predictor lineal η_i

$$g(\mu_i) = \eta_i. \tag{2.2}$$

La función liga determina la escala en la cual se asume la linealidad y mapea cualquier rango de los parámetros del modelo en estudio hacia el intervalo $(-\infty, \infty)$; la elección de dicha función se determina por el ajuste que haga a los datos, la facilidad en la interpretación de los parámetros en el predictor lineal y por la existencia de estadísticos suficientes en la familia exponencial. Finalmente se hace una especificación para la forma de la varianza en términos de la media μ_i.

La especificación de la función liga y de la estructura para la varianza usualmente depende de la distribución de la respuesta, la cual se asume que pertenece a la familia exponencial de distribuciones. Las distribuciones que pertenecen a esta familia se caracterizan porque sus funciones de densidad son de la forma

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\} \tag{2.3}$$

donde θ_i es llamado parámetro natural, ϕ es el parámetro de dispersión y $b(\cdot)$ y $c(\cdot)$ son funciones específicas correspondientes al tipo de distribución.

El parámetro natural θ es una función de la media, esto es, $\theta_i = \theta(\mu_i)$ la cual es determinada por medio de la relación $\mu = b'(\theta)$. Más aún, la varianza de y es de la forma

$$\text{var}(y_i) = \phi v(\mu_i)$$

donde la *función varianza* $v(\mu)$ es determinada por la relación $v(\mu) = b''(\theta)$.

La elección de la función liga apropiada depende del tipo de respuesta y de la aplicación particular. Para cada familia exponencial existe una liga particular llamada la *liga canónica* la cual relaciona directamente al parámetro natural con el predictor lineal

$$\theta = \theta(\mu) = \eta = \mathbf{x}^T\beta,$$

esto es, $g(\mu) \equiv \theta(\mu)$.

Algunas distribuciones perteneciente a la familia exponencial junto con la distribución normal son la distribución binomial, la Poisson y la gamma. Las características de estas distribuciones en términos de los elementos de (2.3) están expresados en la tabla 2.1.

(a)

Distribución		$\theta(\mu)$	$b(\theta)$	ϕ
Normal	$N(\mu, \sigma^2)$	μ	$\sigma^2/2$	σ^2
Bernoulli	$B(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Poisson	$P(\lambda)$	$\log\lambda$	$\exp(\theta)$	1
Gamma	$G(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}

(b)

Distribución	$E(y) = b'(\theta)$	$v(\mu) = b''(\theta)$	$var(y) = b''(\theta)\phi$
Normal	$\mu = \theta$	1	σ^2
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)$
Poisson	$\lambda = \exp(\theta)$	λ	λ
Gamma	$\mu = -1/\theta$	μ^2	$\mu^2\nu^{-1}$

Cuadro 2.1: Algunas distribuciones de la familia exponencial: (a) Parámetro natural, función acumulada y parámetro de dispersión; (b) Esperanza, función varianza y varianza

En lo subsiguiente de este trabajo, se considerarán a los modelos y metodologías presentadas en base al enfoque de los modelos lineales generalizados que se acaban de exponer.

2.2. Modelos marginales

En muchos estudios longitudinales el objetivo principal es analizar la *esperanza marginal* de las respuestas en términos de las variables explicativas; entiéndase como esperanza marginal la respuesta promedio de la muestra estudiada que comparten un mismo valor del vector de variables explicativas. En este sentido, los parámetros de regresión marginales tienen la misma interpretación que los parámetros de regresión de los modelos de corte transversal, con el extra de que en los modelos marginales es necesario tomar en cuenta la correlación existente entre las respuestas.

Para la regresión se modela la esperanza marginal de la respuesta, $\mathrm{E}(Y_{ij})$ como una función de las variables explicativas, considerando las siguientes suposiciones.

- La esperanza marginal de la respuesta, $\mathrm{E}(Y_{ij}) = \mu_{ij}$, depende de las variables explicativas, $\mathbf{x}_{ij}$, de la forma $g(\mu_{ij}) = \mathbf{x}_{ij}^T\beta$, como lo indica la ecuación (2.2) donde g es una función liga conocida;

- la varianza marginal depende de la media marginal: $\mathrm{var}(Y_{ij}) = v(\mu_{ij})\phi$ donde v es alguna función de varianza y ϕ es un parámetro de dispersión, el cual podría necesitar ser estimado;

- la correlación entre Y_{ij} y Y_{ik} es una función de las medias marginales y tal vez de un vector de parámetros adicionales α,

$$\mathrm{cor}(Y_{ij}, Y_{ik}) = c(\mu_{ij}, \mu_{ik}; \alpha)$$

  donde c es una función conocida cuyo rango está en $[-1, 1]$.

Las respuestas correspondientes a diferentes individuos se asumen independientes y los parámetros β y α son los mismos para cada sujeto.

Como una característica importante de los modelos marginales, el vector de regresores β puede estimarse de manera consistente, i. e., el vector de estimadores $\hat{\beta}$ se acerca cada vez más al vector de parámetros "reales" β

conforme el tamaño de muestra se va haciendo cada vez más grande, siempre y cuando la función de correlación sea especificada correctamente. En la literatura, esta función de correlación se conoce como una *correlación de trabajo (working correlation)* para la asociación entre Y_{ij} y Y_{ik}. Particularmente, Liang y Zeger [46] y Prentice [62] usan una *matriz de correlación de trabajo (working correlation matrix)* $R_i(\alpha)$, la cual es simétrica y sus entradas j, k son

$$[R_i]_{jk} = \text{cor}(Y_{ij}, Y_{ik}).$$

Con esta matriz de correlación es posible definir la *matriz de covarianza de trabajo* de la forma

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}/\phi \tag{2.4}$$

donde

$$A_i = \begin{pmatrix} \text{var}(y_{i1}) & 0 & \cdots & 0 \\ 0 & \text{var}(y_{i2}) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \text{var}(y_{in_i}) \end{pmatrix}$$

y $A_i^{1/2}$ es la matriz tal que ${A_i^{1/2}}^T A_i^{1/2} = A_i$. Liang y Zeger [46] sugieren algunas estructuras para $R_i(\alpha)$. La elección más simple es un modelo de independencia, por lo que $\text{cor}(Y_{ij}, Y_{ik}) = 0$ para $j \neq k$, de forma que

$$R_i(\alpha) = \mathbf{I} \tag{2.5}$$

donde $\mathbf{I}$ es la matriz identidad. Otra elección es un modelo de equicorrelación, es decir que $\text{cor}(Y_{ij}, Y_{ik}) = \alpha$ para toda $j \neq k$, así la matriz de correlación es de la forma:

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha \\ \alpha & \cdots & \alpha & 1 \end{pmatrix} \tag{2.6}$$

También se puede considerar una estructura de correlación completamente no especificada, donde $\alpha_{j,k} = \text{cor}(Y_{ij}, Y_{ik}), \quad j < k$ para la correspondiente matriz R_i, y dado que $\text{cor}(Y_{ij}, Y_{ik}) = \text{cor}(Y_{ik}, Y_{ij})$ la matriz de correlación

puede expresarse como

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha_{1,2} & \cdots & \alpha_{1,n_i} \\ \alpha_{1,2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_{n_i-1,n_i-1} \\ \alpha_{1,n_i} & \cdots & \alpha_{n_i-1,n_i-1} & 1 \end{pmatrix}. \tag{2.7}$$

Otra estructura de correlación es el caso donde se considera una correlación común entre una observación y la siguiente únicamente. En este sentido se tiene

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & 0 & \cdots & 0 \\ \alpha & 1 & \alpha & \ddots & \vdots \\ 0 & \alpha & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha \\ 0 & \cdots & 0 & \alpha & 1 \end{pmatrix}$$

Este es el modelo uni-dependiente de correlación, el cual puede extenderse fácilmente al modelo k-dependiente en el que se consideran las correlaciones entre una observación y las siguientes k observaciones.

Además de las estructuras mencionadas antes, se puede considerar una estructura de autocorrelación de primer orden cuando las respuestas son Gaussianas. La especificación de las entradas de la matriz R es de la forma

$$\text{cor}(Y_{ij}, Y_{ik}) = \alpha^{|t_{ij}-t_{ik}|}.$$

La eficiencia de la estimación depende del diseño de la correlación y de las variables explicativas. Fitzmaurice [20] demostró que la suposición de independencia puede llevar a una pérdida de eficiencia considerable, cuando las respuestas están fuertemente correlacionadas y el diseño incluye al menos una variable explicativa entre los individuos, pues el coeficiente de regresión asociado a dicha variable podría ser estimado ineficientemente.

Se pueden considerar distintos modelos. Los siguientes son ejemplos de modelos marginales.

Ejemplo 2.1. *Respuestas continuas*

- $\mu_{ij} = \mathrm{E}(Y_{ij}) = \mathbf{x}_{ij}^T\beta$,
- $\mathrm{var}(Y_{ij}) = \phi = \sigma^2$,
- Correlación no especificada; esto es la matriz de correlación resulta de las correlaciones entre las respuestas, $\mathrm{cor}(y_{ij}, y_{ik})$ como en (2.7).

Ejemplo 2.2. *Respuestas binarias*

- $\mu_{ij} = \pi_{ij} = \mathrm{P}(Y_{ij} = 1), \mathrm{logit}\ \pi_{ij} = \mathbf{x}_{ij}^T\beta$,
- $\mathrm{var}(Y_{ij}) = \pi_{ij}(1 - \pi_{ij})$,
- Modelo de independencia para correlación.

Ejemplo 2.3. *Datos de conteo*

- $\log(\mu_{ij}) = \mathbf{x}_{ij}^T\beta$,
- $\mathrm{var}(Y_{ij}) = \mu_{ij}\phi$,
- Equicorrelación, cuya matriz es de la forma (2.6).

Un ejemplo práctico que será desarrollado explícitamente más adelante corresponde a un estudio de ataques epilépticos (ejemplo 1.2). En este estudio se observa el número de ataques epilépticos en distintos periodos para 59 pacientes, por lo se puede asumir que las respuestas tienen una distribución Poisson. Se considera entonces un modelo marginal con liga log:

$$\log(\mu_{ij}) = \log(t) + \beta_0 + \beta_1 x + \beta_2 trt + \beta_3 x * trt$$

donde t representa el tamaño del periodo de tiempo en que se hizo la observación, x se refiere al tipo de periodo en que se realizaron las observaciones (0 para el periodo base, 1 para el periodo de tratamiento) y trt identifica el tipo de tratamiento.

Este ejemplo queda a modo para ser modelado con un modelo marginal para datos longitudinales, ya que el objetivo de este estudio es averiguar si

un nuevo medicamento es mejor para el tratamiento de ataques epilépticos comparado con un placebo, sin importar las características particulares de cada paciente. En la siguiente sección se presenta el mismo ejemplo para ilustrar los modelos de efectos aleatorios indicando quienes son las variables adicionales que corresponderán a los efectos aleatorios.

2.3. Modelos de efectos aleatorios

Como se mencionó anteriormente, los estudios longitudinales son diseñados para investigar cambios sobre el tiempo de una característica, la cual es medida repetidamente para cada uno de los participantes del estudio. En estudios médicos, las medidas podrían ser presión sanguínea, nivel de colesterol, volumen pulmonar, etc. No siempre se puede tener control completo de las circunstancias bajo las cuales se toman las medidas, y existe la posibilidad de alguna variación considerable entre los individuos respecto al número y calendarización de las observaciones, por ello se puede presentar una situación en la que el número de observaciones para un individuo es diferente del número de observaciones para otro. En términos matemáticos $n_j \neq n_k$ para algún $j \neq k$ donde n_i es el número de observaciones para el individuo i. En este caso se dice que el conjunto de datos está *desbalanceado*. Otra forma de datos desbalanceados es cuando los intervalos de tiempo entre observaciones para un sujeto es diferente de los intervalos para otro sujeto.

Cuando el objetivo de un estudio longitudinal es analizar las características individuales, la mejor suposición es que algunos parámetros de regresión, como la ordenada y el efecto de una que otra variable explicativa podrían variar de un sujeto a otro, como por ejemplo, el efecto de un tratamiento médico varía aleatoriamente de un individuo a otro.

A menudo, los parámetros individuales tienen una interpretación natural la cual es relevante para los objetivos del estudio, y sus estimaciones pueden ser usadas para análisis exploratorio. Es aquí donde la heterogeneidad de los individuos debe tomarse en cuenta.

Los modelos marginales son a menudo difíciles de aplicar a datos altamente desbalanceados cuando se usan algunas estructuras de correlación como el caso autorregresivo de primer orden, además de que no toman en cuenta efectos particulares de cada individuo; mientras que los modelos de efectos aleatorios pueden ser usados fácilmente pues tienen varias características deseables como poder trabajar con datos desbalanceados, ya sea que éste

desbalance sea en el número de observaciones entre individuos, o bien, en los intervalos entre las observaciones. Otra característica de estos modelos es que permite el modelado explícito y el análisis de variación entre individuos y sobre cada individuo, entre otras.

Los modelos de efectos aleatorios son más usados cuando el objetivo es hacer inferencias sobre los individuos más que sobre la población. En estos modelos se asume que existe un conjunto de variables explicativas que tienen un efecto aleatorio en la respuesta de diferentes individuos, i. e., el efecto de dichas variables, varía aleatoriamente de un individuo a otro.

La especificación general de los modelos de efectos aleatorios en un contexto de modelos lineales generalizados es como sigue:

- Dados $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})$ parámetros de efectos aleatorios, las respuestas $Y_{i1}, \ldots, Y_{in_i}$ son mutuamente independientes y siguen una distribución de la familia exponencial tal que

$$g(\mu_{ij}) = g[\mathrm{E}(Y_{ij}|\mathbf{b}_i)] = \mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i$$

  con $\mathrm{E}(Y_{ij}|\mathbf{b}_i) = \mu_{ij}$ y $\mathrm{var}(Y_{ij}|\mathbf{b}_i) = v(\mu_{ij})\phi$ y donde $\mathbf{x}_{ij}$ y $\mathbf{z}_{ij}$ son vectores de dimensión p y q respectivamente con $p \geq q$ ya que generalmente se asume que $\mathbf{z}_{ij}$ es subvector de $\mathbf{x}_{ij}$.

- Los efectos aleatorios, $\mathbf{b}_i$, $i = 1, \ldots, m$, son mutuamente independientes con una distribución multivariada común, F. Dicha distribución depende de la suposición que se hace respecto de la distribución de la respuesta. Por ejemplo, cuando se asume que la respuesta sigue una distribución Poisson, lo más adecuado es suponer una distribución Gamma para los efectos aleatorios. Sin embargo, los paquetes existentes en R únicamente admiten la distribución Gaussiana para los efectos aleatorios con media $\mathbf{0}$ y matriz de covarianza $\mathbf{D}$, la cual es desconocida. En lo subsiguiente se asumirá dicha distribución para los efectos aleatorios.

Ejemplo 2.4. *Modelo logarítmico para respuestas de conteo*

En el estudio de ataques epilépticos (ejemplo 1.2) se observa el número de ataques epilépticos en distintos periodos, por lo que se asume que las respuestas tienen una distribución Poisson. Considérese entonces el modelo con liga log

$$\log \mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2} + b_{i1} + b_{i2} x_{ij2} + \log(t_{ij})$$

para $i = 1, \ldots, 59, \quad j = 0, 1, \ldots, 4$, donde los efectos fijos son

$$x_{ij1} = \begin{cases} 1 \text{ si el } i\text{-ésimo sujeto es asignado al grupo de progabide} \\ 0 \text{ si el } i\text{-ésimosujeto es asignado al grupo de placebo} \end{cases}$$

$$x_{ij2} = \begin{cases} 1 \text{ si la observación es durante el tratamiento, i.e., } j = 1, 2, 3, 4 \\ 0 \text{ si la observación es del periodo base, i.e., } j = 0 \end{cases}$$

En el caso de los efectos aleatorios b_{i1} es la ordenada y b_{i2} es otro efecto aleatorio para considerar la heterogeneidad entre sujetos en la proporción de ataques esperados antes y después de la selección de grupo de tratamiento. Se asume que $\mathbf{b}_i = (b_{i1}, b_{i2})^T$ sigue una distribución gaussiana con media $(0, 0)^T$ y matriz de covarianza

$$\mathbf{D} = \begin{pmatrix} \text{var}(b_{i1}) & \text{cov}(b_{i1}, b_{i2}) \\ \text{cov}(b_{i1}, b_{i2}) & \text{var}(b_{i2}) \end{pmatrix}$$

En resumen, la idea básica subyacente a un modelo de efectos aleatorios es que existe heterogeneidad entre todos los individuos en sus coeficientes de regresión, y que esta heterogeneidad puede ser representada por una distribución de probabilidad.

El método de estimación de parámetros que será usado más adelante en éste trabajo para ajustar un modelo de efectos aleatorios a un conjunto de datos longitudinales es el método de máxima verosimilitud. Bajo tal enfoque, se considera a los efectos aleatorios como un conjunto de variables no observables las cuales se integran fuera de la verosimilitud. Este proceso de estimación será tratado a detalle en el siguiente capítulo bajo la suposición de que la distribución de los efectos aleatorios es Gaussiana con media $\mathbf{0}$ y matriz de covarianza $\mathbf{D}$.

2.4. Modelos de transición

Los modelos de transición consideran la información de las variables explicativas, así como de las observaciones pasadas en la distribución condicional de la respuesta Y_{ij}.

Para especificar el modelo general de transición, sea

$$H_{ij} = \{y_{i1}, \ldots, y_{ij-1}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij-1}, \mathbf{x}_{ij}\}$$

la historia de las respuestas pasadas y el pasado y presente de las variables explicativas del i-ésimo sujeto.

Los modelos autorregresivos generalizados están caracterizados por la siguiente estructura, análoga a la de los modelos lineales generalizados para datos independientes

- Las densidades condicionales $f(y_{ij}|H_{ij}), j = 1, 2, \ldots$ son del tipo de la familia exponencial.
- La esperanza condicional en base al modelo de Zeger y Qaqish [83] es de la forma
$$g(\mu_{ij}) = \mathbf{x}_{ij}^T\beta + \sum_{r=1}^{q} f_r(H_{ij}; \alpha)$$
donde $\mu_{ij} = \mathrm{E}(Y_{ij}|H_{ij})$ y las f_r son funciones conocidas, α es un vector de coeficientes de regresión adicionales a β.
- $\mathrm{var}(Y_{ij}|H_{ij}) = v_{ij} = v(\mu_{ij})\phi$.

Las f_r transforman la historia pasada para que esta información sea tratada como variables explicativas adicionales. Sin embargo, en la mayoría de experimentos, sólo se incluye un número finito de observaciones pasadas, $y_{ij-1}, \ldots, y_{ij-q}$. A éste se le llama *modelo autorregresivo generalizado* o *modelo de Markov de orden q*.

A menudo, el estudio de datos longitudinales mediante modelos de transición se basa en el ajuste de un modelo de Markov multi-estados en tiempo continuo. Kalbfleisch y Lawless [38] proponen algoritmos para la estimación por máxima verosimilitud de los parámetros de regresión de este tipo de modelos y presentan algunos ejemplos. La idea básica de un modelo de Markov multi-estados es como sigue:

Supóngase que los individuos pueden encontrarse en cualquiera de k estados posibles. Entiéndase por estado un nivel o categoría en que se clasifica la respuesta observada de los individuos estudiados, por ejemplo, enfermo y sano ($k = 2$) o sano, enfermo leve y enfermo grave ($k = 3$).

Sea $Y(t)$ el estado ocupado al tiempo t por un individuo elegido al azar. Entonces se puede definir una *matriz de probabilidad de transición* $P(s, t)$ de tamaño $k \times k$ con entradas

$$p_{uv}(s, t) = \mathrm{P}[Y(t) = v|Y(s) = u] \tag{2.8}$$

para $u, v = 1, \ldots, k$. Este proceso puede ser especificado en términos de las intensidades de transición definidas como:

$$q_{uv}(t) = \lim_{\Delta t \to 0} p_{uv}(t, t + \Delta t)/\Delta t, \quad u \neq v. \tag{2.9}$$

Por conveniencia también se define $q_{uu}(t) = -\sum_{v \neq u} q_{uv}(t)$ para $u = 1, \ldots, k$ y se tiene entonces $Q(t)$ que es la *matriz de intensidades de transición* con entradas $q_{uv}(t)$. Obsérvese que $q_{uv} \geq 0$ para todo $u \neq v$. Cuando ninguna de las intensidades de transición depende del tiempo, el modelo es llamado *homogéneo en el tiempo.*

La estructura de la matriz Q depende de las restricciones o características del proceso estudiado, por ejemplo, si se consideran tres estados: sano(1), enfermo(2) y muerto(3), la matriz de intensidad quedaría especificada de la siguiente manera:

$$Q = \begin{pmatrix} -(q_{12} + q_{13}) & q_{12} & q_{13} \\ q_{21} & -(q_{21} + q_{23}) & q_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

Ajustar el modelo multi-estados a los datos es el proceso de encontrar valores para las intensidades q_{uv} que describan mejor el comportamiento de estos. Véase Kalbfleisch y Lawless [38] para los métodos de estimación.

El siguiente ejemplo representa una pequeña ilustración general de las ideas expuestas.

Ejemplo 2.5. *Modelo logístico*

Un modelo de regresión logística para respuesta binarias que constituye un modelo de Markov de primer orden ([12],[43],[82]) es

$$\text{logitP}(Y_{ij} = 1 | H_{ij}) = \mathbf{x}_{ij}^T \beta + \alpha y_{ij-1} \tag{2.10}$$

En este caso

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right), \qquad v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$$

y

$$f_1(H_{ij}; \alpha) = \alpha y_{ij-1}$$

el proceso es descrito por la matriz de transición

$$\begin{pmatrix} \frac{1}{1+\exp(\mathbf{x}_{ij}^T\beta)} & \frac{\exp(\mathbf{x}_{ij}^T\beta)}{1+\exp(\mathbf{x}_{ij}^T\beta)} \\ \\ \frac{1}{1+\exp(\mathbf{x}_{ij}^T\beta+\alpha)} & \frac{\exp(\mathbf{x}_{ij}^T\beta+\alpha)}{1+\exp(\mathbf{x}_{ij}^T\beta+\alpha)} \end{pmatrix}$$

Una extensión a un modelo de orden q tiene la forma

$$\text{logitP}(Y_{ij} = 1|H_{ij}) = \mathbf{x}_{ij}^T\beta + \sum_{r=1}^{q} \alpha_r y_{ij-r}.$$

Cuando se trata de datos de conteo se asume un modelo logarítmico donde Y_{ij} dado H_{ij} sigue una distribución Poisson. Zeger y Qaqish [83] discuten un modelo de Markov de primer orden con $f_1 = \alpha(\log(y^*_{ij-1} - \mathbf{x}_{ij-1}^T\beta)$, donde $y^*_{ij} = \text{máx}(y_{ij}, c), 0 < c < 1$. Esto lleva a

$$\mu_{ij} = \exp(\mathbf{x}_{ij}^T\beta) \left(\frac{y^*_{ij-1}}{\exp(\mathbf{x}_{ij-1}^T\beta)} \right)^{\alpha}.$$

La constante c previene que $y_{ij-1} = 0$ sea un estado de absorción, esto es, que $y_{ij-1} = 0$ force a que todas las respuestas futuras sean cero. Cuando $\alpha > 0$ y $y_{ij-1} > \exp(\mathbf{x}_{ij-1}^T\beta)$ existe una esperanza condicional incrementada. En cambio, si $\alpha < 0$, un valor muy alto al tiempo $j-1$ provoca un valor muy bajo al tiempo j.

Es común encontrar en la literatura modelos de regresión para cadenas de Markov binarias ([43], [82]); para datos de conteo, Wong [73] y Zeger y Qaqish [83] realizan análisis detallados y presentan algunos ejemplos.

Se han presentado los principales modelos para el estudio de datos longitudinales. Más adelante se observará el poder y alcances que tiene cada uno de ellos en el ajuste de datos longitudinales provenientes de problemas reales.

Capítulo 3

Inferencia

En este capítulo se revisarán distintos métodos usados para la estimación de parámetros de regresión de los modelos expuestos en el capítulo anterior, resaltando el enfoque a los modelos lineales generalizados, para después extender estos métodos a los modelos particulares de datos longitudinales.

3.1. Máxima verosimilitud

El análisis de regresión con modelos lineales generalizados está basado en verosimilitud. La contribución de la i-ésima observación a la función de verosimilitud en un modelo lineal generalizado es como en la ecuación (2.3). De acuerdo con esto, la contribución de log-verosimilitud de la observación y_i es

$$l_i(\theta_i) \propto \log\ f(y_i|\theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{\phi}$$

La función $c(y_i, \phi)$ se omite por no contener el parámetro de interés θ.

Dado que $\theta_i = \theta(\mu_i)$, se puede escribir cada contribución como

$$l_i(\mu_i) = \frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{\phi}$$

De igual manera, al establecer la relación $h(\cdot) = g^{-1}(\cdot)$ y considerando la estructura $\mu_i = h(\mathbf{x}_i^T\beta)$ se tiene

$$l_i(\beta) = l_i(h(\mathbf{x}_i^T\beta))$$

como una función de β. Ya que se considera que las observaciones $\mathbf{y}$ son independientes, la log-verosimilitud de la muestra es la suma de las contribuciones individuales:

$$l(\beta) = \sum_i l_i(\beta) \tag{3.1}$$

El estimador de máxima verosimilitud de β es el valor $\hat{\beta}$, el cual maximiza la función de log-verosimilitud (3.1). Esto es, para cualquier valor de β,

$$l(\beta) \leq l(\hat{\beta})$$

Usualmente el estimador $\hat{\beta}$ es obtenido por maximización directa de $l(\beta)$, diferenciando la función log-verosimilitud respecto a β y resolviendo el conjunto de ecuaciones

$$\mathbf{S}(\beta) = \frac{\partial l}{\partial \beta} = 0 \tag{3.2}$$

La función $\mathbf{S}(\beta)$ es conocida como la *función score* para β. Para comprobar que la solución corresponde a un máximo de $l(\beta)$ se verifica que la matriz de información observada

$$\mathbf{i}_{obs}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} \quad j, k = 1, \ldots, p$$

evaluada en $\beta = \hat{\beta}$, es negativa definida.

Una propiedad importante de los estimadores de máxima verosimilitud es que si $g(\beta)$ es cualquier función del vector de parámetros β, entonces el estimador de máxima verosimilitud de $g(\beta)$ es $g(\hat{\beta})$. Esta es llamada la *propiedad de invarianza* de los estimadores de máxima verosimilitud.

Otras propiedades de estos estimadores son las llamadas *propiedades asintóticas*. Bajo algunas condiciones de regularidad, se tienen las siguientes propiedades:

Existencia y unicidad asintótica La probabilidad de que $\hat{\beta}$ exista y sea único (localmente) tiende a 1 cuando $n \to \infty$.

Consistencia Si β denota el valor verdadero, entonces cuando $n \to \infty$ se tiene que $\hat{\beta} \to \beta$ en probabilidad (consistencia débil) o con probabilidad 1 (consistencia fuerte).

Normalidad asintótica La distribución del estimador de máxima verosimilitud se convierte en normal cuando $n \to \infty$, o, dicho de otra forma, para n muy grande

$$\hat{\beta} \sim N(\beta, \mathbf{i}^{-1}(\hat{\beta}))$$

La matriz $\mathbf{i}$ es conocida como la *matriz de información de Fisher* e $\mathbf{i}^{-1}$ es la matriz de covarianza de la distribución asintótica para β. La matriz de Fisher y la matriz de información observada están relacionadas por

$$\mathbf{i}(\beta) = \mathrm{E}(\mathbf{i}_{obs}(\beta)).$$

Además, el estimador de máxima verosimilitud es asintóticamente eficiente comparado con muchos otros estimadores, es decir, tiene la mínima varianza posible.

El siguiente ejemplo es un caso en el que los estimadores de máxima verosimilitud pueden obtenerse de manera explícita. Sin embargo, muy frecuentemente se requieren métodos numéricos para evaluar los estimadores de máxima verosimilitud.

Ejemplo 3.1 (Diggle, Liang y Zeger [14])**.**

Considérese que Y_1 y Y_2 son dos muestras binomiales independientes con índices y probabilidades (n_1, p_1) y (n_2, p_2), respectivamente. Un ejemplo típico para el cual este ajuste es apropiado es un experimento clínico donde se comparan dos tratamientos. Aquí, Y_i denota el número de pacientes que respondió negativamente al tratamiento $i = 1, 2$ al cual se le asignaron n_i sujetos, y p_i es la probabilidad correspondiente de respuestas negativas. Por conveniencia, se transforman los parámetros (p_1, p_2) a (θ_1, θ_2)como sigue:

$$\theta_1 = \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right), \quad \theta_2 = \log\left(\frac{p_2}{1-p_2}\right)$$

Esto lleva a una función de verosimilitud para $\theta = (\theta_1, \theta_2)$ de la forma

$$\begin{aligned} L(\theta|y_1, y_2) \quad &\propto \quad p_1^{y_1}(1-p_1)^{n_1-y_1} p_2^{y_2}(1-p_2)^{n_2-y_2} = \\ &= \left(\frac{p_1}{1-p_1}\right)^{y_1}\left(\frac{p_2}{1-p_2}\right)^{y_2}(1-p_1)^{n_1}(1-p_2)^{n_2} \\ &= \exp\left[\theta_1 y_1 + \theta_2 y_1 + \theta_2 y_2 - n_1 \log(1+e^{\theta_1+\theta_2}) - n_2 \log(1+e^{\theta_2})\right] \end{aligned}$$

El parámetro θ_1 es llamado la *log-razón de probabilidad*. Un valor cero de θ_1 denota igualdad de p_1 y p_2. El estimador de máxima verosimilitud, $\hat{\theta}$, puede ser derivado como la solución del par de ecuaciones

$$y_1 - \frac{n_1 \exp(\theta_1 + \theta_2)}{1 + \exp(\theta_1 + \theta_2)} = y_1 - n_1 p_1 = 0$$

y

$$y_1 + y_2 - \frac{n_1 \exp(\theta_1 + \theta_2)}{1 + \exp(\theta_1 + \theta_2)} - \frac{n_2 \exp(\theta_2)}{1 + \exp(\theta_2)} = y_1 + y_2 - n_1 p_1 - n_2 p_2 = 0$$

Lo que lleva a

$$\hat{\theta}_1 = \log\left(\frac{y_1(n_2 - y_2)}{y_2(n_1 - y_1)}\right), \quad \hat{\theta}_2 = \log\Big(y_2/(n_2 - y_2)\Big)$$

La matriz de información de Fisher para θ puede obtenerse algebraicamente, y está expresada por:

$$\mathbf{i}^{-1} = \begin{pmatrix} \frac{n_1 \exp(\theta_1+\theta_2)}{[1+\exp(\theta_1+\theta_2)]^2} & \frac{n_1 \exp(\theta_1+\theta_2)}{[1+\exp(\theta_1+\theta_2)]^2} \\ \\ \frac{n_1 \exp(\theta_1+\theta_2)}{[1+\exp(\theta_1+\theta_2)]^2} & \frac{n_1 \exp(\theta_1+\theta_2)}{[1+\exp(\theta_1+\theta_2)]^2} + \frac{n_2 \exp(\theta_2)}{[1+\exp(\theta_2)]^2} \end{pmatrix}$$

Las ecuaciones de verosimilitud son en general no lineales y tienen que ser resueltas iterativamente. El proceso de iteración mas usado es el de puntuación de Fisher, el cual es una simple modificación del método de Newton-Raphson (Para una revisión más extensa y comparativa de éstos y otros métodos iterativos, véase por ejemplo [24]). Recuérdese que el método de Newton-Raphson está basado en la expansión en series de Taylor de la función objetivo. De esta forma el proceso iterativo está definido de la siguiente manera:

$$\hat{\beta}_{i+1} = \hat{\beta}_i - \mathbf{S}(\hat{\beta}_i)\mathbf{H}^{-1}(\hat{\beta}_i) \tag{3.3}$$

donde $\mathbf{S}(\beta)$ denota el vector de primeras derivadas de $l(\beta)$ como en (3.2) y $\mathbf{H}(\beta)$ denota la matriz de segundas derivadas (matriz hessiana) de forma que $\mathbf{H}(\beta) = -\mathbf{i}_{obs}(\beta)$.

Al sustituir cada $\mathbf{H}(\hat{\beta}_i)$ por $\mathrm{E}[\mathbf{H}(\hat{\beta}_i)]$ en la ecuación (3.3), observando que $\mathrm{E}[\mathbf{H}(\beta)] = -\mathrm{E}[\mathbf{i}_{obs}(\beta)] = -\mathbf{i}(\beta)$, se tiene la expresión para el proceso iterativo del método de puntuación de Fisher, el cual queda especificado de la siguiente manera:

$$\hat{\beta}_{i+1} = \hat{\beta}_i + \mathbf{S}(\hat{\beta}_i)\mathbf{i}^{-1}(\hat{\beta}_i) \tag{3.4}$$

El método de Fisher tiene dos ventajas sobre el método de Newton-Raphson: primero, sólo es necesario calcular las primeras derivadas de la función de log-verosimilitud, ya que se tiene que

$$\mathrm{E}\left[\frac{\partial^2 l}{\partial\beta_i\partial\beta_j}\right] = -\mathrm{E}\left[\left(\frac{\partial l}{\partial\beta_i}\right)\left(\frac{\partial l}{\partial\beta_j}\right)\right],$$

la segunda ventaja es que se garantiza que $-\mathbf{i}(\beta)$ es positiva definida, eliminando posibles problemas de no-convergencia del método de Newton-Raphson.

Considerando la estructura de los datos longitudinales, se puede reescribir la expresión (3.4) de la siguiente forma:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left(\sum_{i=1}^{m} \hat{\mathbf{D}}_i^{T(k)} \hat{\mathbf{V}}_i^{-1(k)} \hat{\mathbf{D}}_i^{(k)}\right)^{-1} \left(\sum_{i=1}^{m} \hat{\mathbf{D}}_i^{T(k)} \hat{\mathbf{V}}_i^{-1(k)} (\mathbf{y}_i - \hat{\mu}_i^{(k)})\right) \tag{3.5}$$

donde

$$\begin{aligned} \hat{\mathbf{D}}_i^{(k)} &= \mathbf{D}_i(\hat{\beta}^{(k)}) \\ \hat{\mathbf{V}}_i^{(k)} &= \mathbf{V}_i(\hat{\beta}^{(k)}) \\ \hat{\mu}_i^{(k)} &= \mu_i(\hat{\beta}^{(k)}) \end{aligned} \tag{3.6}$$

En esta expresión los componentes de $\mathbf{D}_i$ son $D_{irs} = \partial\mu_r/\partial\beta_s$, $\mathbf{V}_i$ es la matriz de varianza-covarianza del i-ésimo sujeto y μ_i es la media de las observaciones del i-ésimo sujeto.

Obsérvese que para la k-ésima iteración y el i-ésimo individuo

$$\mathbf{i}_i(\hat{\beta}^{(k)}) = \hat{\mathbf{D}}_i^{T(k)} \hat{\mathbf{V}}_i^{-1(k)} \hat{\mathbf{D}}_i^{(k)}$$

así como

$$\mathbf{S}_i(\hat{\beta}^{(k)}) = \hat{\mathbf{D}}_i^{T(k)} \hat{\mathbf{V}}_i^{-1(k)} (\mathbf{y}_i - \hat{\mu}_i^{(k)}).$$

Con lo que se tiene

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left(\sum_{i=1}^{m} \mathbf{i}_i(\hat{\beta}^{(k)})\right)^{-1} \left(\sum_{i=1}^{m} \mathbf{S}_i(\hat{\beta}^{(k)})\right)$$

la cual es una generalización de (3.4) para el ajuste de datos longitudinales por máxima verosimilitud.

Ahora bien, el método de máxima verosimilitud permite obtener estimaciones de los parámetros de regresión β bajo el supuesto de que se conoce el valor del parámetro de dispersión. Sin embargo, cuando éste no se conoce, es necesario también hacer una estimación de él. En el enfoque de los modelos lineales generalizados, un estimador consistente de

$$\phi = \frac{\text{var}(y_i)}{v(\mu_i)}$$

es

$$\tilde{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{3.7}$$

Así, por ejemplo, para n variables aleatorias normalmente distribuidas $Y_1, \ldots, Y_n$ con $\text{E}[Y_i] = \mu_i$ y $\text{var}(Y_i) = \sigma^2$, se tiene el estimador

$$\tilde{\phi} = \hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

3.1.1. Máxima verosimilitud para modelos de efectos aleatorios

En el caso particular en que se usa el método de máxima verosimilitud para el ajuste de modelos de efectos aleatorios de datos longitudinales, la estructura de la función score para los parámetros de regresión tiene características particulares.

Por principio de cuentas, se trata a las variables de efectos aleatorios, $\mathbf{b}_i$, como una muestra de variables independientes no observables con cierta distribución común perteneciente a la familia exponencial. Como se mencionó en el capítulo anterior, por simplicidad se asumirá que el conjunto de variables de efectos aleatorios tiene una distribución Gaussiana con media $\mathbf{0}$ y matriz de covarianza $\mathbf{D}$, la cual se estimará más adelante.

Así pues, la función de verosimilitud está definida como sigue:

$$L(\beta, \mathbf{D}; \mathbf{y}) = \prod_{i=1}^{m}\int\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i;\beta)f(\mathbf{b}_i;\mathbf{D})d\mathbf{b}_i \tag{3.8}$$

Encontrar la estimación de máxima verosimilitud significa resolver las ecuaciones score para β y para $\mathbf{D}$. Bajo el supuesto de que se usan ligas

canónicas, la función score para β tiene la forma particular

$$\mathbf{S}(\beta|\mathbf{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}[y_{ij} - \mathrm{E}(\mu_{ij}(\mathbf{b}_i)|\mathbf{y}_i)] = 0 \tag{3.9}$$

con $\mu_{ij}(\mathbf{b}_i) = g^{-1}(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}\mathbf{b}_i)$. Las ecuaciones score para $\mathbf{D}$ pueden obtenerse de manera similar como:

$$\mathbf{S}(\mathbf{D}|\mathbf{y}) = \frac{1}{2}\mathbf{D}^{-1}\left[\sum_{i=1}^{m}\mathrm{E}(\mathbf{b}_i^T\mathbf{b}_i|\mathbf{y}_i)\right]\mathbf{D}^{-1} - \frac{m}{2}\mathbf{D}^{-1} = 0 \tag{3.10}$$

La estrategia más usada para resolver numéricamente la estimación de máxima verosimilitud para β y $\mathbf{D}$ es usar el algoritmo EM. Este algoritmo itera entre la fase E, en la cual se evalúan las esperanzas en las ecuaciones score anteriores usando valores presentes estimados de los parámetros; y una fase M, en la cual se resuelven las ecuaciones score para actualizar las estimaciones de los parámetros.

Los detalles del algoritmo EM están fuera de los objetivos de este trabajo. Pueden encontrarse investigaciones del tema en Lindstrom y Bates [48] y Diggle, Liang y Zeger [14].

A continuación se revisan otros métodos de estimación, cuyo enfoque puede considerarse como una "reducción" o simplificación de la metodología de máxima verosimilitud al establecer ciertos supuestos sobre el comportamiento de los datos.

3.2. Cuasi-verosimilitud

A menudo aparecen experimentos en los cuales la información existente es insuficiente para construir una función de verosimilitud. Para definir una verosimilitud se tiene que especificar la forma de la distribución de las observaciones.

En el enfoque de *cuasi-verosimilitud* que a continuación se expone sólo es necesario especificar una relación entre la media y la varianza de las observaciones y usar el método de cuasi-verosimilitud en la misma forma que se usa el método de máxima verosimilitud para hacer las estimaciones.

Sea $\mathbf{y} = (y_1, \ldots, y_n)$ un vector de respuestas independientes con vector de medias $\mu = (\mu_1, \ldots, \mu_n)$ y matriz de covarianza $\mathbf{V}(\mu)$. Se asume que

los parámetros de interés $\beta = (\beta_1, \ldots, \beta_p)$ relacionan la dependencia de μ respecto de las variables explicativas $\mathbf{x}$. Ya que los componentes de $\mathbf{y}$ son independientes, $\mathbf{V}(\mu)$ es diagonal con

$$V_{ii}(\mu) = \text{var}(\mu_i), \quad i = 1, \ldots, n$$

donde $\text{var}(\mu_i)$ es una estimación de la varianza de y_i en términos de su media μ_i. En lo subsecuente se escribirá $V(\mu_i)$ para denotar a $\text{var}(\mu_i)$.

Bajo estas suposiciones, se define la función de cuasi-verosimilitud, o más correctamente, la función de log cuasi-verosimilitud $Q(y_i, \mu_i)$ por la relación

$$\frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \tag{3.11}$$

o equivalentemente

$$Q(y_i, \mu_i) = \int^{\mu_i} \frac{y_i - t}{V(\mu_i)} dt \tag{3.12}$$

Esta idea fue propuesta inicialmente por Wedderburn [72] en 1974, y parte de un teorema que aparece en el mismo artículo la cual dice lo siguiente:

Teorema 3.2.1. *Para una observación de y, la función de log verosimilitud l tiene la propiedad*

$$\frac{\partial l}{\partial \mu} = \frac{y - \mu}{V(\mu)},$$

donde $\mu = \text{E}[y]$ y $V(\mu) = \text{var}(y)$, sí y sólo sí la densidad de y con respecto a alguna medida puede ser escrita en la forma $\exp\{y\theta - g(\theta)\}$, donde θ es alguna función de μ.

Este teorema muestra que la función de log verosimilitud es idéntica a la de cuasi-verosimilitud sí y sólo sí pertenece a la familia exponencial. La demostración del teorema se presenta en el artículo de Wedderburn [72] y se omite en este trabajo.

Considérese por el momento, al igual que en el teorema anterior, que se tiene sólo una observación, así que y y μ serán referidos como una observación y su esperanza, respectivamente. El siguiente teorema muestra algunas propiedades de la función Q, las cuales son similares a las que tiene la log verosimilitud.

Teorema 3.2.2. *Sea* $\mathbf{y}$ *una observación y* Q *definida como en (3.11) y supóngase que* μ *es expresada como una función de los parámetros* $\beta_1, \ldots, \beta_p$. *Entonces* Q *tiene las siguientes propiedades:*

1. $\mathrm{E}\left(\dfrac{\partial Q}{\partial \mu}\right) = 0,$
2. $\mathrm{E}\left(\dfrac{\partial Q}{\partial \beta_i}\right) = 0,$
3. $\mathrm{E}\left(\dfrac{\partial Q}{\partial \mu}\right)^2 = -\mathrm{E}\left(\dfrac{\partial^2 Q}{\partial \mu^2}\right) = \dfrac{1}{V(\mu)},$
4. $\mathrm{E}\left(\dfrac{\partial Q}{\partial \beta_i}\dfrac{\partial Q}{\partial \beta_j}\right) = \dfrac{\partial^2 Q}{\partial \beta_i \partial \beta_j} = \dfrac{1}{V(\mu)}\dfrac{\partial \mu}{\partial \beta_i}\dfrac{\partial \mu}{\partial \beta_j}$

Considérese nuevamente el caso de varias observaciones. Ya que se supone independencia entre los componentes de $\mathbf{y}$, la cuasi-verosimilitud para el total de los datos es la suma de las contribuciones individuales

$$Q(\mu, \mathbf{y}) = \sum Q(\mu_i, y_i)$$

Las ecuaciones estimadoras de cuasi-verosimilitud para los parámetros de regresión β pueden escribirse en la forma $\mathbf{U}(\hat{\beta}) = 0$, donde

$$\mathbf{U}(\beta) = \mathbf{D}^T \mathbf{V}^{-1}(\mathbf{Y} - \mu) \tag{3.13}$$

es llamada la función *cuasi-score*. En esta expresión los componentes de $\mathbf{D}$ son $D_{ir} = \partial \mu_i / \partial \beta_r$ y esta matriz tiene dimensión $n \times p$. La matriz de covarianza de $\mathbf{U}(\beta)$ es

$$\mathbf{i}_\beta = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} \tag{3.14}$$

El proceso para obtener los estimadores de cuasi-verosimilitud para el vector de parámetros β comienza con un valor arbitrario $\hat{\beta}^{(0)}$ suficientemente cercano a $\hat{\beta}$ para ser usado en el proceso iterativo

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} + (\hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i)^{-1} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} (\mathbf{y} - \hat{\mu}^{(i)}) \tag{3.15}$$

donde

$$\begin{aligned} \hat{\mathbf{D}}_i &= \mathbf{D}(\hat{\beta}^{(i)}) \\ \hat{\mathbf{V}}_i &= \mathbf{V}(\hat{\beta}^{(i)}) \\ \hat{\mu}^{(i)} &= \mu(\hat{\beta}^{(i)}) \end{aligned} \tag{3.16}$$

y $\hat{\beta}^{(i)}$ es el vector de valores estimados para β en la i-ésima iteración. Este proceso continúa hasta que se cumpla algún criterio de convergencia. Se tiene entonces que el vector de estimadores $\hat{\beta}$ es aproximadamente insesgado y tiene una distribución asintóticamente normal.

3.3. Ecuaciones Estimadoras Generalizadas

Las ecuaciones estimadoras generalizadas (GEE por sus siglas en ingles) son una extensión del enfoque de cuasi-verosimilitud para datos longitudinales bajo el esquema de modelos marginales.

En ausencia de condiciones para establecer una función de verosimilitud, se puede estimar el vector de parámetros β resolviendo un sistema de ecuaciones cuasi-score análogo a (3.13):

$$\mathbf{S}_\beta(\beta, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i) = 0 \tag{3.17}$$

donde $\mathbf{V}_i$ es la matriz de covarianza correspondiente al vector $\mathbf{Y}_i$.

Existe la dificultad adicional de que $\mathbf{S}_\beta$ depende tanto de β como de α ya que $\mathbf{V}_i = \mathbf{V}_i(yi; \alpha, \beta)$ y está definida de igual manera que (2.4). La alternativa en reemplazar α por un estimador consistente $\hat{\alpha}(\hat{\beta})$. Liang y Zeger [46] y Gourieroux et. al. [28] mostraron que la solución de la ecuación resultante es asintóticamente tan eficiente como si α fuera conocida. El proceso iterativo es muy semejante al empleado en las estimaciones de cuasi-verosimilitud como en (3.15).

Dados los estimadores actuales $\hat{\alpha}$ y $\hat{\phi}$, los cuales pueden estimarse inicialmente de los datos en base a la estructura de la matriz de correlación $R(\alpha)$ que se haya elegido, se comienza con una estimación $\hat{\beta}_0$ de los parámetros de regresión para usarse en el proceso iterativo

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left(\sum_{i=1}^{m} \hat{\mathbf{D}}_i^{T(k)} \hat{\mathbf{V}}_i^{-1(k)} \hat{\mathbf{D}}_i^{(k)} \right)^{-1} \left(\sum_{i=1}^{m} \hat{\mathbf{D}}_i^{T(k)} \hat{\mathbf{V}}_i^{-1(k)} (\mathbf{y}_i - \hat{\mu}_i^{(k)}) \right) \tag{3.18}$$

donde el superíndice (k) indica la k-ésima iteración y las matrices $\hat{\mathbf{D}}_i^{(k)}$ y $\hat{\mathbf{V}}_i^{(k)}$ y el vector $\hat{\mu}_i^{(k)}$ se definen de manera semejante a (3.16).

En una iteración dada, los parámetros α y ϕ pueden ser estimados de los

residuales Pearson definidos por

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\left(\hat{\text{var}}(y_{ij})\right)^{1/2}} \tag{3.19}$$

que dependen directamente del valor actual para β. Se puede estimar ϕ por

$$\hat{\phi}^{-1} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{r}_{ij}^2/(N-p)$$

donde $N = \sum n_i$. El enfoque general para estimar $R(\alpha)$ es considerar las entradas

$$\hat{R}_{uv} = \sum_{i=1}^{m} \hat{r}_{iu}\hat{r}_{iv}/(N-p)$$

Estimadores específicos de α dependen de la elección de la estructura de $R(\alpha)$. Obviamente, cuando se trata de una estructura de independencia $\alpha = 0 \ \forall j \neq k$. En el caso de equicorrelación, se tiene que $\alpha = \text{cor}(Y_{ij}, Y_{ik})$ es un valor constante para todo $j \neq k$. Este valor es estimado por

$$\hat{\alpha} = \frac{\phi \sum_{i=1}^{m} \sum_{j>k} \hat{r}_{ij}\hat{r}_{ik}}{\sum_{i=1}^{m} \frac{1}{2} n_i(n_i - 1) - p}$$

Algunos otros ejemplos de estas elecciones se discuten en Liang y Zeger [46].

En general, se puede decir que los tres esquemas o, mejor dicho, los métodos antes expuestos, son semejantes en cuanto a la idea de establecer una función que contenga la información referente a los parámetros de estimación, la cual es necesario maximizar para obtener los mejores estimadores (en el estricto sentido de la palabra) de dichos parámetros que explican el comportamiento de la población en estudio.

La única diferencia radica, como se ha explicado a lo largo de este capítulo, en la forma en que se establece la función 'objetivo'. Para el caso de la función de verosimilitud es necesario, si no es que preciso, definir completamente la función de distribución que describe el comportamiento de los datos; en el caso de cuasi-verosimilitud y GEE's, se propone sólo una relación entre la media y la varianza para definir una función 'objetivo' la cual ha de ser maximizada de manera semejante a como se hace en el caso de verosimilitud.

El siguiente capítulo permite conciliar la teoría expuesta hasta el momento mediante el análisis de datos reales, estableciendo distintos modelos y enfoques adecuados a la estructura de tales datos y experimentando con ellos.

Capítulo 4

Ilustraciones

En este capítulo se hace uso de los modelos de datos longitudinales mediante su implementación en el paquete estadístico de distribución libre R a fin de analizar ejemplos de datos reales y presentar detalladamente los procesos de exploración de datos, estimación y diagnósticos.

4.1. Modelo marginal

Como se mencionó en el capítulo anterior, las GEE son un método general para ajustar modelos marginales a datos longitudinales con la ventaja de que sólo se necesita una especificación de la media marginal para que el estimador de parámetros sea consistente y asintóticamente normal.

El paquete `geepack` para R implementa el enfoque básico de Liang y Zeger [46] de GEE's y algunas extensiones [74]. Dicho paquete contiene una función interfaz llamada `geeglm` la cual está diseñada para ser tan similar como sea posible a la función `glm`, usada para el ajuste de modelos lineales generalizados; el paquete también dispone de un estimador de la varianza *jackknife* como una alternativa al estimador *sandwich* y las variables explicativas pueden ser incorporadas en los parámetros de escala y correlación en una forma similar al modelado de la media.

Para ilustrar el ajuste de modelos marginales a datos longitudinales usando el método de GEE se usarán los datos del ejemplo 1.2 sobre tratamiento de ataques epilépticos. La justificación de por qué se prefiere usar modelos marginales a usar otros modelos es porque el objetivo del estudio es averiguar si el nuevo medicamento progabide reduce efectivamente el número

de ataques epilépticos en pacientes con este padecimiento, en comparación con el uso de un placebo.

Las tablas 4.1 y 4.2 muestran las observaciones de los 59 pacientes considerados en este estudio.

id	y1	y2	y3	y4	trt	base	age
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	36
5	7	18	9	21	0	66	22
6	5	2	8	7	0	27	29
7	6	4	0	2	0	12	31
8	40	20	23	12	0	52	42
9	5	6	6	5	0	23	37
10	14	13	6	0	0	10	28
11	26	12	6	22	0	52	36
12	12	6	8	5	0	33	24
13	4	4	6	2	0	18	23
14	7	9	12	14	0	42	36
15	16	24	10	9	0	87	26
16	11	0	0	5	0	50	26
17	0	0	3	3	0	18	28
18	37	29	28	29	0	111	31
19	3	5	2	5	0	18	32
20	3	0	6	7	0	20	21
21	3	4	3	4	0	12	29
22	3	4	3	4	0	9	21
23	2	3	3	5	0	17	32
24	8	12	2	8	0	28	25
25	18	24	76	25	0	55	30
26	2	1	2	1	0	9	40
27	3	1	4	2	0	10	19
28	13	15	13	12	0	47	22

Cuadro 4.1: Datos de ataques epilépticos del grupo de tratamiento placebo

id	y1	y2	y3	y4	trt	base	age
29	11	14	9	8	1	76	18
30	8	7	9	4	1	38	32
31	0	4	3	0	1	19	20
32	3	6	1	3	1	10	20
33	2	6	7	4	1	19	18
34	4	3	1	3	1	24	24
35	22	17	19	16	1	31	30
36	5	4	7	4	1	14	35
37	2	4	0	4	1	11	57
38	3	7	7	7	1	67	20
39	4	18	2	5	1	41	22
40	2	1	1	0	1	7	28
41	0	2	4	0	1	22	23
42	5	4	0	3	1	13	40
43	11	14	25	15	1	46	43
44	10	5	3	8	1	36	21
45	19	7	6	7	1	38	35
46	1	1	2	4	1	7	25
47	6	10	8	8	1	36	26
48	2	1	0	0	1	11	25
49	102	65	72	63	1	151	22
50	4	3	2	4	1	22	32
51	8	6	5	7	1	42	25
52	1	3	1	5	1	32	35
53	18	11	28	13	1	56	21
54	6	3	4	0	1	24	41
55	3	5	4	3	1	16	32
56	1	23	19	8	1	22	26
57	2	3	0	1	1	25	21
58	0	0	0	0	1	13	36
59	1	4	3	2	1	12	37

Cuadro 4.2: Datos de ataques epilépticos del grupo de tratamiento progabide

La identificación del paciente se representa por `id`, en este caso, el grupo de tratamiento placebo se considera como el primer grupo conteniendo o considerando 28 pacientes mientras que el grupo de tratamiento progabide es el segundo grupo y cuenta con 31 pacientes; cada una de las `y1, y2, y3, y4` corresponde al número de ataques en el periodo bisemanal correspondiente, `trt` es el tipo de tratamiento (0 para placebo, 1 para progabide), `base` es el conteo del número de ataques en el periodo base de ocho semanas y `age` es la edad del paciente en años.

Al reestructurar los datos, se obtiene la tabla 4.3 en la cual se han ordenado las observaciones base y subsecuentes para cada paciente en una estructura longitudinal. Adicionalmente a las variables antes explicadas, aparece la variable `t`, la cual representa la longitud de tiempo en semanas de las observaciones en curso, así como la variable `x` la cual está estructurada como un indicador: si la observación es del periodo base, entonces `x=0`, si es de las observaciones de tratamiento, es decir, mientras se suministran los respectivos medicamentos para cada grupo, entonces `x=1`. De esta forma, se puede ver que si `x=0`, entonces `t=8`, y si `x=1`, se tendrá que `t=2`.

id	trt	age	time	y	t	x
1.0	0	31	0	11	8	0
1.1	0	31	1	5	2	1
1.2	0	31	2	3	2	1
1.3	0	31	3	3	2	1
1.4	0	31	4	3	2	1
2.0	0	30	0	11	8	0
2.1	0	30	1	3	2	1
2.2	0	30	2	5	2	1
2.3	0	30	3	3	2	1
2.4	0	30	4	3	2	1

Cuadro 4.3: Arreglo longitudinal para los dos primeros pacientes en el grupo placebo

Obsérvese la forma en que cambio el número de identificación del paciente. En este caso se tiene un número de la forma `i.j`, con lo cual el renglón correspondiente representa los datos del i-ésimo paciente en la j-ésima observación. Por ejemplo, si `id=1.0` esto indica que se tienen los valores de las distintas variables en estudio para el paciente número 1 en la observación del

periodo base, mientras que `id=1.1` se refiere a los datos del paciente número 1 pero ahora extraídos de la primera observación bisemanal del periodo de tratamiento.

Para iniciar la exploración de los datos, un buen principio es observar la gráfica de serie de tiempo múltiple. En la figura 4.1 se muestran las series de los 59 pacientes del estudio de ataques epilépticos. El conjunto de observaciones de todos los pacientes que se representa por `Tiempo=0` corresponde al periodo base y las demás, cuando `Tiempo=1, 2, 3, 4` son las observaciones correspondientes a cada uno de los periodos de tratamiento para cada paciente considerado en el estudio.

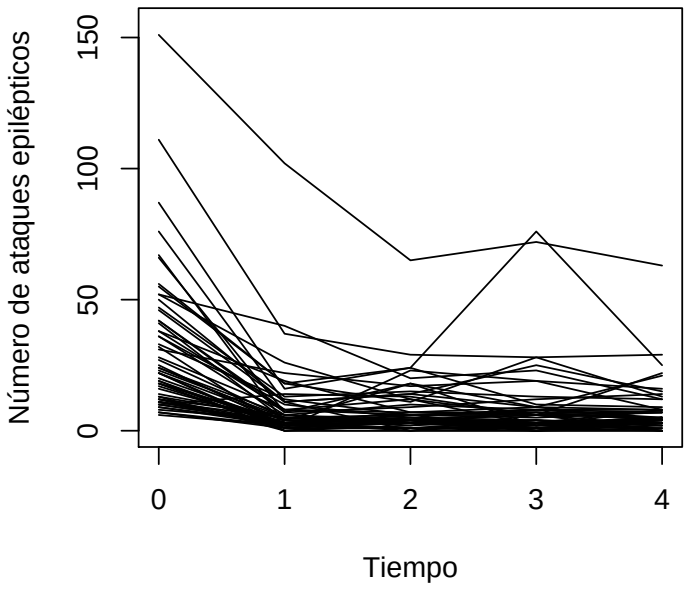
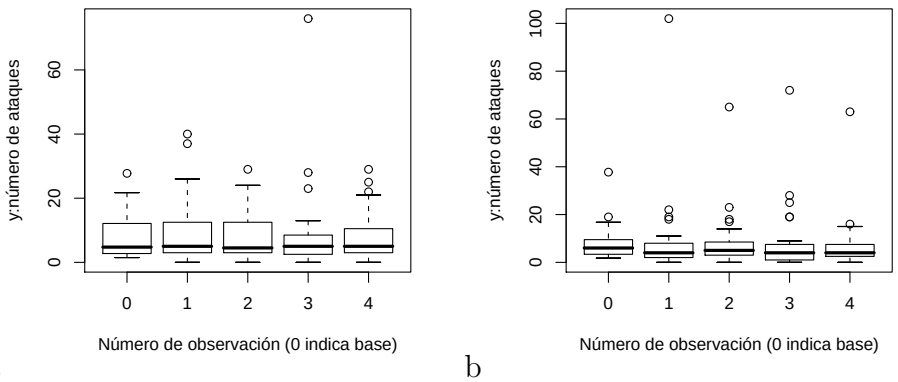


Figura 4.1: Series de tiempo individuales para pacientes con ataques epilépticos

Se observa que los conteos de ataques epilépticos de los pacientes estudiados siguen un comportamiento muy parecido, hasta cierta forma homogéneo, con dos excepciones a las que es necesario analizar: uno de los pacientes pre-

senta un 'pico' del número de ataques en la tercera observación de tratamiento y otro paciente tiene un comportamiento 'sospechosamente' alto en todas sus observaciones pues el conteo de ataques epiléptico en todas y cada una de las observaciones es muy alto comparado con los conteos obtenidos de otros pacientes. Este último paciente, pertenece al grupo de tratamiento progabide y está identificado como el paciente número 49. Más adelante se tratará por separado el caso en el que no se toma en cuenta a este paciente a fin de observar la influencia que tiene en los resultados de los ajustes que se hagan.

Ya que objetivo principal de este estudio es averiguar si el progabide reduce el índice de ataques epilépticos comparado con un placebo, la exploración de la distribución de observaciones para cada tratamiento dará indicio de la eficacia de éste fármaco. En la figura 4.2 se muestra un *boxplot* que compara los dos tratamientos durante el tiempo que duró el estudio. Para el periodo base se dividió el número de ataques entre 4 a fin de obtener el número de ataques promedio bisemanal de ese periodo para poder compararlo con los promedios de las observaciones bisemanales del periodo de tratamiento. Este gráfico sugiere una muy pequeña reducción en el número promedio de ataques epilépticos para el grupo de tratamiento progabide mientras que para el grupo de tratamiento placebo no parece existir reducción alguna en el promedio de episodios de epilepsia.



a b

Figura 4.2: Boxplot para el estudio de ataques epilépticos: (a) placebo, (b) progabide

Para el ajuste del modelo marginal, la variables explicativas que obviamente debe incluirse en el modelo es la variable tratamiento `trt`. Otra de estas variables es la correspondiente al número ordinal de observación, en este caso `time`. Sin embargo, la figura 4.3 muestra que ésta variable no es significativa para el modelo pues se observa que el promedio de ataques por periodo bisemanal (nuevamente se dividió el número de ataques del periodo base entre 4) permanece constante cuando se considera conjuntamente a los dos grupos de estudio.

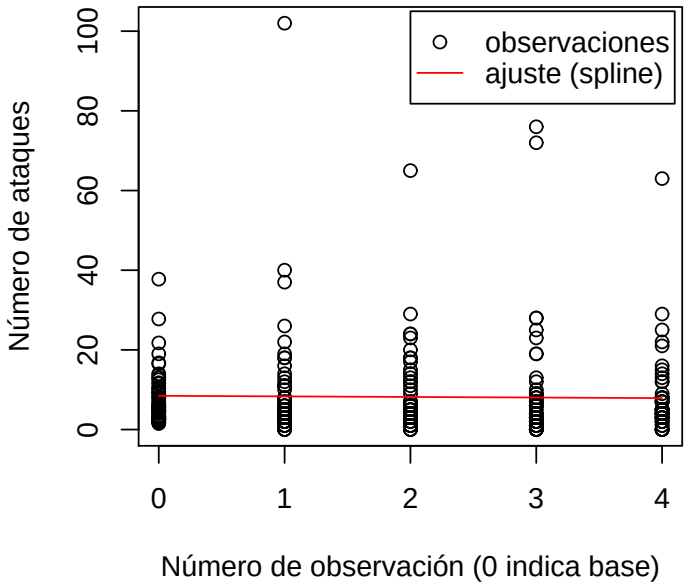


Figura 4.3: Ajuste *spline* de los datos de epilepsia

Ahora bien, cuando se consideran los dos grupos de manera separada y a cada uno se le ajusta una curva "spline"se observa de manera semejante que en la gráfica anterior, que el promedio de ataques por periodo bisemanal permanece constante en cada uno de los grupos de tratamiento, más aun, el comportamiento de los dos grupos es semejante como se observa en la siguiente gráfica:

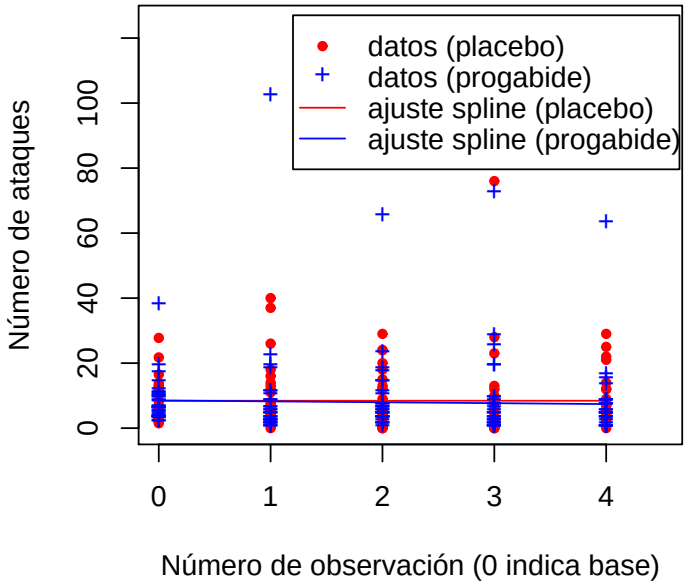


Figura 4.4: Ajuste *spline* para cada grupo de tratamiento

En las dos gráficas anteriores se muestra un ajuste por medio de *splines* para el número de ataques epilépticos en términos de las observaciones subsecuentes. Dicho ajuste define que para el periodo de tratamiento el número de ataques permanece constante, con lo que se concluye que la variable que sí debe incluirse es la `x` que es una variable dicotómica explicada antes. Otra variables explicativas considerada en el ajuste es la interacción `x:trt`, es decir, considerar el comportamiento pre y post-tratamiento para cada uno de los tratamientos. Un valor negativo de este coeficiente corresponde a una reducción significativa en el conteo de ataques para el grupo de progabide. Entonces ésta interacción se convierte en la variable explicativa más importante a considerar en el análisis del modelo.

Un ajuste preliminar (sólo para observar qué sucede si se consideran los datos como si fueran todos independientes) usando modelos lineales generalizados se expresa de la siguiente forma:

```
s <- glm(y ~ offset(log(t)) + x + trt + x:trt,
         data = seiz.l, family = poisson)
```

El término `log(t)` es necesario para tomar en cuenta los diferentes periodos de observación, ya que la variable `t` indica la longitud del periodo de tiempo considerado: 8 para el periodo base de ocho semanas, 2 para cada uno de los cuatro periodos bisemanales de tratamiento. Un `offset` es un término agregado a un predictor lineal con coeficiente conocido igual a 1. Se asume una distribución Poisson para las respuestas, ya que son datos de conteo. El ajuste del modelo anterior se realiza por máxima verosimilitud y los valores de los estimadores de regresión se presentarán más adelante.

En base a este ajuste se puede obtener la matriz de correlación de los residuales Pearson estandarizados para un paciente, estos residuales pueden obtenerse en base a la expresión (3.19). Dicha matriz, que se muestra a continuación, indica una muy apreciable correlación entre las observaciones de un paciente.

	1	2	3	4	5
1	1.00	0.79	0.83	0.68	0.15
2	0.00	1.00	0.87	0.74	−0.00
3	0.00	0.00	1.00	0.81	−0.02
4	0.00	0.00	0.00	1.00	−0.03
5	0.00	0.00	0.00	0.00	1.00

Una de las características de los modelos marginales mencionadas en el capítulo 2 es que se puede elegir entre distintas estructuras de correlación. Más adelante se presentarán los resultados del ajuste considerando las diferentes estructuras de correlación que permite el paquete `geepack`, pero antes, se explicará la sintaxis de la función `geeglm` que se usará para el ajuste de los datos de ataques epilépticos.

La función `geeglm` permite ajustar modelos marginales a datos agrupados, en esta caso longitudinales, usando un enfoque de GEE's el cual se enfoca en modelos para la media de las observaciones correlacionadas dentro de grupos sin la necesidad de especificar completamente la distribución conjunta de las observaciones. Esta función sigue en gran parte la sintaxis de la función `glm`, y muchos de los métodos disponibles para los objetos `glm`, también se disponen para los objetos `geeglm` (los métodos que se pueden aplicar a objetos `glm` pueden revisarse en la ayuda de R [63]).

La estructura general de la función `geeglm` es la siguiente:

```
geeglm(formula, family = gaussian, data = parent.frame(),
       id, corstr = "independence", std.err="san.se")
```

donde se define

`formula`: De la forma `y~x`, donde `y` representa la respuesta y `x`, la o las variables explicativas consideradas para el ajuste. Para este ejemplo, la fórmula es la misma que la presentada anteriormente en el ajuste usando `glm`.

`family`: El nombre de alguna distribución de la familia exponencial. Las distribuciones disponibles son: `gaussian, binomial, poisson` y `gamma`. Si no se especifica, se toma por default `gaussian`.

`data`: El nombre de la base de datos a los que se ajustará el modelo. En este caso la base de datos se llama `seiz.l`.

`id`: Un vector que identifica las unidades individuales. La longitud de 'id' debe ser la misma que el número de observaciones. Se asume que los datos están ordenados así que las observaciones en un individuo son renglones contiguos para todas las entidades en la fórmula. Recuérdese que para este ejemplo fue necesario reestructurar la base de datos original a fin de transformarla en una estructura de datos longitudinales (como en la tabla 4.3) para poder hacer los ajustes en R.

`corstr`: Una palabra que identifique la estructura de la matriz de correlación. Las siguientes están permitidas: 'independence' como en la expresión (2.5), 'exchangeable' como en la expresión (2.6), 'ar1' cuya expresión puede verse en [46] y 'unstructured' como en la expresión (2.7).

`std.err`: El tipo de error estándar a ser calculado. Por default 'san.se' es el estimador robusto usual (sandwich). Otras opciones son 'fij': es la estimación de la varianza jackknife completamente iterado, además de 'jack': para la aproximación jackknife de la varianza o 'j1s': si se quiere el estimador jackknife de un paso, los cuales son computacionalmente menos demandantes que 'fij'.

Para un número pequeño de grupos ($m \leq 30$) el estimador sandwich de la varianza exhibe un sesgo y por eso se recomienda usar el estimador jackknife de la varianza ([29]). Este está definido por

$$\frac{m-p}{m}\sum_{i=1}^{m}(\hat{\beta}_{-i}-\hat{\beta})^T(\hat{\beta}_{-i}-\hat{\beta}),$$

donde p es el número de parámetros en la estructura de la media y $\hat{\beta}_{-i}$ son los estimadores de β sin tomar en cuenta la información del i-ésimo individuo.

En la exploración de los datos se observó una considerable correlación en las observaciones para un paciente. Ahora se hará el ajuste del modelo marginal usando la función liga identidad para respuestas Poisson y considerando las cuatro estructuras de correlación antes mencionadas. Para el caso en el que se considera una estructura de equicorrelación en el ajuste de los datos de ataques epilépticos, la función `geeglm` se usa de la siguiente forma:

```
s.eqi <- geeglm(y ~ offset(log(t)) + x + trt + x:trt,
    data=seiz.l, family=poisson, id=id, corstr="exchangeable")
```

Las estimaciones de los parámetros, sus correspondientes errores estándar y estadísticos de prueba se presentan en la tabla 4.4 para los distintos ajustes, incluyendo el ajuste con `glm`. Véase que en el caso del modelo ajustado con `glm`, el estadístico de prueba corresponde a un valor z con distribución normal estándar, mientras que el ajuste con la función `geeglm` usa el estadístico de Wald el cual se distribuye χ^2_1.

La idea básica respecto del uso del estadístico de Wald es el hecho de probar la hipótesis nula de que el valor del parámetro en cuestión es cero, contra la alternativa de que el valor es distinto de cero. Bajo el enfoque de las GEE's cada parámetro estimado $\hat{\beta}_i$ se distribuye asintóticamente $N(\beta_i, \sigma_i^2)$ donde σ_i^2 es la estimación de la varianza de β_i usando el valor estimado $\hat{\beta}_i$ correspondiente, con $i = 1, \ldots, p$.

Con esta estructura, el estadístico de prueba se define como

$$W_i = \left(\frac{\hat{\beta}_i - \beta_i}{\sigma_i}\right)^2 \equiv \left(\frac{\hat{\beta}_i}{\sigma_i}\right)^2$$

el cual se distribuye χ^2_1. La descripción del caso multivariado de este estadístico se puede encontrar en Dobson [15], Hardin y Hilbe [31], Fahrmeir y Tutz [18] y Garthwaite [24].

(a)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.35	0.03	39.57	0.00
x	0.11	0.05	2.39	0.02
trt	0.03	0.05	0.59	0.56
x:trt	−0.10	0.07	−1.61	0.11

(b)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.35	0.16	73.34	0.00
x	0.11	0.12	0.93	0.33
trt	0.03	0.22	0.02	0.90
x:trt	−0.10	0.22	0.23	0.63

(c)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.35	0.16	73.34	0.00
x	0.11	0.12	0.93	0.33
trt	0.03	0.22	0.02	0.90
x:trt	−0.10	0.22	0.23	0.63

(d)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.31	0.16	65.37	0.00
x	0.16	0.11	1.88	0.17
trt	0.02	0.21	0.01	0.94
x:trt	−0.13	0.27	0.23	0.63

(e)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	0.92	0.43	4.65	0.03
x	0.26	0.19	2.03	0.15
trt	0.11	0.50	0.05	0.82
x:trt	−0.15	0.36	0.17	0.68

Cuadro 4.4: Parámetros estimados usando MGL's (a) y GEE's con distintas estructuras de correlación: (b) independencia, (c) equicorrelación, (d) autorregresivo de orden 1 y (e) no estructurado.

Obsérvese que la estimación de los parámetros es muy semejante para los ajustes hechos con MLG's y para GEE's con estructuras de correlación de independencia, equicorrelación y autorregresivo de primer orden, lo cual no sucede con la estructura de correlación no especificada (e). En este sentido, el ajuste del modelo marginal con cualquiera de las estructuras de correlación lleva a la conclusión preliminar de que la única variable explicativa signi-

ficativa en el estudio es el intercepto u ordenada. Sin embargo, recuérdese la importancia que tiene un valor negativo en la interacción `x:trt`. Dicha interacción será analizada más adelante a fin de ver su significancia en el modelo.

En la tabla 4.5 se presenta una comparación del estimador de varianza 'sandwich' y varios estimadores de varianza 'jackknife' para el caso en que se considera una estructura de equicorrelación. Obsérvese que en todos los casos los valores son prácticamente los mismos.

	san	fij	j1s	jack
(Intercept)	0.16	0.16	0.15	0.15
x	0.12	0.12	0.11	0.11
trt	0.22	0.22	0.22	0.22
x:trt	0.22	0.24	0.21	0.21

Cuadro 4.5: Errores estándar para el ajuste de datos de epilepsia usando diferentes estimadores de varianza en base a una estructura de equicorrelación

Un análisis adicional que se puede hacer con el paquete `geepack` es un análisis de varianza en base al estadístico de Wald para comparar dos modelos: el modelo completo, que incluye todas las variables de interés y un 'sub-modelo' o modelo reducido, el cual resulta de haber eliminado alguna o algunas variables del modelo completo. Este análisis permite observar si una variable es significativa o no en el ajuste de los datos.

Para este ejemplo, primero se obtiene un modelo en el que se eliminó la interacción (`x:trt`) en base a los ajustes presentados en la tabla 4.4 y al análisis que se pretende hacer sobre esta interacción, esto se hace 'actualizando' el modelo completo (se considera una estructura de equicorrelación):

```
s.eqi.0<-update(s.eqi,.~.-x:trt)
```

Ahora se procede a efectuar el análisis de varianza:

```
anova(s.eqi,s.eqi.0)
```

Cabe señalar que no importa el orden en que se coloquen los modelos, el modelo 1 siempre será el que tenga más variables (se puede hacer el ANOVA a dos 'sub-modelos' siempre y cuando uno sea sub-modelo del otro) y el modelo 2 será el modelo 'reducido'.

El resultado del análisis se basa entonces en la definición de los modelos:
`Model 1 y ~ offset(log(t)) + x + trt + x:trt`
`Model 2 y ~ x + trt + offset(log(t))`
y se obtienen los siguientes resultados:

	Df	X2	P(>|Chi|)
1	1	0.22978	0.63169

El valor P del ANOVA resulta ser alto respecto respecto del nivel de significancia de, por ejemplo, $\alpha = 0{,}05$, lo que indica que la variable `x:trt` puede ser eliminada del modelo. Dicho de otro modo, a pesar que el parámetro estimado para la interacción antes mencionada tiene un valor negativo ($\hat{\beta}_3 = -0{,}10 \pm 0{,}22$), indicando una disminución en el conteo de ataques epilépticos con el tratamiento progabide, dicha disminución no es significativa respecto del tratamiento placebo. Esta conclusión se sospechaba desde la observación de la gráfica 4.4 y de los resultados de la tabla 4.4.

Los valores de los parámetros de regresión de la tabla 4.4, así como el análisis de varianza anterior, sugieren que no hay una diferencia significativa entre los grupos de tratamiento progabide y de control (placebo) en el cambio del conteo de ataques antes y después del tratamiento, por lo cual, médicamente el progabide no ofrece una mejor alternativa para el tratamiento de ataques epilépticos. Sin embargo, en el análisis hecho para llegar a esta conclusión, se han tomado en cuenta las observaciones atípicas del paciente 49. Entonces, el siguiente paso es realizar el mismo análisis pero excluyendo las observaciones de este paciente a fin de poder comparar resultados y conclusiones.

Lo primero en hacerse es ajustar con splines los datos de los dos grupos sin la participación de los datos del paciente 49. En la gráfica 4.5 se puede observar que la diferencia entre los ajustes spline de cada grupo de tratamiento es más amplia en comparación a la diferencia entre grupos mostrada en la gráfica 4.4. Este análisis gráfico no es un medio contundente para decir que hay una diferencia significativa entre grupos de tratamiento, pero indica que es necesario hacer un análisis más detallado.

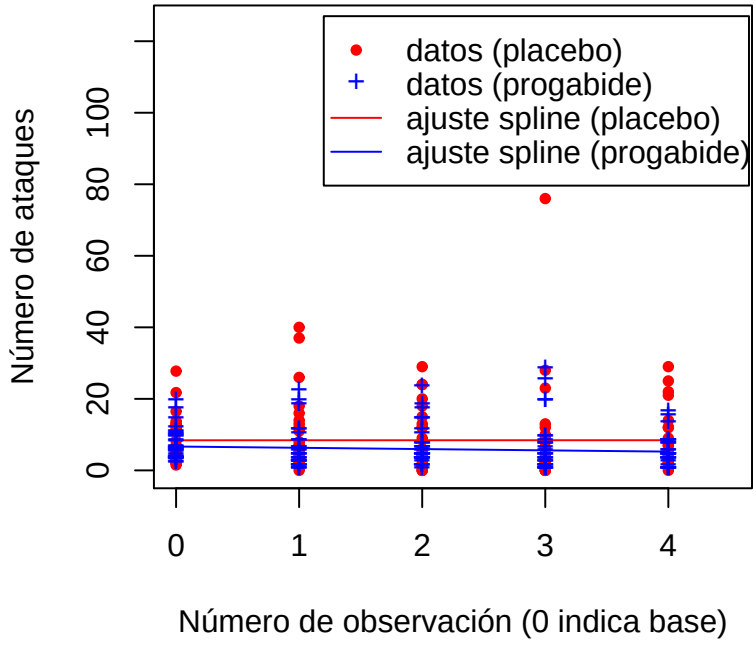


Figura 4.5: Ajuste *spline* para cada grupo de tratamiento (paciente No. 47 excluido)

Entonces, el siguiente paso es hacer el ajuste marginal al nuevo grupo de datos. Los resultados con cada una de las estructuras de correlación para modelos marginales disponibles en el paquete `geepack` se muestran en la tabla 4.6. Obsérvese que ahora la interacción `x:trt` es significativa al nivel $\alpha = 0{,}10$ para todas las estructuras de correlación. Considerando la estructura de equicorrelación, se observa que ahora la reducción del número de ataques con el tratamiento progabide respecto del tratamiento placebo es más significativa que en el caso en que se considera la participación del paciente 49.

La conclusión general es que si no se considera la información del paciente 49, los resultados indican que el progabide es una alternativa viable para el tratamiento de ataques epilépticos.

(a)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.35	0.16	73.34	0.00
x	0.11	0.12	0.93	0.33
trt	−0.11	0.20	0.30	0.58
x:trt	−0.30	0.17	3.01	0.08

(b)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.35	0.16	73.34	0.00
x	0.11	0.12	0.93	0.33
trt	−0.11	0.20	0.30	0.58
x:trt	−0.30	0.17	3.01	0.08

(c)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.31	0.16	66.10	0.00
x	0.15	0.11	1.86	0.17
trt	−0.08	0.20	0.16	0.69
x:trt	−0.40	0.18	5.01	0.03

(d)

	Estimate	Std.err	Wald	p(>W)
(Intercept)	1.33	0.16	69.88	0.00
x	0.11	0.10	1.42	0.23
trt	−0.10	0.20	0.27	0.60
x:trt	−0.31	0.16	4.08	0.04

Cuadro 4.6: Parámetros estimados para los datos de epilepsia (sin el paciente 49) usando GEE's: (a) independencia, (b) equicorrelación, (c) autorregresivo de orden 1 y (d) no estructurado.

4.2. Modelos de efectos aleatorios

4.2.1. Respuestas gaussianas: programa de seguro médico

Para exponer los modelos de efectos aleatorios se seguirá el esquema ilustrativo de la sección anterior, describiendo primero los datos indicando cuál es la variable respuesta y cuál o cuáles son las variables explicativas para después establecer un análisis exploratorio que permita definir el tipo de ajuste a realizar y/o las variables que intervendrán en dicho ajuste, y finalizar con el diagnóstico del o los modelos ajustados a los datos. La base ilustrativa

de los modelos de efectos aleatorios proviene del ejemplo 1.4 de datos sobre cargos al seguro médico *Medicare* que se analiza en Frees, Young y Lou [23]. Como se mencionó en la descripción de los datos, este ejemplo corresponde a un estudio realizado para el programa de seguro médico Medicare en 54 Estados de la Unión Americana.

Los datos correspondientes a los dos primeros Estados en el estudio se presentan a continuación en la tabla 4.7.

Estado	Tiempo	NA	RC	EP
1	1	230015	9435.212	8.402378
1	2	234739	10514.928	8.251458
1	3	245027	11927.713	8.229109
1	4	243947	12911.598	7.987091
1	5	258384	13097.968	7.455318
1	6	261738	13344.014	7.057500
2	1	6636	9379.487	7.824442
2	2	6940	9737.740	7.644236
2	3	7646	10011.797	7.218284
2	4	7610	11077.412	7.007753
2	5	8229	12035.143	7.008385
2	6	8940	13140.126	7.013311

Cuadro 4.7: Datos del programa de seguro médico *Medicare*

La variable respuesta de interés es RC, la cantidad de indemnizaciones cubiertas por el programa Medicare (en dólares por paciente). Las variables explicativas son Tiempo (en años), NA (número de pacientes dados de alta) y EP (estancia promedio en hospitalización). Adicionalmente, 'Estado' es la variable indicadora referente al Estado de la Unión Americana al que corresponde la observación obtenida [1].

En la figura 4.6 se muestra una gráfica de series de tiempo múltiple de RC donde las unidades de tiempo están consideradas en años a partir de 1995. En esta gráfica se ha remarcado en rojo la serie de observaciones correspondiente al estado No. 31 para indicar que el comportamiento de la serie de este Estado es algo inusual en comparación con las series de los demás Estados. Esta conjetura se basa en el hecho de que se observa una pendiente positiva muy

[1] No existe información sobre qué número corresponde a cada Estado.

pronunciada comparada con las pendientes de las otras series. Se podría decir que cada año el índice de indemnizaciones cubiertas por Medicare aumenta considerablemente y no hay indicios de una disminución, de ahí su extrañeza.

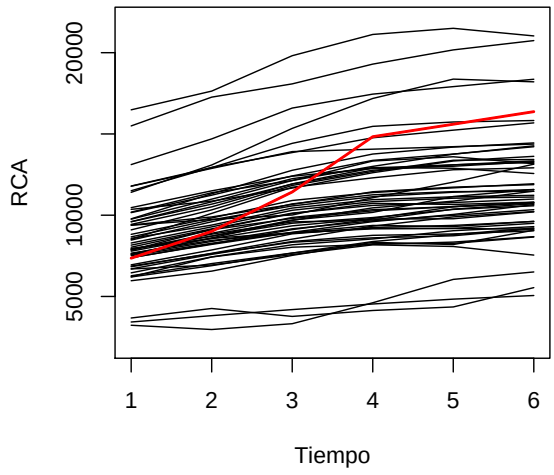


Figura 4.6: Gráfica de serie de tiempo múltiple para RC

En el ajuste del modelo de efectos aleatorios se considerará la interacción `tiempo:Estado 31` para tomar en cuenta el comportamiento inusual de la serie del Estado 31. Se considera que `Estado 31=1` cuando efectivamente se están considerando las observaciones de ese Estado, y `Estado 31=0` en otro caso.

Al asumir que las respuestas siguen una distribución Gaussiana, se usará el enfoque de Laird y Ware [44] en el que el modelo de datos longitudinales con efectos aleatorios se define como:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$$

para el i-ésimo individuo (Estado), donde las β son parámetros correspondientes a los efectos fijos o de la población, y las $\mathbf{b}_i$ son los correspondientes a los efectos aleatorios o de sujeto específico.

El modelo queda especificado por

$$\begin{aligned} \mathrm{RC_i} &= \beta_0 + \beta_1 \mathrm{Tiempo_i} + \beta_2 \mathrm{NA_i} + \beta_3 \mathrm{EP_i} + \beta_4 \mathrm{Tiempo_i} * (\mathrm{Estado} = 31) \\ &\quad + \mathrm{b_{i1}} + \mathrm{b_{i2}} \mathrm{Tiempo_i} \end{aligned}$$

donde se consideran como efectos fijos los elementos en el primer renglón de la fórmula, incluyendo la interacción $Tiempo_i * (Estado = 31)$ con la cual se considera el comportamiento inusual de RC en el Estado 31 (línea roja en la gráfica 4.6), y se considera como variables con efectos aleatorios a la ordenada y a 'Tiempo', los cuales están considerados en el segundo renglón de la fórmula.

Antes de exponer la forma en que se realizaron las regresiones y los ajustes de los modelos para estos datos, se expone la sintaxis de la función `lme` del paquete `nlme` para R, la cual será usada para analizar estos procesos.

La función `lme` permite ajustar modelos lineales de efectos aleatorios (modelos mixtos) en base a la formulación de Laird y Ware [44] para datos longitudinales. Puede aplicarse a datos que presentan correlación dentro de grupos y/o cuyos errores tienen varianzas desiguales. El ajuste con esta función está basado en máxima verosimilitud. Los métodos computacionales están descritos en Pinheiro y Bates [61] y siguen la estructura general de Lindstrom y Bates [48].

La estructura básica de la función `lme` es como sigue:

```
lme(fixed, data, random, correlation, method)
```

donde

`fixed` Una fórmula lineal que describa la parte de los efectos fijos en el modelo, de la forma `y~x1+x2+...+xp`, un objeto `lmList`, o un objeto `groupedData`. La metodología de las funciones `lme.lmList` así como de `lme.groupedData` están documentadas separadamente y se pueden encontrar en la ayuda de R.

`data` Una lista de datos que contiene las variables nombradas en `fixed`, `random`, `correlation`. Es el conjunto de datos longitudinales a los cuales se pretende ajustar un modelo.

`random` Una fórmula de un sólo lado de la forma `~x1+...+xn | g1/.../gm`, con `x1+...+xn` especificando el modelo para los efectos aleatorios y `g1/.../gm` la estructura (identificación) de grupos (`m` puede ser igual a 1, en tal caso no se necesita `/`).

`correlation` Un objeto opcional `corStruct` que describe la estructura de correlación dentro de grupos. Por default se asume que no existen correlaciones dentro de grupos.

`method` Una palabra indicadora. Si `method='REML'` el modelo es ajustado maximizando la log-verosimilitud restringida. Si `method='ML'` la log-verosimilitud es maximizada. Por default se considera `'REML'`.

Entonces, para los datos de Medicare, el modelo de efectos aleatorios queda especificado en R de la siguiente manera:

```
modelo1 <- lme(RC ~ Tiempo + Tiempo:(Estado==31) + NA + EP,
          Medicare, random = ~Tiempo|Estado, method = "ML")
```

En este caso, `Medicare` representa el conjunto de datos para el análisis. La tabla 4.8 muestra los valores de los parámetros de efectos fijos. En ella se observa que, a un nivel de significancia de 0.05, la variable NA no es significativa para el modelo.

	Value	Std.Error	DF	t-value	p-value
(Intercept)	4406.221	574.414	265	7.67	0.00
Tiempo	738.119	34.572	265	21.35	0.00
NA	0.003	0.001	265	1.84	0.07
EP	340.296	45.647	265	7.45	0.00
Tiempo:Estado==31	1530.421	184.038	265	8.32	0.00
log-verosimilitud	-2567.522				

Cuadro 4.8: Estimadores de los efectos fijos en el modelo para el programa *Medicare* y valor de log-verosimilitud del ajuste.

Se considera ahora el modelo en el cual la variable NA ha sido eliminada en base a los estadísticos de la tabla 4.8. La formulación en R queda especificada entonces de la siguiente manera:

```
modelo2 <- lme(RC ~ Tiempo + Tiempo:(Estado==31) + EP,
          Medicare, random = ~Tiempo|Estado, method = "ML")
```

La tabla 4.9 muestra los valores de parámetros para este ajuste en la cual se observa que todos estos parámetros son significantes en el modelo.

	Value	Std.Error	DF	t-value	p-value
(Intercept)	4827.04	536.76	266	8.99	0.00
Tiempo	753.11	33.93	266	22.20	0.00
EP	348.32	44.87	266	7.76	0.00
Tiempo:Estado == 31	1540.89	180.24	266	8.55	0.00
log-verosimilitud	-2569.085				

Cuadro 4.9: Estimadores de efectos fijos para el modelo que no incluye la variable NA.

Se puede comprobar la no-significancia de la variable NA en el modelo mediante el cálculo del estadístico de razón de verosimilitud para dos modelos usando `anova(modelo1,modelo2)`. Los resultados de esta prueba para el caso en que se consideran los dos modelos antes expuestos para los datos del programa Medicare aparecen en la tabla 4.10.

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
modelo1	8	5154.17	5184.39	−2569.09			
modelo2	9	5153.04	5187.04	−2567.52	1 vs 2	3.13	0.08

Cuadro 4.10: Valores de la prueba de razón de verosimilitud para los dos ajustes a los datos del programa Medicare.

Con un nivel de significancia de 0.05, estos datos corroboran la exclusión de la variable NA en el modelo, por lo tanto el modelo considerado en lo subsecuente es `modelo2`. Otro análisis que se puede hacer es la verificación de la significancia de los efectos aleatorios, eliminando uno de ellos del modelo en cada ocasión; esto sirve para ver si alguna de las variables de efectos aleatorios es significativa en el modelo:

```
modelo3<-update(modelo2,random=~1|Estado)
modelo4<-update(modelo2,random=~Tiempo-1|Estado)
```

El siguiente paso es realizar nuevamente una prueba de razón de verosimilitud entre `modelo2` y `modelo3`, y entre `modelo2` y `modelo4`. Los resultados se presentan en la tabla 4.11 e indican que los dos parámetros de efectos aleatorios son muy significativos para el modelo.

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
modelo2	8	5154.17	5184.39	−2569.09			
modelo3	6	5273.44	5296.11	−2630.72	1 vs 2	123.27	0.00
modelo2	8	5154.17	5184.39	−2569.09			
modelo4	6	5685.51	5708.18	−2836.76	1 vs 2	535.34	0.00

Cuadro 4.11: Valores de las pruebas de razón de verosimilitud para verificar la significancia de los parámetros de efectos aleatorios.

De hecho, otra forma de ver la significancia de los efectos aleatorios es observar la magnitud de la desviación estándar de cada uno de los parámetros de efectos aleatorios. Entre más cerca este este valor de cero, se considera a dicho efecto aleatorio como no significativo. Dicho de otro modo, si la desviación estándar de un efecto aleatorio es cercana a cero, quiere decir que no hay variación en el efecto de la variable aleatoria correspondiente de un individuo a otro. En la siguiente tabla se muestra intuitivamente que los efectos aleatorios son significativos, lo cual se comprobó con el análisis anterior.

	StdDev	Corr
(Intercept)	2341.2892	
		0.744
Tiempo	201.4691	

Cuadro 4.12: Desviación estándar y correlación de los efectos aleatorios.

Entonces, el modelo final queda de la siguiente forma:

$$\begin{aligned} \mathrm{RC_i} &= 4827 + 753\mathrm{Tiempo_i} + 348\mathrm{EP_i} + 1541\mathrm{Tiempo_i} * (\mathrm{Estado} = 31) + \\ &\quad \mathrm{b_{i1}} + \mathrm{b_{i2}} * \mathrm{Tiempo_i} \end{aligned}$$

donde las $\mathrm{b_{i1}}$ y $\mathrm{b_{i2}}$ están definidas para el i-ésimo estado y no serán mostradas aquí. Estas pueden obtenerse usando `modelo2$coefficients`.

Con esta regresión se pueden hacer pronósticos a futuro de la cantidad de indemnizaciones que serán cubiertas por el programa de seguro en cada Estado como se muestra en la figura 4.7.

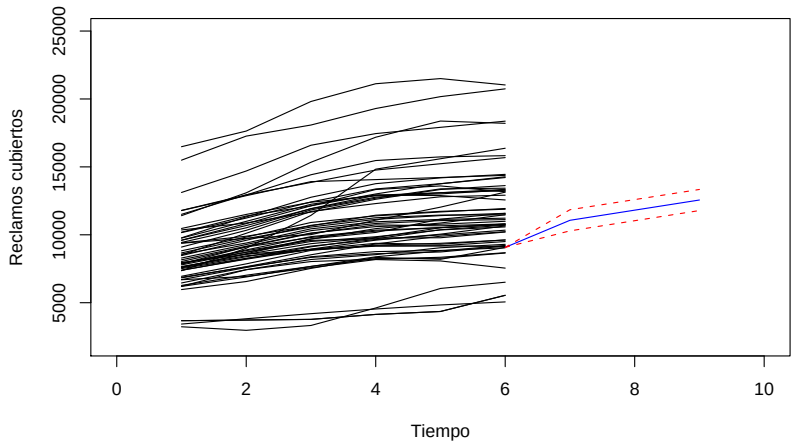


Figura 4.7: Pronósticos para el estado No. 52

En el artículo de Frees, Young y Luo [23] se exponen las consideraciones para hacer estos pronósticos y en Fox [22] y la ayuda de R [63], se encuentra la descripción de los comandos para hacer predicciones en base a un ajuste por `lme`.

4.2.2. Respuestas no gaussianas: ataques epilépticos

El análisis de datos de la subsección anterior se hizo bajo el supuesto de que las respuestas seguían una distribución Gaussiana, por lo que en el siguiente ejemplo, se analizará el caso en que esto no sucede, en particular se ilustrará el caso de respuestas de conteo donde se asume una distribución Poisson usando el conjunto de datos de ataques epilépticos y la función `lmer` del paquete `lme4` para R.

La sintaxis de esta función es como sigue:

```
lmer(formula, data, family, method)
```

donde

`formula` Una fórmula lineal de la forma `y~x1+x2+...+xp+(X1+...+Xq|id)`, donde x_i, $i = 1, \ldots, p$ son los efectos fijos y X_j, $j = 1, \ldots, q$ son los efectos aleatorios para cada individuo identificado con `id`.

`data` Una lista de datos que contiene las variables nombradas en `formula`. Es el conjunto de datos longitudinales a los cuales se pretende ajustar un modelo.

`family` El nombre de una de las distribuciones de la familia exponencial que se asume tienen las respuestas. Respecto a la distribución de los efectos aleatorios, hasta la fecha solamente se ha programado la función lmer bajo la suposición de que la distribución de los efectos aleatorios es Gaussiana.

`method` Una palabra indicadora. Si `method='REML'` el modelo es ajustado maximizando la log-verosimilitud restringida. Si `method='ML'` la log-verosimilitud es maximizada. Para modelos lineales mixtos por default se considera `'REML'`. Para modelos lineales mixtos generalizados (como es el caso en este ejemplo), por default se considera `'ML'`.

Bajo este enfoque, el modelo de efectos aleatorios para los datos de epilepsia queda especificado en R de la siguiente manera:

```
sei <- lmer(y ~ offset(log(t))+x+trt+x:trt+(x|id),seiz.l,
            poisson, method = "ML")
```

En este caso, las variables con efectos aleatorios son la ordenada y `x`, que es la variable que indica si la observación correspondiente es del periodo base (`x=0`) o del periodo de tratamiento (`x=1`), la cual se incluye para tomar en cuenta la posible heterogeneidad entre sujetos en la razón de conteos de ataques epilépticos antes y después del tratamiento. Se incluye la información del paciente 49 que se consideró anteriormente atípica.

Los resultados del ajuste se muestran en la siguiente tabla:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	1.0714924	0.1400200	7.652	1.97e-14
x	-0.0006207	0.1079951	-0.006	0.9954
trt	0.0499083	0.1929297	0.259	0.7959
x:trt	-0.3061411	0.1502999	-2.037	0.0417

Cuadro 4.13: Estimadores de los efectos fijos para los datos de epilepsia.

Estos resultados muestran que al nivel $\alpha = 0{,}05$ la ordenada y la interacción `x:trt` son significativas en el modelo. La significancia de los efectos

aleatorios se puede verificar ajustando un submodelo que no incluya alguno de estos efectos aleatorios y haciendo un análisis de varianza entre el modelo completo y este submodelo. En el siguiente submodelo, se excluye el efecto aleatorio de la variable `x`:

```
sei2 <- lmer(y ~ offset(log(t))+x+trt+x:trt+(1|id),seiz.l,
           poisson, method = "ML")
```

y el análisis de varianza correspondiente, indica que los efectos aleatorios de `x` son significativos para el modelo:

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
sei2	5	970.66	989.10	−480.33			
sei	7	802.79	828.60	−394.40	171.87	2	0.0000

Cuadro 4.14: Análisis de varianza para la significancia de los efectos aleatorios de `x`.

Ahora bien, para finalizar esta subsección se analizará el caso donde no se considera al paciente 49 en el modelo de efectos aleatorios para respuestas de conteo. Los resultados del ajuste bajo estas condiciones se muestra a continuación:

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	1.069807	0.133559	8.010	1.15e-15
x	0.006423	0.104936	0.061	0.9512
trt	-0.008988	0.185584	-0.048	0.9614
x:trt	-0.344123	0.147820	-2.328	0.0199

Cuadro 4.15: Estimadores de los efectos fijos para los datos de epilepsia (sin el paciente 49).

La prueba de significancia de los efectos aleatorios para este modelo arroja las mismas conclusiones que en el caso donde se incluye al paciente 49.

Bajo el modelo ajustado al conjunto de datos completo (incluye al paciente 49), se puede concluir que el tratamiento progabide tiene un efecto modesto para la disminución en el conteo de ataques epilépticos ($\hat{\beta}_3 = -0{,}3061$). Si no se considera al paciente 49, la evidencia de que el progabide

es efectivo en la reducción de ataques epilépticos es un poco más fuerte que en el caso de datos completos, con $\hat{\beta}_3 = -0{,}3441$.

4.3. Análisis de niveles de gravedad de una enfermedad

El modelo particular de transición que se estudiará aquí es el modelo de Markov multi-estados en tiempo continuo en el cual se supone que el valor esperado de la respuesta presente, sólo depende de la respuesta anterior. Para ello se hará uso de un conjunto de datos de monitoreo de transplante de corazón [65]. Los datos son especificados como una serie de observaciones agrupadas por paciente (cada paciente tiene un número de identificación).

El historial de los dos primeros pacientes (de un total de 622) se muestra enseguida.

	PTNUM	age	years	dage	sex	pdiag	cumrej	state
1	100002	52.50	0.00	21	0	IHD	0	1
2	100002	53.50	1.00	21	0	IHD	2	1
3	100002	54.50	2.00	21	0	IHD	2	2
4	100002	55.59	3.09	21	0	IHD	2	2
5	100002	56.50	4.00	21	0	IHD	3	2
6	100002	57.49	5.00	21	0	IHD	3	3
7	100002	58.35	5.85	21	0	IHD	3	4
8	100003	29.51	0.00	17	0	IHD	0	1
9	100003	30.70	1.19	17	0	IHD	1	1
10	100003	31.52	2.01	17	0	IHD	1	3
11	100003	32.50	2.99	17	0	IHD	2	4

`PTNUM` es el identificador del sujeto. Aproximadamente cada año después del transplante, cada paciente tiene un examen angiográfico en el cual se puede diagnosticar el estado de CAV. El resultado de esta prueba están en la variable `state`, con los posibles valores 1, 2, 3 y 4 los cuales representan un estado libre de CAV, CAV moderado, CAV severo y muerte respectivamente. `years` da la fecha de la prueba en años desde el transplante de corazón. Otras variables incluyen `age` (edad del paciente a la fecha de visita), `dage` (la edad del donante), `sex` (el sexo del paciente 0=hombre, 1=mujer), `pdiag`

(diagnóstico primario o razón del transplante - IHD representa Isquemia, IDC representa Cardiopatía idiopática) y `cumreg` (número acumulado de episodios de rechazo).

El paquete de R que se usará para ajustar estos datos se llama `msm` y es un paquete de funciones para modelado multi-estados. La función `msm`, que será la función principal a utilizar, implementa la estimación por máxima verosimilitud para modelos de Markov multi-estados en tiempo continuo en base al método descrito por Kalbfleisch y Lawless [38]. El desarrollo del paquete `msm` fue motivado por aplicaciones al modelado del curso de una enfermedad, pues muchas enfermedades crónicas tienen una interpretación natural en términos de estados o niveles progresivos. Como ejemplo de estas situaciones médicas que han sido modeladas usando modelos multi-estados se puede mencionar Jackson, et al. [36], Gentleman, et al. [26] entre otros.

Una forma usual de resumir datos multi-estados es con una tabla de frecuencias de pares de estados consecutivos. Esta cuenta sobre todos los individuos, el número de veces que un individuo tuvo una observación del estado r seguida por una observación del estado s. La función `statetable.msm` puede ser usada para producir dicha tabla como sigue:

```
statetable.msm(state, PTNUM, data = heart)
```

lo que resulta en

```
    to
from    1    2    3    4
   1 1367  204   44  148
   2   46  134   54   48
   3    4   13  107   55
```

En el artículo de Sharples y otros [65] se considera que inicialmente el paciente puede avanzar o recuperarse de estados consecutivos mientras esté vivo y puede pasar a la muerte desde cualquier estado.

Un modelo de este tipo es especificado por una matriz de intensidad de transición. Para este caso de cuatro estados:

$$Q = \begin{pmatrix} -(q_{12}+q_{14}) & q_{12} & 0 & q_{14} \\ q_{21} & -(q_{21}+q_{23}+q_{24}) & q_{23} & q_{24} \\ 0 & q_{32} & -(q_{32}+q_{34}) & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Es necesario definir valores iniciales para esta matriz. Por ejemplo,

```
> ini.q <- rbind(c(0, 0.25, 0, 0.25), c(0.166, 0, 0.166, 0.166),
                 c(0, 0.25, 0, 0.25), c(0, 0, 0, 0))
```

Ajustar el modelo es un proceso de encontrar valores para las intensidades: $q_{12}, q_{14}, q_{21}, q_{23}, q_{24}, q_{32}, q_{34}$, los cuales maximicen la verosimilitud. Esto se hace usando la función `msm` con los argumentos apropiados:

```
ajuste <- msm(state ~ years, subject = PTNUM,
            data = heart, qmatrix = ini.q, death = 4)
```

En este ejemplo el día de la muerte se asume que es registrado exactamente, como es usual en estudios de enfermedades crónicas. Se especifica también `death=4` para indicar a la función `msm` que el estado 4 es un estado de 'muerte'.

Mientras la función `msm` se está ejecutando, ésta busca el valor máximo de la verosimilitud de los parámetros desconocidos. Internamente, hace uso de la función `optim` de R para minimizar el valor negativo de log-verosimilitud(véase la ayuda de R para `optim`).

Para mostrar los estimadores de máxima verosimilitud y los intervalos de confianza al 95 % (entre paréntesis) se escribe el nombre del objeto (en este caso `ajuste`) y se obtiene la matriz de intensidad que a continuación se presenta:

	State 1	State 2
State 1	-0.17 (-0.19,-0.15)	0.13 (0.11,0.15)
State 2	0.22 (0.17,0.30)	-0.61 (-0.71,-0.52)
State 3	0	0.13 (0.08,0.22)
State 4	0	0
	State 3	State 4
State 1	0	0.04 (0.03,0.05)
State 2	0.34 (0.27,0.43)	0.04 (0.01,0.14)
State 3	-0.44 (-0.55,-0.35)	0.30 (0.24,0.39)
State 4	0	0

Cuadro 4.16: Valores estimados de las entradas de la matriz intensidad con sus correspondientes intervalos de confianza

El valor para -2log-verosimilitud en este ajuste es de 3968.803. Se puede obtener la estimación de la matriz de probabilidad de transición P dentro de un tiempo dado, por ejemplo, dentro de 10 años, `pmatrix.msm(ajuste,t=10)` da las estimaciones:

	State 1	State 2	State 3	State 4
State 1	0.31	0.10	0.09	0.50
State 2	0.17	0.07	0.08	0.68
State 3	0.06	0.03	0.05	0.86
State 4	0.00	0.00	0.00	1.00

Cuadro 4.17: Matriz estimada de probabilidad de transición a 10 años

Entonces, una paciente promedio que se encuentre en el estado 1 en este momento, tiene una probabilidad de 0.5 de morir dentro de diez años, una probabilidad de 0.31 de permanecer libre de padecimientos y una probabilidad de 0.1 y 0.09 de tener CAV moderado y CAV severo, respectivamente. En este caso se asume que la matriz de intensidad permanece constante dentro del intervalo considerado, en este caso 10 años.

Se puede estimar también el tiempo de permanencia promedio en un estado mediante:

```
totlos.msm(ajuste)
```

obteniendo

State 1	8.82377 años
State 2	2.23689 años
State 3	1.74679 años

Cuadro 4.18: Tiempo promedio estimado de estancia en cada estado

Entonces para el modelo `ajuste`, cada paciente tiene un pronóstico de permanecer 8.8 años libre de CAV, 2.2 años con CAV moderado y 1.8 años con CAV severo para finalmente morir.

En estudios de enfermedades crónicas, un uso importante de los modelos multi-estados es predecir la probabilidad de sobrevivencia de los pacientes para algún tiempo en el futuro. Esta puede ser obtenida directamente de

la matriz de probabilidad de transición P. Se puede obtener una gráfica de probabilidad esperada de sobrevivencia sobre el tiempo para cada estado transitorio, esto se hace mediante `plot(ajuste)` con lo que se obtiene la gráfica presentada en la figura 4.8.

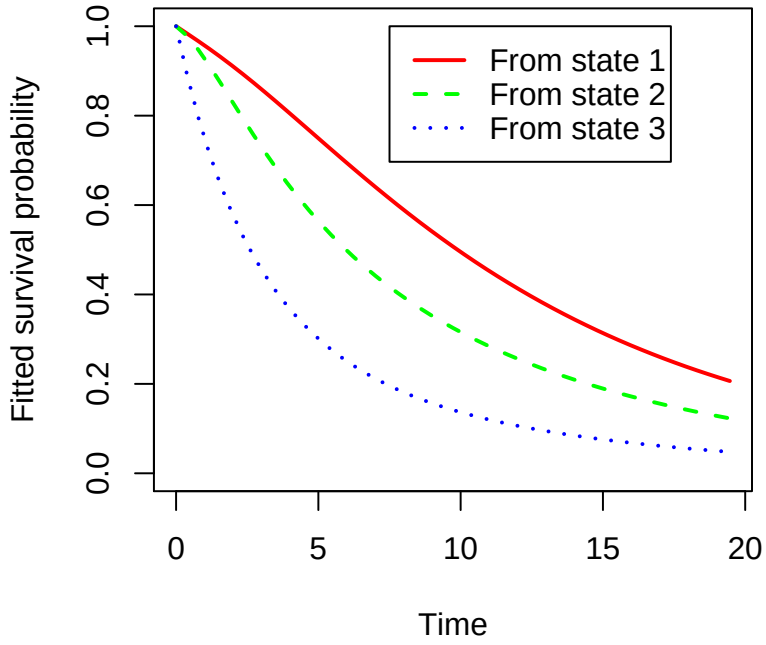


Figura 4.8: Probabilidad de sobrevivencia

Esta gráfica muestra que la probabilidad de sobrevivencia dentro de 10 años considerando que actualmente se tiene CAV severo es aproximadamente 0.1, en contraposición a 0.3 si se tiene CAV moderado y 0.5 si no se tiene CAV. Con CAV severo, la probabilidad de sobrevivencia disminuye muy rápido a alrededor de 0.3 en los primeros 5 años después del transplante, mientras que si se está libre de CAV, ésta probabilidad en 5 años es de aproximadamente 0.8.

La conclusión principal de los resultados obtenidos es que la probabilidad de muerte, dado un estado transitorio es mayor conforme el estado es más grave. Otras conclusiones pueden obtenerse analizando la matriz de probabilidad P.

El siguiente paso es analizar el efectos de variables explicativas en las razones de transición. Se tiene ahora una matriz de intensidad $Q(x)$ la cual depende de un vector de variables explicativas x. Para cada entrada de $Q(x)$, la intensidad de transición para el paciente i al tiempo de observación j es $q_{rs}(x_{ij}) = q_{rs}^{(0)} \exp(\beta_{rs}^T x_{ij})$ donde $q_{rs}^{(0)}$ es la matriz de intensidad sin considerar variables explicativas. En la función `msm`, las variables explicativas se especifican a través del argumento `covariates`. Si x_{ij} es dependiente del tiempo, se asume como constante dentro de los intervalos entre observaciones. `msm` calcula la probabilidad de una transición del tiempo $t_{i,j-1}$ a t_{ij} usando el valor de las variables explicativas al tiempo $t_{i,j-1}$.

En este trabajo se considerará solamente la variable explicativa `sex` para el ajuste del modelo. De los 622 pacientes del estudio, 535 son hombres y 87 son mujeres. Entonces, mediante la siguiente instrucción se realiza el ajuste considerando la participación de la variable `sex`.

```
ajuste.sex <- msm(state ~ years, subject = PTNUM,
            data = heart, qmatrix = ini.q, death = 4,
            covariates=~sex)
```

A partir de este ajuste, se pueden obtener por separado las matrices intensidad para cada sexo a fin de compararlas. Esto se hace con la instrucción

```
qmatrix.msm(ajuste.sex,covariates=list(sex=0))
```

para el grupo de hombres (`sex=1` para las mujeres). Las matrices de intensidad para cada grupo se muestran a continuación.

	State 1	State 2	State 3	State 4
State 1	−0.18	0.14	0.00	0.04
State 2	0.24	−0.66	0.36	0.05
State 3	0.00	0.16	−0.44	0.28
State 4	0.00	0.00	0.00	0.00

Cuadro 4.19: Matriz de intensidad para pacientes hombres (`sex=0`)

	State 1	State 2	State 3	State 4
State 1	−0.13	0.08	0.00	0.05
State 2	0.24	−0.90	0.56	0.10
State 3	0.00	0.35	−0.90	0.55
State 4	0.00	0.00	0.00	0.00

Cuadro 4.20: Matrices de intensidad para pacientes mujeres (`sex=1`)

Estas matrices de intensidad tanto para hombres como para mujeres están basadas en un modelo de riesgos proporcionales (*proportional hazards*) descrito en Marshall y Jones [49] y mencionado anteriormente.

El vector de efectos β se presenta en forma matricial para observar en que entrada de la matriz intensidad influye. Así, la matriz de efectos para la variable explicativa sexo, junto con los intervalos de confianza de los parámetros, se presenta a continuación:

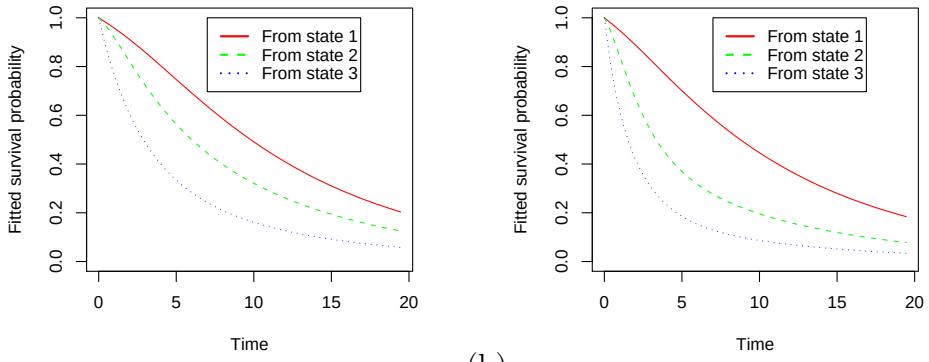
	State 1	State 2
State 1	0	-0.63 (-1.14,-0.12)
State 2	-0.02 (-1.05,1.02)	0
State 3	0	0.78 (-1.91,3.47)
State 4	0	0
	State 3	State 4
State 1	0	0.21 (-0.36,0.79)
State 2	0.45 (-0.50,1.39)	0.59 (-1.27,2.44)
State 3	0	0.67 (-0.16,1.50)
State 4	0	0

Cuadro 4.21: Matriz de efectos de la variable `sex`

Respecto de esta matriz, si el valor de una entrada tiene signo negativo, significa que la intensidad de transición denotada por dicha entrada matricial, es menor para las mujeres, y es mayor en caso contrario. Se puede ver por ejemplo que la razón de inicio de CAV, es decir, la intensidad de transición del estado 1 al estado 2, es menor para las mujeres respecto a la de los hombres, pues la entrada [1,2] en la matriz de efectos tiene signo negativo (-0.6276). Otra observación es que la intensidad de transición hacia la muerte partiendo de cualquier estado transitorio, es mayor para las mujeres respecto de los

hombres, pues las entradas [1,4], [2,4] y [3,4] de la matriz de efectos tienen valores positivos (0.21, 0.59 y 0.67 respectivamente). Esto quiere decir, en términos coloquiales, que las mujeres morirán más rápido que los hombres. La forma más sencilla de verificar estas observaciones es comparando las matrices de intensidad para cada sexo. Obsérvese que $q_{rs}(x = 1) = q_{rs}(x = 0)\exp(\beta_{rs})$ que es justamente el modelo de riesgos proporcionales.

También se puede observar gráficamente la probabilidad de sobrevivencia de cada sexo como en la gráfica 4.8. En la siguiente gráfica se observa que ya desde el quinto año la probabilidad de sobrevivencia para las mujeres disminuyó drásticamente a un valor de aproximadamente 0.4 para el estado 2 y de 0.2 para el estado 3.



(a) (b)

Figura 4.9: Probabilidad de sobrevivencia por sexo: (a) Hombres (b) Mujeres

Finalmente, la inclusión de la variable `sex` en el modelo permite observar si existen diferencias significativas en el proceso de enfermedad cardíaca entre hombres y mujeres. Médicamente este análisis indica que las mujeres necesitan más atención y cuidados a fin de poder mejorar su esperanza de vida a un nivel similar al de los hombres.

Capítulo 5

Conclusiones y perspectivas

En este trabajo se presentaron los principales enfoques en el tratamiento de datos longitudinales. Se analizó la forma en que fueron desarrollados estos modelos, con sus respectivas metodologías, y cuáles fueron las herramientas que sirvieron de base para la estructuración de un tipo de modelo particular, por ejemplo, la generalización de la función de cuasi-verosimilitud hacia el caso de las GEE's, las cuales se utilizaron en los modelos marginales de datos longitudinales. Las conclusiones extraídas a lo largo de este trabajo se presentan a continuación:

- La elección de un modelo particular para tratar un conjunto de datos longitudinales depende de los objetivos del estudio, pues la interpretación de los resultados varía de un modelo a otro. Esta elección también depende de la estructura de los datos, quiénes son las variables explicativas y qué tipo de relación tienen con la variable respuesta.

- Los modelos para datos longitudinales presentados en este trabajo, están enmarcados por una estructura basada en los modelos lineales generalizados, ya que cada uno de estos modelos puede ajustarse a respuestas tanto gaussianas como no gaussianas, discretas o continuas, mediante el uso de funciones liga en base a las distribuciones pertenecientes a la familia exponencial.

- El uso de las distintas estructuras de correlación en los modelos marginales para datos longitudinales, permite agregar la correlación que pueden presentar las respuestas como información adicional en el ajuste del modelo; a diferencia de los modelos lineales generalizados usados en

estudios de corte transversal, en los cuales se supone la independencia entre las respuestas. En el caso de un modelo marginal con estructura de correlación de independencia los valores estimados de los parámetros de regresión son los mismos que los resultantes de un tratamiento con modelos lineales generalizados.

- Los modelos de efectos aleatorios permiten la inclusión de variables adicionales que expliquen características particulares del conjunto de datos. En el caso en que estos modelos son usados en el tratamiento de datos longitudinales, tales variables adicionales, las cuales definen los efectos aleatorios, podrían ser la ordenada y los intervalos de tiempo en que se realizan las observaciones, y proporcionarían información extra que permitiría un mejor ajuste del modelo así como la exploración detallada del comportamiento de una unidad experimental particular.

- Los modelos de efectos aleatorios tienen la ventaja sobre los modelos marginales de que pueden ser utilizados incluso con datos en los cuales el número de observaciones para una unidad experimental difiere del número de observaciones para otra unidad, así como datos en los que las observaciones no están igualmente espaciadas en el tiempo, mientras que en los modelos marginales, no todas las estructuras de correlación permiten trabajar con este tipo de datos.

- Los modelos de transición permiten la observación de patrones globales de comportamiento de las respuestas en un entorno de múltiples respuestas posibles. En este sentido, una característica de los modelos de transición, que podría considerarse como ventaja, es su enorme aplicabilidad en estudios longitudinales de seguimiento, donde se observa el curso que toman en promedio las respuestas de varias unidades experimentales. Una posible desventaja es la necesidad de categorizar todas las respuestas posibles en un número finito y pequeño de clases de respuesta, con lo que en cierta forma se puede presentar pérdida de información.

Bibliografía

[1] Albert, P. S. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine 18* (1999), 1707–1732.

[2] Albert, P. S., and Waclawiw, M. A. A two-state markov chain for heterogeneous transitional data: a quasi-likelihood approach. *Statistics in Medicine 17* (1998), 1481–1493.

[3] Artes, R., and Jørgensen, B. Longitudinal data estimating equations for dispersion models. *Scandinavian Journal of Statistics 27* (2000), 321–334.

[4] Berchtold, A., and Raftery, A. E. The mixtute transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science 17* (2002), 328–356.

[5] Breslow, N. E., and Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88* (1993), 9–25.

[6] Brockwell, P., and Davis, R. *Introduction to time series and forecasting*. Springer, 2002.

[7] Cantoni, E., Flemming, J. M., and Ronchetti, E. Variable selection for marginal longitudinal generalized linear models. *Biometrics 61* (2005), 507–514.

[8] Chaganty, N. R. An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference 63* (1997), 39–54.

[9] Chaganty, N. R., and Joe, H. Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society B 66* (2004), 851–860.

[10] Chambers, J. M., and Hastie, T. J. *Statistical models in S.* wadsworth Brooks/Cole, 1992.

[11] Chang, Y. C. Residual analysis of the generalized linear models for longitudinal data. *Statistics in Medicine 19* (2000), 1277–1293.

[12] Cox, D. R. Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics 8* (1981), 93–115.

[13] Crainiceanu, C. M., and Rupert, D. Restricted likelihood ratio test for longitudinal models. *Statistica Sinica* (2004). Por publicarse.

[14] Diggle, P. J., Liang, K. Y., and Zeger, S. L. *Analysis of longitudinal data.* Clarendon Press, 1994.

[15] Dobson, A. J. *An introduction to generalized linear models.* Chapman-Hall, 2001.

[16] Dunn, P. K., and Smyth, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics 5* (1996), 236–244.

[17] Durban, M., Harezlak, J., Wand, M. P., and Carroll, R. J. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine 24* (2005), 1153–1167.

[18] Fahrmeir, L., and Tutz, G. *Multivariate statistical modelling based on generalized linear models.* Springer, 2001.

[19] Faraway, J. J. A graphical method for exploring the mean structure in longitudinal data analysis. *Journal of Computational and Graphical Statistics 8* (1999), 60–68.

[20] Fitzmaurice, G. M. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics 80* (1995), 141–151.

[21] Fitzmaurice, G. M., and Laird, N. M. A likelihood-based method for analyzing longitudinal binary responses. *Biometrika 80* (1993), 141–151.

[22] Fox, J. The nmle package. Appendix to *An R and S-PLUS Companion to Applied Regression*, 2002.

[23] Frees, E. W., Young, V. R., and Luo, Y. Case studies using panel data models. *North American Actuarial Journal 5* (2004), 24–42.

[24] Garthwaite, P. H., et al. *Statistical inference*. Prentice Hall, 1995.

[25] Gentleman, R. C. A users guide to panel. Tech. rep., Department of Statistics, University of Auckland, 2006. Manual que acompaña al paquete panel.

[26] Gentleman, R. C., et al. Multi-state markov models for analyzing incomplete disease history data with illustrations for hiv disease. *Statistics in Medicine 13*, 3 (1994), 805–821.

[27] Geyer, C. J. Maximum likelihood in R. Disponible en www.r-project.org, 2003.

[28] Gourieroux, C., Monfort, A., and Trognon, A. Pseudo-maximum likelihood methods: theory. *Econometrica 52* (1984), 681–700.

[29] Halekon, U., Højsgaard, S., and Yan, J. The R package geepack for generalized estimating equations. *Journal of Statistical Software 15*, 2 (2006).

[30] Hall, D. B. On the application of extended quasilikelihood to the clustered data case. *The Canadian Journal of Statistics 29*, 2 (2001), 1–22.

[31] Hardin, J. W., and Hilbe, J. M. *Generalized estimating equations*. Chapman-Hall, 2002.

[32] Heagerty, P. J., and Zeger, S. L. Marginal regresion models for clustered ordinal measurements. *Journal of the American Statistical Association 91* (1996), 1024–1036.

[33] Horton, N. J., and Lipsitz, S. R. Review of software to fit generalized estimating equation regression models. *The American Statistician 53* (1999), 160–169.

[34] Hougaard, P. *Analysis of multivariate survival data*. Springer-Verlag, 2000.

[35] Jackson, C. Multi-state modelling with r: the msm package. Tech. rep., Department of Epidemiology and Public Health, Imperial College, 2006. Manual que acompaña al paquete msm.

[36] Jackson, C., et al. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society, Series D - The Statistician 52*, 2 (2003), 193–209.

[37] Jørgensen, B., et al. State-space models for multivariate longitudinal data of mixed types. *Canadian Journal of Statistics 24* (1996), 385–402.

[38] Kalbfleisch, J. D., and Lawless, J. F. The analysis of panel data under a markov assumption. *Journal of the American Statistical Association 80*, 392 (1985), 863–871.

[39] Kaslow, R. A., et al. The multicenter aids cohort study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology 126* (1987), 310–318.

[40] Kauermann, G. Modelling longitudinal data with ordinal response by varyng coefficients. *Biometrics 56* (2000), 692–698.

[41] Kelly, P. J. A review of software packages for analyzing correlated survival data. *American Statistician 58* (2004), 337–342.

[42] Kittelson, J. M., Sharples, K., and Emerson, S. S. Group sequential clinical trials for longitudinal data with analyses using summary statistics. *Statistics in Medicine 24* (2005), 2457–2475.

[43] Korn, E. L., and Whittemore, A. S. Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics 35* (1979), 795–802.

[44] Laird, N. M., and Ware, J. W. Random-effects models for longitudinal data. *Biometrics 38* (1982), 963–974.

[45] Lee, S. J., Kim, K., and Tsiatis, A. A. Repeated significance testing in longitudinal clinical trials. *Biometrika 83* (1996), 779–789.

[46] Liang, K. Y., and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika 73* (1986), 13–22.

[47] Liang, K. Y., Zeger, S. L., and Qaqish, B. Multivariate regression analysis for categorical data (with discussion). *Journal of the Royal Statistical Society B 54* (1992), 3–40.

[48] Lindstrom, M. J., and Bates, D. M. Newton-raphson and em algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association 83* (1988), 1014–1022.

[49] Marshall, G., and Jones, R. H. Multi-state markov models and diabetic retinopathy. *Statistics in Medicine 14* (1995).

[50] McCullagh, P., and Nelder, J. A. *Generalized linear models.* Chapman-Hall, 1989.

[51] McCulloch, C. E., and Searle, S. R. *Generalized, linear, and mixed models.* Wiley, 2001.

[52] Meester, S. G., and MacKay, J. A parametric model for cluster correlated categorical data. *Biometrics 50* (1994), 954–963.

[53] Nelder, J. A., and Lee, Y. Conditional and marginal models: another view. *Statistical Science 19* (2004), 219–238.

[54] Nelder, J. A., and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society, Series A 135* (1972), 370–384.

[55] Nelsen, R. B. *An introduction to copulas.* Springer-Verlag, 1999.

[56] Neuhaus, J. M. Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research 1* (1992), 249–273.

[57] Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. A comparison of cluster specific and population averaged approaches for analyzing correlated binary data. *International Statistical Review 59* (1991), 25–36.

[58] Nunez-Anton, V., and Zimmerman, D. L. Modeling nonstationary longitudinal data. *Biometrics 56* (2000), 699–705.

[59] Pan, W. Akaike's information criterion in generalized estimating equations. *Biometrics 57* (2001), 120–125.

[60] Pinheiro, J., et al. The nmle package, 2006. Manual que acompaña al paquete nmle.

[61] Pinheiro, J. C., and Bates, D. M. *Mixed-effects models in S and S-PLUS*. Springer, 2000.

[62] Prentice, R. L. Correlated binary regression with covariates specific to each binary observation. *Biometrics 44* (1988), 1033–1048.

[63] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. http://www.R-project.org/.

[64] Rahiala, M. Random coefficient autoregressive models for longitudinal data. *Biometrika 86* (1999), 718–722.

[65] Sharples, L. D., Jackson, C. H., et al. Diagnostic accuracy of coronary angiopathy and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation 76*, 4 (2003), 679–682.

[66] Spiessens, B., et al. An overview of groups sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research 9* (2000), 497–515.

[67] Sribney, W. M. What is the difference between random-effects and population-averaged estimators? http://www.stata.com/support/faqs/stat/repa.html, 1999.

[68] Sutradhar, B. C., and Das, K. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika 86* (1999), 459–465.

[69] Thall, P. F., and Vail, S. C. Some covariance models for longitudinal count data with overdispersion. *Biometrics 46* (1990), 657–671.

[70] Toscas, P. J., and Faddy, M. J. Likelihood-based analysis of longitudinal count data using a generalized poisson model. *Statistical Modelling 3* (2003), 99–108.

[71] Verbyla, A. P., et al. The analysis of designed experiments and longitudinal data using smoothing splines. *Applied Statistics 48* (1999), 269–312.

[72] Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models and the gauss-newton method. *Biometrika 61* (1974), 439–447.

[73] Wong, W. H. Theory of partial likelihhod. *Annals of Statistics 14* (1986), 88–123.

[74] Yan, J. **geepack**: Yet another package for generalized estimating equations. *R News 2* (2002), 12–14.

[75] Zeger, S. L. A regression model for time series of counts. *Biometrika 75* (1988), 621–629.

[76] Zeger, S. L., and Diggle, P. J. Semi-parametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics 50* (1994), 689–699.

[77] Zeger, S. L., and Liang, K.-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics 42* (1986), 121–130.

[78] Zeger, S. L., and Liang, K. Y. A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association 84* (1989), 447–451.

[79] Zeger, S. L., and Liang, K.-Y. Feedback models for discrete and continuous time series. *Statistica Sinica 1* (1991), 51–64.

[80] Zeger, S. L., and Liang, K. Y. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine 11* (1992), 1825–1839.

[81] Zeger, S. L., Liang, K.-Y., and Albert, P. S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics 44* (1988), 1049–1060.

[82] Zeger, S. L., Liang, K.-Y., and Self, S. G. The analysis of binary longitudinal data with time-independent covariates. *Biometrika 72* (1985), 31–38.

[83] Zeger, S. L., and Qaqish, B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics 44* (1988), 1019–1031.

[84] Zheng, B. Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine 19* (2000), 1265–1275.

[85] Zheng, Y., and Heagerty, P. J. Partly conditional survival models for longitudinal data. *Biometrics 61* (2005), 379–391.