PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELO LINEAL MIXTO CONJUNTO DE CLASES LATENTES APLICADO A UN CONJUNTO DE DATOS LONGITUDINALES DEL SECTOR SALUD

TESIS PARA OPTAR POR EL GRADO DE MAGISTER EN ESTADÍSTICA

Presentado por:

Carmen Stéfany Neciosup Vera

Asesor: Dr. Luis Hilmar Valdivieso Serrano

Miembros del jurado:
Dr. Cristian Luis Bayes Rodriguez
Dr. Luis Hilmar Valdivieso Serrano
Dr. Victor Giancarlo Sal y Rosas Celi

Lima, 2018

Dedicatoria

A mis padres, William y Carmen, por su gran amor, confianza y apoyo incondicional.

A mi hermano Oscar, por transmitirme tanta alegría y ser mi confidente desde siempre.

Agradecimentos

Agradezco a mi asesor de tesis, Dr. Luis Valdivieso, por sus enseñanzas, guía, apoyo y respaldo durante este proceso.

Muchas gracias a MSc. Claudio J. Flores, de la Dirección Científica y Académica de Oncosalud-AUNA, por habernos apoyado en la aplicación del modelo desde el punto de vista de la Bioestadística.

Le doy las gracias también a mis profesores de la Maestría en Estadística, por todo el conocimiento que compartieron con nosotros.

Finalmente, agradezco a todas las personas que de alguna u otra manera contribuyeron con el desarrollo de este trabajo.

Resumen

Los modelos lineales mixtos conjuntos de clases latentes, propuestos por Proust-Lima et al. (2015), permiten modelar de manera conjunta un proceso longitudinal y un proceso de supervivencia, calculando también la probabilidad de pertenencia a determinadas clases latentes que puedan existir en la población en estudio. En el presente trabajo se describen los componentes que conforman este modelo, y mediante un estudio de simulación se evalúa y analiza la implementación de su estimación. El modelo se aplica finalmente a un conjunto de datos longitudinales de pacientes diagnosticados con Cáncer de Próstata, permitiéndonos la identificación de clases latentes que se asocian luego con el estadío clínico de los pacientes.

Palabras-clave: JointLCMM, Modelo Lineal Mixto de Clases Latentes, LCMM, Clases Latentes, Modelo de Riesgo Proporcional de Cox, Cáncer de Próstata.

Abstract

The joint latent class mixed model, proposed by Proust-Lima et al. (2015), allows to jointly model a longitudinal process and a survival process, also calculating the probability of belonging to certain latent classes in the study population. In our study, we describe the components that make up this model (Proust-Lima et al. (2017)) and through a simulation study we assesed the implementation of its estimation. The model is finally applied to a set of longitudinal data of Prostate Cancer diagnosed patients allowing us to identify latent classes that are then associated with the clinical stage of the patients.

Keywords: JointLCMM, Latent Class Mixed Model, LCMM, Latent Classes, Proportional Hazard Model, Prostate Cancer.

Índice general

Lista de abreviaturas

AIC	Criterio de información de Akaike.
AR	Proceso Auto-Regresivo.
BIC	Criterio de información Bayesiano.
EC	Estadío clínico.
EM	Algoritmo de la familia denominada *Expectation Maxi-mization*.
FBQ	Falla bioquímica.
JointLCMM	Modelo lineal mixto conjunto de clases latentes.
LCMM	Modelo lineal mixto de clases latentes.
LMM	Modelo lineal mixto.
OU	Proceso de Ornstein-Uhlembeck.
PHM	Modelo de riesgo proporcional.
PSA	Antígeno prostático específico (del inglés, *prostate-specific antigen*).
REML	Estimación por máxima verosimilitud restringida.

Lista de símbolos

μ	Media.
σ^2	Varianza.
ρ	Correlación.
$\stackrel{d}{=}$	Igualdad en distribución.
t_{ij}	Tiempo de observación del sujeto i en el momento j.
α_i	Vector de efectos aleatorios.
ϵ_{ij}	Vector de errores de medición.
$W_i(t_{ij})$	Proceso estocástico Gaussiano.
Λ_{ij}	Variable latente para el i-ésimo elemento en el tiempo T_{ij}
Σ	Matriz de varianzas y covarianzas de los errores de medición.
$\mathbb{H}$	Matriz de varianzas y covarianzas de los efectos aleatorios.
$\lambda_i(t)$	Función de riesgo.
$\lambda_0(t)$	Función de riesgo basal.
$S(t)$	Función de supervivencia en el tiempo t.
$1_{[T_i^* < \tilde{T}_i]}$	Función indicadora de censura.
$L(\Theta)$	Verosimilitud.
$l(\Theta)$	Log-Verosimilitud.
∇	Gradiente.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En diversos campos de las Ciencias se desea explorar la relación existente entre variables respuesta y variables explicativas considerando tanto la correlación entre las medidas tomadas de una misma unidad de análisis, así como la correlación entre las diversas unidades de análisis que pertenecen a una misma clase o familia.

En algunos casos, las variables respuesta y explicativas son observables y medibles directamente de la unidad de análisis, pero en otros casos existen ciertos constructos necesarios para la investigación que no son directamente medibles (por ejemplo, la capacidad de aprendizaje, la ansiedad, la motivación, etc.) pero que pueden ser aproximados mediante *variables manifiestas* u observables. Según Cupani (2012) a este tipo de variables se les conoce con el nombre de *variables latentes*.

Sabemos que los modelos lineales mixtos (LMM por sus siglas en inglés "Linear Mixed Models") permiten explorar la relación existente entre una variable respuesta y una o varias variables predictoras. A la influencia de las segundas sobre las primeras se le denomina *efecto*. Estos pueden ser fijos o aleatorios. Los efectos fijos son valores establecidos de antemano e independientes entre una observación y otra. Por el contrario, los efectos aleatorios son valores que pueden cambiar si la observación se realiza en un contexto diferente (por ejemplo, diferentes tiempos, diferentes lugares), por tanto las variables asociadas a ellos se definen como variables de efectos aleatorios y se asume la existencia de correlación entre las observaciones que comparten un mismo contexto, tiempo o lugar (McCulloch, 2003).

Sea $i=1,\ldots,n$ cada uno de los elementos de estudio, y t_{ij} el tiempo de medición j en el que se observa al elemento i, que sin pérdida de generalidad lo consideraremos como un sujeto. Denotemos a Y_{ij} como el valor de la variable respuesta para el sujeto i, observado en el momento t_{ij} y consideramos un modelo lineal mixto de la forma:

$$Y_{ij} = X_{ij}^T\beta + Z_{ij}^T\alpha_i + \epsilon_{ij},$$

donde X_{ij}^T y Z_{ij}^T son vectores de covariables conocidas observadas en el tiempo t_{ij}, β es un vector de coeficientes de regresión desconocidos llamado de efectos fijos, α_i es el vector de efectos aleatorios, y ϵ_{ij} es el vector de errores de medición.

Esta forma general de un modelo lineal mixto, asume que la relación entre la variable observable respuesta y el tiempo sigue una trayectoria lineal diferente para cada uno de los

sujetos del estudio, con errores independientes y normales en cada punto de observación. Como lo indica Stirrup et al. (2015), esta suposición podría no ser muy adecuada cuando se trata de estudios biomédicos ya que el momento de las observaciones puede ser muy diferente entre y dentro de cada uno de los pacientes o individuos en estudio. Así, si en lugar de utilizar la forma general del modelo lineal mixto se incluyera un componente que permitiera identificar algún patrón en la variabilidad temporal de los datos y describir así el proceso estocástico oculto en este patrón, esto permitiría encontrar mejores estimadores, sin que además tales estimaciones se vieran afectados por la presencia de datos faltantes. Fue por ello que Taylor et al. (1994) propuso la adición de un componente de *Movimiento Browniano* escalado a un modelo lineal mixto de pendiente aleatoria, encontrando que este permitirá mejorar significativamente el ajuste del modelo en términos del Criterio de Información de Akaike (AIC).

El adicionar a un modelo lineal mixto un término $B_{t_{ij}}$ en el tiempo t_{ij} definido como un proceso estocástico Gaussiano o componente de correlación en serie, permite además mejor modelar la correlación intrínseca de los datos. El modelo lineal mixto extendido queda entonces expresado de la forma:

$$Y_{ij} = X_{ij}^T\beta + Z_{ij}^T\alpha_i + B_{t_{ij}} + \epsilon_{ij}$$

De esta forma, el modelo permite observaciones espaciadas y desiguales, y un modelamiento de dependencia más flexible de estas últimas.

Un problema con el modelamiento anterior es que los modelos lineales mixtos solo son aplicables cuando se considera la existencia de una sola variable respuesta observable que sigue además una distribución normal. Si se desea analizar más de una variable respuesta y estas no satisfacen el supuesto de normalidad al ser por naturaleza respuestas binarias, ordinales o acotadas, ya no sería posible trabajar bajo este modelo. Más aún, a pesar de que los LMM son útiles para estudiar el cambio en el tiempo de alguna variable longitudinal, estos modelos no consideran:

1. La existencia de una posible estructura latente en el conjunto de datos.

2. La posibilidad de que una variable respuesta esté alterada por ciertos eventos, tales como la muerte del individuo en estudio, la exclusión voluntaria del estudio, etc.

Una familia de modelos lineales mixtos que permite explorar la existencia de una posible estructura latente en la data son los modelos lineales mixtos de clases latentes (Commenges y Jacqmin-Gadda, 2015), los cuales además de incorporar al estudio constructos no observables como respuestas, admiten también la incorporación de variables manifiestas a estos constructos no necesariamente de distribución normal. Al igual que en la teoría de las ecuaciones estructurales, estos modelos poseen dos componentes: un modelo de medición y un modelo estructural. El modelo de medición es aquel que vincula las variables manifiestas con la variable dependiente no observable; mientras que el modelo estructural indica la forma en la que el proceso latente es explicado de acuerdo al tiempo y a las covariables. El proceso latente en su componente estructural está definido como un modelo lineal mixto sin error de

medición, de la forma:

$$\Lambda_{ij} = X_{ij}^T\beta + Z_{ij}^T\alpha_i + B_{t_{ij}} \tag{1.1}$$

donde Λ_{ij} denota a la variable latente de interés para el i-ésimo sujeto en el tiempo t_{ij}.

Con la finalidad de estudiar esta última variable a través de una variable manifiesta no necesariamente de distribución normal, el modelo de medición vincula al proceso latente subyacente con la variable manifiesta Y_{ij} en el tiempo t_{ij}, (Proust-Lima et al., 2015) mediante:

$$Y_{ij} = g(\tilde{Y}_{ij};\eta) = g(\Lambda_{ij} + \epsilon_{ij};\eta) \tag{1.2}$$

siendo g una función de enlace con inversa, monótona, creciente y continua e $\tilde{Y}_{ij}$ un proceso de ruido latente de la forma:

$$\tilde{Y}_{ij} = \Lambda_{ij} + \epsilon_{ij}$$

y η un conjunto de parámetros del modelo. Esta notación se puede extender para más de una variable respuesta.

De otro lado, los modelos lineales mixtos de clases latentes (LCMM por sus siglas en inglés "Latent Class Mixed Model") incorporan dos relaciones primordiales: la probabilidad de pertenencia de un sujeto a una determinada clase latente a través de un modelo logístico multinomial en el que ℓ identifica la clase a la que pertenece el sujeto i y $\tilde{X}_i$ es un vector de covariables asociadas a la clase, como lo muestra (1.3); y un modelo lineal mixto específico para cada clase latente, el cual se diferencia del LMM estándar porque ahora tanto los efectos fijos como aleatorios son específicos para cada una de las clases latentes ℓ, como se muestra en (1.4) y en donde incluso se identifican covariables generales (X_{1ij}) y covariables de clase (X_{2ij}) para los efectos fijos.

$$\pi_{i\ell} = \frac{e^{\tilde{X}_i^T\xi_\ell}}{1 + e^{\tilde{X}_i^T\xi_1} + e^{\tilde{X}_i^T\xi_2} + ... + e^{\tilde{X}_i^T\xi_{(L-1)}}} \tag{1.3}$$

$$Y_{ij}|_{C_i=\ell} = X_{1ij}^T\beta + X_{2ij}^T\nu|_{C_i=\ell} + Z_{ij}^T\alpha_i|_{C_i=\ell} + B_{t_{ij}} + \epsilon_{ij} \tag{1.4}$$

Un caso especial de los modelos lineales mixtos de clases latentes se presenta cuando el proceso longitudinal está estrechamente asociado con un proceso de supervivencia; en otras palabras, cuando la variable respuesta es afectada por eventos tales como muerte, exclusión voluntaria, etc. En estos casos, el modelo sufre una modificación para poder capturar esta correlación adicional. Este submodelo se denomina modelo lineal mixto conjunto de clases latentes (JLCMM por sus siglas en inglés "Joint latent class mixed model"). Se basa en la idea de que para cada una de las clases latentes existe un modelos lineal mixto para el proceso longitudinal y un modelo de supervivencia para el proceso de supervivencia (Proust-Lima et al., 2015). De este modo, el JLCMM consta de cuatro submodelos, los tres primeros son los mismos que los del LCMM: la probabilidad de pertenencia de un individuo a una

determinada clase latente a través del modelo logístico multinomial (1.3), la de un modelo lineal mixto específico para cada clase de ruido latente (1.4) y la de su modelo de medición correspondiente (1.2). Adicionalmente se presenta un cuarto submodelo, el indicado en (1.5), el cual hace referencia a un modelo de supervivencia para cada una de las clases latentes.

El beneficio adicional de modelar un proceso de supervivencia en los modelos JLCMM es que nos permite utilizar otra variable respuesta que afecta directamente a la variable respuesta longitudinal pero que, luego de establecerse la variable latente, ambas variables respuesta se vuelven independientes condicionalmente a esta variable latente. Para cada una de las clases de la variable latente se calcula el riesgo de ocurrencia de un determinado evento mediante un modelo de riesgo proporcional de Cox, de la forma:

$$\lambda_i(t)|_{C_i=\ell} = \lambda_0(t)|_{C_i=\ell}\, e^{\check{X}_{1i}^T\delta + \check{X}_{2i}^T\delta_\ell} \tag{1.5}$$

En el presente proyecto de tesis, se buscará aplicar este modelo al campo de la salud debido a que, tal como lo señala Pedrero et al. (2015), actualmente no existen muchos trabajos de este tipo en nuestra región. Para esto, se empleará el paquete *"lcmm"* (Proust-Lima et al., 2017) el cual se encuentra implementado en el software libre *R*.

1.2. Objetivos

El objetivo general de la tesis es presentar el modelo lineal mixto conjunto de clases latentes, estudiar sus fundamentos y propiedades y aplicarlo a un problema de nuestro entorno en el sector salud. De manera específica:

- Realizar un estudio preliminar de los modelos lineales mixtos, la incorporación de un proceso estocástico en él, los modelos de variables latentes, los modelos lineales mixtos de clases latentes y los modelos de supervivencia.
- Desarrollar el modelo teórico de los modelos lineales mixtos conjuntos de clases latentes.
- Realizar un estudio de simulación para evaluar el desempeño de los estimadores del modelo lineal mixto conjunto de clases latentes.
- Aplicar el modelo lineal mixto conjunto de clases latentes a una base de datos con información de pacientes diagnosticados con cáncer de próstata, de quienes se registraron sus valores de PSA en el tiempo luego de haber sido intervenidos mediante una prostatectomía radical.

1.3. Organización del trabajo

En el Capítulo 2, presentamos un estudio preliminar al desarrollo de los modelos lineales mixtos conjuntos de clases latentes, el cual incluye los modelos lineales mixtos, modelos lineales mixtos con componentes estocásticos de error, modelos de variables latentes, modelos lineales mixtos de clases latentes y modelos de supervivencia. En el Capítulo 3, se muestra el estudio de las propiedades, procedimiento de estimación de parámetros y análisis de bondad de ajuste de los modelos lineales mixtos conjuntos de clases latentes. El Capítulo 4 presenta

el método de estimación de los parámetros del modelo. El Capítulo 5 presenta un estudio de simulación en diferentes escenarios, comparando el modelo lineal mixto conjunto de clases latentes con modelos tradicionales. En el Capítulo 6, se muestra la aplicación del modelo lineal mixto conjunto de clases latentes a una base de datos de pacientes diagnosticados con cáncer de próstata, de quienes se analiza la trayectoria en el tiempo de sus valores de PSA en la sangre, y de manera conjunta se modela un proceso de supervivencia mediante el análisis del tiempo transcurrido desde la prostatectomía radical hasta el momento en el que el valor de PSA del paciente supera los 0.20 ng/ml en la sangre. En el Capítulo 7 discutimos algunas conclusiones obtenidas en este trabajo, analizando las ventajas y desventajas del modelo lineal mixto conjunto de clases latentes. Incluimos finalmente un anexo presentando los programas utilizados para el estudio de simulación y para la aplicación al conjunto de datos reales (Apéndice A).

Capítulo 2

Estudio preliminar

2.1. Modelos lineales mixtos

Los modelos lineales mixtos (LMM) permiten estudiar la relación existente entre determinadas variables explicativas sobre una variable respuesta haciendo posible inclusive, mediante sus componentes aleatorias, el análisis del cambio de esta variable respuesta en el tiempo o entre conglomerados. Para poder realizar este análisis en el contexto temporal, uno de los principales supuestos del modelo es que la variable respuesta debe tener carácter longitudinal y seguir una distribución normal. Además, estos modelos asumen la existencia de correlación en los datos, correlación tanto entre observaciones de un mismo individuo como correlación entre individuos pertenecientes a un mismo grupo o familia (McCulloch, 2003).

A la influencia de las variables explicativas sobre la variable respuesta se le llama *efecto*. Los LMM consideran dos tipos de efectos en su estructura: efectos fijos y efectos aleatorios. Los *efectos fijos* indican la influencia de la variación en cada variable explicativa sobre la variable respuesta, siendo este efecto idéntico para todos los sujetos en estudio. Por otro lado, los *efectos aleatorios* son diferentes para cada sujeto y se asume que cada individuo en estudio tiene un nivel diferente de respuesta "no observado" que persiste en todas sus mediciones. Este efecto es tratado como aleatorio; además, es posible considerar no solo que los sujetos varían en la respuesta desde el tiempo inicial (intercepto aleatorio), sino que también en sus tasa de cambio (pendientes aleatorias).

En general, sea Y_{ij} el valor de la variable respuesta para el sujeto i en la medición j la cual se da en el tiempo t_{ij}, siendo $i = 1, 2, ..., n$, $j = 1, 2, ..., n_i$. Aquí n_i denota al número de mediciones del sujeto i y n al número total de sujetos en el estudio. La medición j corresponderá en nuestro contexto al tiempo t_{ij} en que se registre la variable respuesta para el sujeto i. Sea también X_{ij}^T el vector de covariables del sujeto i observadas en el tiempo t_{ij} asociado al vector de efectos fijos β, y sea Z_{ij}^T el vector de covariables del sujeto i observadas en el tiempo t_{ij} asociado al vector de efectos aleatorios α_i. Tenemos que un modelo lineal mixto toma la forma:

$$Y_{ij} = X_{ij}^T\beta + Z_{ij}^T\alpha_i + \epsilon_{ij} \tag{2.1}$$

donde

$$X_{ij}^T = [X_{1ij}, X_{2ij}, ..., X_{pij}]$$
$$Z_{ij}^T = [Z_{1ij}, Z_{2ij}, ..., Z_{qij}]$$

De modo matricial, el LMM se puede representar de la forma:

$$\mathbf{Y}_i = \mathbb{X}_i\beta + \mathbb{Z}_i\alpha_\mathbf{i} + \epsilon_\mathbf{i} \qquad , \tag{2.2}$$

donde

$$\mathbf{Y}_i := \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}_{n_i \times 1} \quad , \quad \mathbb{X}_i := \begin{pmatrix} X_{1ij} & X_{2ij} & \cdots & X_{pij} \\ \vdots & \vdots & & \vdots \\ X_{1in_i} & X_{2in_i} & \cdots & X_{pin_i} \end{pmatrix}_{n_i \times p}$$

$$\mathbb{Z}_i := \begin{pmatrix} Z_{1ij} & Z_{2ij} & \cdots & Z_{pij} \\ \vdots & \vdots & & \vdots \\ Z_{1in_i} & Z_{2in_i} & \cdots & Z_{pin_i} \end{pmatrix}_{n_i \times q} \quad , \quad \alpha_\mathbf{i} := \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{iq} \end{pmatrix}_{q \times 1} \quad y \quad \varepsilon_\mathbf{i} := \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}_{n_i \times 1} .$$

Se asume además que $\alpha_\mathbf{i}$ y $\varepsilon_\mathbf{i}$ son vectores aleatorios, independientes, con distribuciones normales multivariadas:

$$\alpha_\mathbf{i} \sim N_q\left(0, \mathbb{H}\right)$$
$$\varepsilon_\mathbf{i} \sim N_{n_i}\left(0, \Sigma\right)$$

Se tiene entonces que $\mathbf{Y}_i \sim N_{n_i}\left(\mathbb{X}_i\beta, \mathbb{V}_i\right)$, donde

$$\mathbb{V}_i = \mathbb{Z}_i\mathbb{H}\mathbb{Z}_i^T + \Sigma$$

La estimación de los valores del vector β, podría realizarse mediante el método de máxima verosimilitud, el cual maximiza la función de distribución conjunta de $\mathbf{Y} = [Y_1, Y_2, ..., Y_n]^T$ y en donde $\mathbb{V}_i$ depende de un vector de parámetros γ. Tenemos aquí una función de verosimilitud de la forma:

$$L\left(\beta,\gamma\right) = \prod_{i=1}^{n} (2\pi)^{-\frac{n}{2}} \left|\mathbb{V}_i\left(\gamma\right)\right|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\mathbf{Y}_i - \mathbb{X}_i\beta)^T {\mathbb{V}_i}^{-1}\left(\gamma\right)\left(\mathbf{Y}_i - \mathbb{X}_i\beta\right)\right) \tag{2.3}$$

siendo $\mathbf{Y}_i$ el vector observado de variables respuesta para el sujeto i.

Esta expresión podría presentar complicaciones computacionales si se tratase con un gran número de parámetros de efectos aleatorios; además, los estimadores por máxima verosimilitud no resultan insesgados, y este sesgo podría resultar significativo.

Otra forma de estimar los parámetros es mediante el método de máxima verosimilitud restringida (REML). La diferencia respecto a la anterior es que esta se realiza por partes. La estimación por REML pretende primero encontrar una combinación lineal de las componentes de $\mathbf{Y}$ cuya media sea 0, de modo que el vector aleatorio resultante solo dependa de γ. Si definimos:

$$\mathbb{Q} = \mathbb{I} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \tag{2.4}$$

la combinación en mención viene dada por $\mathbf{U} = \mathbb{Q}\mathbf{Y}$, siendo $\mathbf{U} \sim N(0, \mathbb{W})$, donde:

$$\mathbb{W} = Var(\mathbf{U}) = \mathbb{Q}Var(\mathbf{Y})\mathbb{Q}^T = \mathbb{Q}\mathbb{V}\mathbb{Q}^T \tag{2.5}$$

El estimador de máxima verosimilitud restringida de γ, $\hat{\gamma}_{REML}$, se obtiene al maximizar el logaritmo de la función de verosimilitud de $\mathbf{U}$, la cual es de la forma:

$$\log L_{REML}(\gamma) \propto -\frac{1}{2}\left(\log|\mathbb{W}(\gamma)| + \mathbf{Y}^T\mathbb{Q}^{-1}\mathbb{W}(\gamma)^{-1}\mathbb{Q}\mathbf{Y}\right) \tag{2.6}$$

Finalmente, el estimador de β, mediante este modelo viene dado por el estimador de mínimos cuadrados generalizados, luego de reemplazar en ella la estimación anterior; esto es,

$$\hat{\beta}_{REML} = \hat{\beta}_{GLS}(\hat{\gamma}_{REML}) = \left(\mathbb{X}^T\mathbb{W}^{-1}(\hat{\gamma}_{REML})\,\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{W}^{-1}(\hat{\gamma}_{REML})\,\mathbf{Y} \tag{2.7}$$

2.2. Modelos lineales mixtos con componente estocástico de error

En diferentes estudios, y específicamente en los del sector salud, la cantidad de medidas tomadas de las unidades de análisis pueden variar; así como también, los tiempos en los que se toman las medidas. De otro lado, es preciso considerar la existencia de correlación entre observaciones provenientes de un mismo individuo.

Con la finalidad de considerar estos detalles al modelar las trayectorias de los individuos y con el objetivo de describir el proceso estocástico subyacente en las trayectorias descritas por el modelo, se incluirá en el LMM un componente estocástico que permita identificar algún patrón en la variabilidad de los datos. El modelo en cuestión toma la forma:

$$Y_{ij} = X_{ij}^T\beta + Z_{ij}^T\alpha_i + B_{t_{ij}} + \epsilon_{ij} \tag{2.8}$$

donde, $B_{t_{ij}}$ es un proceso estocástico que se evalúa en t_{ij}, que podría ser de ruido blanco, un proceso de Ornstein-Uhlenbeck de orden 1, o un movimiento Browniano también llamado proceso de Wiener.

Un proceso estocástico, recordemos, es una familia de variables aleatorias $\{W_t\}$ indexadas por el parámetro temporal t, cuyas características pueden variar en el tiempo. Dado que este parámetro temporal puede observarse continua o discretamente, los procesos estocásticos se clasifican en cuatro tipos como lo muestra la Figura 2.1 .

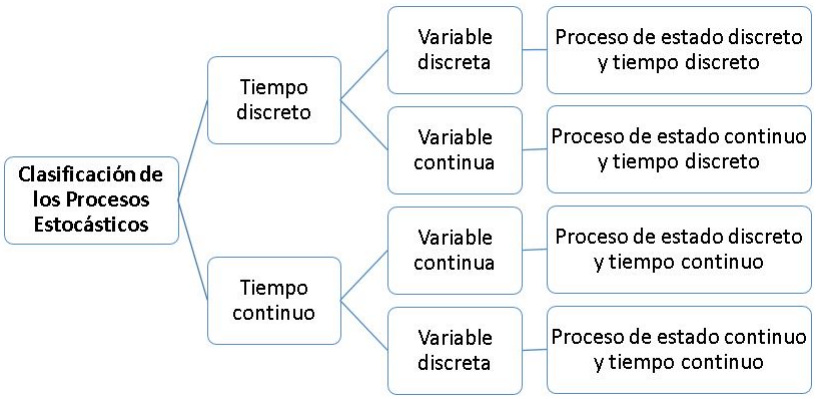

Figura 2.1: Clasificación de los procesos estocásticos

En un proceso estocástico, las medias, varianzas y covarianzas ahora son funciones del tiempo, y pueden variar en él. Así, las principales características de un proceso estocástico $\{W_t\}$ se resumen en:

- Función de medias: $\mu_t = E(W_t)$
- Función de varianzas: $\sigma_t^2 = Var(W_t)$
- Función de autocovarianzas: $Cov(W_t, W_s) = Cov(W_s, W_t)$
- Función de autocorrelación: $\rho_{t,s} = \frac{Cov(W_t,W_s)}{\sigma_t \sigma_s}$

Además, se dice que un proceso tiene incrementos estacionarios cuando las funciones antes mencionadas solo dependen de la distancia entre observaciones y no del instante considerado, cumpliéndose entonces que: $Cov\left(W_t, W_{t+j}\right) = Cov\left(W_s, W_{s+j}\right), \quad \forall j$. Posteriormente se podrá observar que esta condición resulta ser de gran utilidad al momento de tratar datos longitudinales, ya que solo interesa la distancia entre instantes y no el instante en sí, haciendo que la ausencia de datos o la no igualdad en la cantidad de observaciones ya no sea un problema al momento de analizar los datos.

Dentro del grupo de los procesos estocásticos, existe un subconjunto que cumple un determinado supuesto en común, el cual asume que las características transversales del proceso (las funciones listadas anteriormente) permanecen estables a lo largo del tiempo. A este subconjunto de procesos se les denomina *Procesos Estocásticos Estacionarios*. La denominación de estacionarios indica que su comportamiento es constante a lo largo del tiempo.

Se dice que un proceso estocástico es estacionario en un sentido estricto si al realizar un mismo desplazamiento en el tiempo de cualquier subconjunto de sus variables, su distribución no varía. Esto es, para cualquier $i_1, i_2, ..., i_r$, t y r se cumple que las funciones de distribución satisfacen :

$$F_{W_{i_1}, W_{i_2}, \ldots, W_{i_r}} = F_{W_{i_1+t}, W_{i_2+t}, \ldots, W_{i_r+t}} \tag{2.9}$$

Por otro lado, adoptando una condición no tan restrictiva, se dice que un proceso estocástico es débilmente estacionario si mantiene constantes parte o todas sus características a lo largo del tiempo. Estos pueden ser estacionarios de 1º orden o estacionarios de 2º orden.

$$\begin{aligned}
&\text{Estacionario de } 1^{\text{o}} \text{ orden:} && \mu_t = \mu \\
&\text{Estacionario de } 2^{\text{o}} \text{ orden:} && \mu_t = \mu \\
& && \sigma_t^2 = \sigma^2 \\
& && Cov\,(W_t, W_{t+j}) = \gamma_j = Cov(W_j) \\
& && \rho_j = \frac{\gamma_j}{\gamma_0} = \frac{\gamma_j}{\sigma^2}
\end{aligned}$$

La función de autocorrelación simple ρ_j resulta para el último caso ser simétrica ($\rho_j = \rho_{-j}$) y solo depende de la distancia o retardo entre los instantes de tiempo. La representación gráfica de esta función es el correlograma.

2.2.1. Procesos de ruido blanco

El proceso de ruido blanco es un proceso estocástico estacionario $\{\epsilon_t\}$ donde $t \geq 0$, con las siguientes características:

- $E\,(\epsilon_t) = 0$
- $Var\,(\epsilon_t) = \sigma^2$
- $Cov\,(\epsilon_t, \epsilon_s) = 0$; si $t \neq s$

Un proceso de ruido blanco es entonces una sucesión de valores sin relación alguna y que oscilan en torno a 0 y en un margen constante.

2.2.2. Procesos Browniano o proceso de Wiener

Un proceso o movimiento Browniano, modela en Física al movimiento aleatorio que se observa en partículas solidas pequeñas que se encuentran inmersas en un medio fluido como resultado de los choques contra las moléculas del fluido. En el campo del estudio de probabilidades, a este proceso también se le conoce como *proceso de Wiener*, el cual es un proceso estocástico de tipo Gaussiano $\{B_t\}$ que toma valores continuos. Formalmente, un movimiento Browniano es un proceso de estado continuo y tiempo continuo que satisface para cualquier $0 < s < t$ lo siguiente:

(a) Comienza en el origen con probabilidad 1 (a.s.)
$B_0 = 0$, a.s.
$P(B_0 = 0) = 1$

(b) Los incrementos del proceso son Gaussianos de media 0 y varianza proporcional a la longitud del incremento, esto es,
$B_t - B_s \sim N(0, \sigma^2(t-s))$.

(c) $B_{t_1}, B_{t_2} - B_{t_1}, ..., B_{t_n} - B_{t_{n-1}}$ son variables aleatorias independientes para cualquier $0 \leq t_1 \leq t_2 \leq ... \leq t_n$.

(d) B_t tiene incrementos estacionarios, esto es,
$B_{t+s} - B_s \stackrel{d}{=} B_t$
donde $\stackrel{d}{=}$ indica igualdad en distribución.

Algunas propiedades interesantes de este proceso son:

- Función covarianza: $Cov\left(B_t, B_s\right) = min(s,t)\sigma^2$
- B_t es $\frac{1}{2}$ autosemejante $B_{st} \stackrel{d}{=} \sqrt{s}B_t$
- B_t tiene trayectorias continuas a.s., pero no diferenciables en ningún punto.

2.2.3. Proceso de Ornstein-Uhlenbeck

Un proceso de Ornstein-Uhlenbeck (OU) puede ser considerado como una versión en tiempo continuo de un proceso AR(1). Para entenderlo mejor, empezaremos definiendo este último.

Los procesos auto-regresivos de orden AR(1) representan la influencia de un hecho pasado sobre el inmediato siguiente, expresando el valor actual del proceso mediante un modelo de regresión lineal en función del valor anterior del mismo proceso, considerando una constante δ, un parámetro a que está relacionado con la memoria de la serie (a medida que el valor de a aumenta, la dependencia con respecto al pasado es más fuerte) y una perturbación aleatoria con distribución $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. El modelo se expresa por:

$$X_t = \delta + aX_{t-1} + \epsilon_t \tag{2.10}$$

y es usual considerar el caso $\delta = 0$. Un proceso de Ornstein-Uhlenbeck es una generalización del proceso último anterior pero en tiempo continuo, el cual toma la forma:

$$dX_t = -\lambda X_t dt + \sigma dB_t \qquad , \qquad donde \quad \lambda \quad 0 >$$

Esta ecuación diferencial estocástica tiene como solución a

$$X_t = X_0 exp\left(-\lambda t\right) + \sigma \int_0^t exp\left(-\lambda\left(t-s\right)\right) dB_s \qquad , \tag{2.11}$$

lo cual muestra que es Gaussiano (Valdivieso, 2007). El proceso además es estacionario y posee las propiedades siguientes :

$$E\left(X_t\right) = exp\left(-\lambda t\right) E\left(X_0\right) \tag{2.12}$$

$$V\left(X_t\right) = \frac{\sigma^2}{2\lambda} + \left(V\left(X_0\right) - \frac{\sigma^2}{2\lambda}\right) exp\left(-2\lambda t\right) \tag{2.13}$$

$$Cov\left(X_t, X_{t+h}\right) = \left(V\left(X_0\right) + \frac{\sigma^2}{2\lambda}\left(exp\left(2\lambda t\right) - 1\right)\right) exp\left(-\lambda\left(2t+h\right)\right) \tag{2.14}$$

2.3. El modelo de riesgo proporcional de Cox

Cuando se trata de estudios longitudinales, uno de los aspectos de mayor interés es poder medir, a través de probabilidades, la posibilidad de que a un individuo le ocurra un determinado evento a lo largo del tiempo. Una función del tiempo que permite modelar esto es la función de riesgo instantáneo. Si T^* es la v.a. que modela el tiempo de ocurrencia del evento de interés, la función de riesgo para esta variable aleatoria se define por:

$$\lambda(t) = \lim_{h \to 0} \frac{P(t < T^* < t + h)}{h} = \frac{f(t)}{S(t)}$$

siendo f y S las funciones de densidad y supervivencia de T^*, respectivamente.

Es importante considerar que esta función puede ser expresada también en base a la función del riesgo acumulado $\Lambda(t) = \int_0^t \lambda(s)ds$, de este modo, la relación entre la función de riesgo y la función de supervivencia queda expresada como se muestra a continuación.

$$\begin{aligned}
\lambda(t) &= \frac{f(t)}{S(t)} \\
\lambda(t) &= -\frac{d}{dt}\log\left(S(t)\right) \\
\int \lambda(t) &= \int -\frac{d}{dt}\log\left(S(t)\right) \\
-\Lambda(t) &= \log\left(S(t)\right) \\
S(t) &= e^{-\Lambda(t)}
\end{aligned} \tag{2.15}$$

Note que según este desarrollo la función de densidad puede escribirse como:

$$f(t) = \lambda(t)\, e^{-\Lambda(t)} \tag{2.16}$$

El modelo de riesgo proporcional de Cox (Cox, 1972) permite construir un modelo de regresión en función de covariables dependientes o independientes del tiempo. Este se expresa para un sujeto i por:

$$\lambda_i(t) = \lambda_0(t) e^{X_i^T \delta} \tag{2.17}$$

donde:

- $\lambda_0(t)$ es una función de riesgo basal en el tiempo t
- X_i es un vector columna de covariables para el sujeto i
- δ es un vector de parámetros de regresión del modelo

El modelo de regresión de Cox, dado en (2.17), indica que existe una misma función de riesgo basal para todos los sujetos en el modelo. En nuestro caso, sin embargo, estos podrán pertenecer a diferentes clases o estratos y por ello sería más adecuado considerar para cada clase una función de riesgo basal diferente.

Algunas veces, no es posible observar el evento de interés en todos los sujetos. Esto se puede deber a que, por ejemplo, algún sujeto desistió de participar en el estudio, terminó el estudio o quizás el sujeto no presentó el evento de interés. A estos sucesos se les conoce como *censura*, ya que por alguna razón, que no es posible controlar, no se pudo observar el evento de interés en el sujeto durante el periodo de observación. De esta manera, el tiempo de observación sobre el evento de interés para el sujeto i se define como $T_i = min(T_i^*, \tilde{T}_i)$, donde T_i^* es el tiempo hasta que ocurra el evento de interés en el sujeto i y $\tilde{T}_i$ es el tiempo de censura, donde no se observa T_i^*.

Además, se tiene una variable indicadora E_i que da a conocer si se observó en el sujeto el evento o si se observó una censura, siendo $E_i = 1_{[T_i^* < \tilde{T}_i]}$, el cual toma el valor de 1 si se observó al evento de interés y 0 en caso contrario.

La función de verosimilitud para (2.17), en base a (2.15) y (2.16), se define por:

$$\begin{aligned} L_n\left(\delta, \lambda_0(t)\right) &= \prod_{i=1}^{n} f(t_i)^{E_i} P(T_i > t_i)^{1-E_i} \\ &= \prod_{i=1}^{n} \left(\lambda_i(t_i) e^{-\Lambda(t_i)}\right)^{E_i} S(t_i)^{1-E_i} \\ &= \prod_{i=1}^{n} \left(\lambda_i(t_i) e^{-\Lambda(t_i)}\right)^{E_i} \left(e^{-\Lambda(t_i)}\right)^{1-E_i} \\ &= \prod_{i=1}^{n} \lambda_i(t_i)^{E_i} \; e^{-\Lambda(t_i)} \\ &= \prod_{i=1}^{n} \left(\lambda_0(t_i) e^{X_i^T \delta}\right)^{E_i} \; e^{-\int_0^{t_i} \lambda_0(s_i) e^{X_i^T \delta} ds} \end{aligned} \tag{2.18}$$

donde n es el número de sujetos en estudio, t_i es el tiempo observado del evento de interés o censura y $\lambda_0(t_i)$ es la función de riesgo basal en t_i.

En el presente estudio se considerará una función de riesgo basal paramétrica, la cual estará definida en base a un vector de parámetros ζ. Algunas de las funciones paramétricas de riesgo basal, parametrizadas por un vector de parámetros ζ son:

- La función de riesgo Weibull, definida por:

$$\lambda_0(t) \equiv \lambda_0(t;\zeta) = \zeta_1 \zeta_2 t^{\zeta_2 - 1} \tag{2.19}$$

- Función de riesgo constante por partes:

$$\lambda_0(t) \equiv \lambda_0(t;\zeta) = \sum_{d=1}^{n_z - 1} \zeta_d 1_{t \epsilon [t_d, t_{d+1}]} \tag{2.20}$$

siendo n_z el número de *nodos* y $1_{t\epsilon[t_d,t_{d+1}]}$ una función indicadora, donde 1 indica que t se encuentra en el intervalo $[t_d, t_{d+1}]$, y 0 indica lo contrario.

Capítulo 3

Modelo lineal mixto conjunto de clases latentes

Los modelos lineales mixtos de clases latentes son una familia de modelos que permiten explorar la existencia de una posible estructura latente en la data (Commenges y Jacqmin-Gadda, 2015) y que admiten además variables respuestas no necesariamente con distribución normal.

Una extensión de este modelo es el modelo lineal mixto conjunto de clases latentes (JLCMM). El JLCMM se caracteriza porque permite modelar el proceso de interés considerando además que este puede estar asociado a un segundo proceso de supervivencia. De esta manera el modelo permite capturar la correlación existente entre ellos.

Estos modelos asumen que cada clase latente ℓ se caracteriza por un modelo de medición y un modelo estructural; último que está conformado por: un submodelo lineal mixto para el proceso longitudinal, un submodelo logístico multinomial para la probabilidad de pertenencia a la clase y un submodelo de riesgo proporcional de Cox específico para cada clase latente.

Al igual que en la teoría de ecuaciones estructurales, el modelo posee un modelo de medición y uno estructural. El modelo de medición expresa la relación existente entre las variables latentes y las variables observables, y el modelo estructural relaciona el proceso latente con el tiempo y con las covariables. Dada la complejidad del modelo, y las características de nuestra aplicación, en este trabajo se considerará en detalle solo el modelo estructural, asumiendo que nuestra variable dependiente será observable y con distribución normal.

3.1. El submodelo lineal mixto de clases latentes

Las características de las trayectorias de la variable dependiente Y dentro de una clase latente ℓ son representadas mediante un modelo lineal mixto por clase, de la forma:

$$Y_{ij}|_{C_i=\ell} = X_{1ij}^T\beta + X_{2ij}^T\upsilon|_{C_i=\ell} + Z_{ij}^T\alpha_i|_{C_i=\ell} + B_{t_{ij}} + \epsilon_{ij} \tag{3.1}$$

siendo:

- $j = 1, 2, ..., n_i$ el periodo de observación para el sujeto i, el cual se da en un tiempo t_{ij}
- X_{1ij}^T un vector de covariables asociadas a los efectos fijos comunes del i-ésimo sujeto en el periodo j
- X_{2ij}^T un vector de covariables asociadas a los efectos fijos por clase ℓ, del i-ésimo sujeto en el periodo j

- Z_{ij}^T un vector de covariables asociadas a los efectos aleatorios, del i-ésimo sujeto en estudio en el periodo j
- β un vector de parámetros de efectos fijos
- C_i una variable aleatoria con rango $1, 2, ... \ell$ que nos indica la clase de pertenencia del sujeto i
- $\nu|_{C_i=\ell}$ un vector de parámetros de efectos fijos para la clase ℓ
- $\alpha_i|_{C_i=\ell}$ un vector aleatorio de efectos aleatorios para el i-ésimo sujeto en la clase ℓ
- $B_{t_{ij}}$ un proceso estocástico Gaussiano evaluado en el tiempo t_{ij}
- ϵ_{ij} un vector aleatorio de errores de medición para el sujeto i en el periodo j

El modelo asume que $\alpha_i|_{C_i=\ell} \sim N_q(0, w_\ell^2 \mathbb{H})$ y $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ son elementos aleatorios independientes.

Como se explicó en el capítulo anterior, el proceso Gaussiano de media 0, $\{B_t\}$ podría ser: un ruido blanco, un movimiento Browniano con estructura de covarianza: ($Cov(B_t, B_s) = \sigma_\omega^2\ min(t, s)$) o un proceso de Ornstein-Uhlenbeck de orden 1, con estructura de covarianza:

$$Cov\left(B_t, B_{t+h}\right) = \left(\frac{\sigma_w^2}{2\lambda}\left(exp\left(2\lambda t\right) - 1\right)\right) exp\left(-\lambda\left(2t + h\right)\right)$$

donde $\lambda > 0$ es un parámetro del modelo.

El beneficio de adicionar el componente estocástico es poder modelar mejor la correlación serial adicional entre los datos debida a la aproximación o distancia entre las observaciones, permitiendo así obtener estimaciones más flexibles para la correlación.

3.2. El submodelo de regresión logística multinomial

En este modelo se busca explicar las probabilidades de pertenencia a la clase ℓ por parte de un sujeto i en el vector multinomial $C_i \sim Mult\left(1, \pi_1, \pi_2, ..., \pi_\ell\right)$ mediante un vector de covariables $\tilde{X}_i$ por:

$$\pi_{i\ell} = \frac{e^{\tilde{X}_i^T \xi_\ell}}{1 + e^{\tilde{X}_i^T \xi_1} + e^{\tilde{X}_i^T \xi_2} + ... + e^{\tilde{X}_i^T \xi_{(L-1)}}} \tag{3.2}$$

donde ξ_0 es un parámetro de intercepto y ξ_ℓ es un vector de coeficientes de regresión para la clase latente ℓ, asumiendo que existen L clases latentes con $\ell = 1, ..., L-1$. Aquí L define la clase de referencia del modelo y para ella se cumple:

$$\pi_{iL} = \frac{1}{1 + e^{\tilde{X}_i^T \xi_1} + e^{\tilde{X}_i^T \xi_2} + ... + e^{\tilde{X}_i^T \xi_{(L-1)}}} \tag{3.3}$$

3.3. El submodelo de supervivencia

Algunas veces no es posible asumir que la probabilidad de la ausencia de un valor de la variable dependiente en el modelo (3.1) se encuentre explicada por las observaciones. Algunas veces, la ausencia de valores de esta variable se puede deber a que el proceso longitudinal puede estar asociado a un segundo proceso de supervivencia en el que se presenta algún tipo de censura.

Los modelos lineales mixtos conjuntos de clases latentes permiten capturar la correlación existente entre el proceso longitudinal y el proceso de supervivencia hacia un evento de interés.

Se propone entonces un modelo de riesgo proporcional de Cox para el tiempo observado del evento de interés que permite modelar su riesgo en cualquier instante t, dada una determinada clase latente, como:

$$\lambda_i(t)|_{C_i=\ell} = \lambda_0(t)|_{C_i=\ell} \quad e^{\breve{X}_{1i}^T\delta + \breve{X}_{2i}^T\delta|_{C_i=\ell}} \tag{3.4}$$

siendo:

- $\lambda_i(t)|_{C_i=\ell}$ el riesgo de que al individuo i, perteneciente a la clase latente ℓ, le ocurra el evento de interés o la censura en el tiempo t
- $\lambda_0(t)|_{C_i=\ell}$ una función de riesgo basal para la clase ℓ en el tiempo t
- $\breve{X}_{1i}$ un vector de covariables asociadas a los parámetros comunes
- $\breve{X}_{2i}$ un vector de covariables asociadas a los parámetros específicos por clase ℓ
- δ un vector de parámetros fijos comunes
- $\delta|_{C_i=\ell}$ un vector de parámetros fijos específicos por clase ℓ

El modelo de riesgo proporcional calcula el riesgo de ocurrencia de un determinado evento estratificándolo mediante la función de riesgo basal, considerando diferentes riesgos basales para cada clase o estrato ℓ, quedando expresado este como lo muestra (3.4).

Este modelo, puede ser expresado también en base a la función de riesgo acumulado, resultando lo que se indica en (3.5).

$$\Lambda_i(t)|_{C_i=\ell} = \Lambda_0(t)|_{C_i=\ell} \quad e^{\breve{X}_{1i}^T\delta + \breve{X}_{2i}^T\delta|_{C_i=\ell}} \tag{3.5}$$

o equivalentemente, en base a la función de supervivencia:

$$S_i(t)|_{C_i=\ell} = (S_0(t)|_{C_i=\ell})^{\breve{X}_{1i}^T\delta + \breve{X}_{2i}^T\delta|_{C_i=\ell}} \tag{3.6}$$

3.4. El modelo lineal mixto conjunto de clases latentes con variable respuesta observable

En base a todo lo anteriormente expuesto, el planteamiento del modelo lineal mixto conjunto de clases latentes con variable respuesta observable, que es el que detallaremos en este trabajo, quedará expresado por los siguientes submodelos para cualquier sujeto i. Inicialmente, todo sujeto i se asume que pertenece a alguna clase ℓ de las L clases latentes, siendo su probabilidad de pertenecia modelada por la regresión logística multinomial (a). Seguidamente, la respuesta de interés del sujeto i en el tiempo t_{ij}, seguirá para la clase de pertenencia del sujeto, el modelo estructural (b) descrito en (3.1). Simultáneamente, la función de riesgo para el sujeto i se describirá de acuerdo al modelo de riesgos proporcionales (c).

(a) $\pi_{i\ell} = \frac{e^{\tilde{X}_i^T \xi_\ell}}{1+e^{\tilde{X}_i^T \xi_1}+e^{\tilde{X}_i^T \xi_2}+...+e^{\tilde{X}_i^T \xi_{(L-1)}}}$

(b) $Y_{ij}|_{C_i=\ell} = X_{1ij}^T \beta + X_{2ij}^T \upsilon|_{C_i=\ell} + Z_{ij}^T \alpha_i|_{C_i=\ell} + B_{t_{ij}} + \epsilon_{ij}$

(c) $\lambda_i(t)|_{C_i=\ell} = \lambda_0(t)|_{C_i=\ell} \quad e^{\breve{X}_{1i}^T \delta + \breve{X}_{2i}^T \delta|_{C_i=\ell}}$

donde $\alpha_i|_{C_i=\ell} \sim N_q(0, w_\ell^2 \mathbb{H})$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ y $B_{t_{ij}}$ es un proceso Gaussiano en el tiempo t_{ij}. Bajo esta formulación, si definimos como en (2.2) $\mathbf{Y_i} = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^T$, la distribución marginal de las respuestas de un sujeto i queda descrita como:

$$\mathbf{Y_i}|_{C_i=\ell} \sim \quad N_{n_i}(\mu_{i\ell}, \mathbb{V}_{i\ell})$$

con

$$\begin{aligned} \mu_{i\ell} &= X_{1ij}^T \beta + X_{2ij}^T \upsilon|_{C_i=\ell} \\ \mathbb{V}_{i\ell} &= w_\ell^2 Z_i \mathbb{H} Z_i^T + Cov(W_i) + \sigma_\epsilon^2 \mathbb{I}_{n_i} \quad , \end{aligned}$$

donde $Cov(W_i)$ denota a la matriz de varianza-covarianza asociada al proceso Gaussiano en los tiempos de observación para el sujeto i.

De otro lado, si $\mathbf{U_i} = (T_i, E_i)$ es el vector aleatorio conformado por:

- $T_i = min\left\{T_i^*, \tilde{T}_i\right\}$: variable observada en el modelo de supervivencia, siendo T_i^* el tiempo transcurrido hasta la observación del evento de interés y $\tilde{T}_i$ el tiempo de censura
- $E_i = 1_{[T_i^* < \tilde{T}_i]}$: variable indicadora que toma el valor de 1, si se observó el evento de interés en el sujeto i, y 0 en caso contrario,

el modelo asume que $\mathbf{Y_i}$ y $\mathbf{U_i}$ son condicionalmente independientes dada la clase latente de pertenencia del sujeto i.

Si denotamos por $\check{X}_i = \left(\tilde{X}_i, X_{1i}, X_{2i}, Z_i, \breve{X}_{1i}, \breve{X}_{2i}\right)$ al conjunto de variables predictoras, donde:

$$
\begin{aligned}
X_{1i} &= \left(X_{1i1}^T, X_{1i2}^T, ..., X_{1in_i}^T\right)^T \\
X_{2i} &= \left(X_{2i1}^T, X_{2i2}^T, ..., X_{2in_i}^T\right)^T \\
\breve{X}_{1i} &= \left(\breve{X}_{1i1}^T, \breve{X}_{1i2}^T, ..., \breve{X}_{1in_i}^T\right)^T \\
\breve{X}_{2i} &= \left(\breve{X}_{2i1}^T, \breve{X}_{2i2}^T, ..., \breve{X}_{2in_i}^T\right)^T
\end{aligned}
$$

se tiene que la distribución conjunta condicional de las respuestas de un sujeto y de los tiempos de ocurrencia del evento y censura viene dada por:

$$
f\left(\mathbf{y}, \mathbf{u} | \check{X}_i = \check{x}, C_i = \ell\right) = f\left(\mathbf{y} | \check{X}_i = \check{x}, C_i = \ell\right) f\left(\mathbf{u} | \check{X}_i = \check{x}, C_i = \ell\right) \quad , \tag{3.7}
$$

donde recordemos que C_i denota a la variable aleatoria indicadora de la clase latente ℓ.

Este supuesto se basa en dos aspectos:

- La evaluación de la trayectoria de $\mathbf{Y_i}$ y el riesgo de $\mathbf{U_i}$ pueden ser diferentes a través de las clases latentes.
- Una vez identificada la clase en el vector $\mathbf{U_i}$, el tiempo de ocurrencia o censura del evento debe ser independiente de $\mathbf{Y_i}$.

3.5. Clasificación a posteriori

Con la finalidad caracterizar la clasificación de los sujetos y analizar el ajuste del modelo se puede también calcular, a través del Teorema de Bayes, la probabilidad a posteriori para cada sujeto i, de pertenecer a la clase ℓ dada la información recolectada.

Para el caso del modelo desarrollado en este trabajo, la probabilidad a posteriori para cada sujeto i de pertenecer a una clase ℓ dada la información recolectada, se puede calcular mediante el Teorema de Bayes y viene dada por:

$$
\begin{aligned}
\hat{\pi}_{i\ell} &= P\left(C_i = \ell | \check{X}_i = \check{x}, Y_i = y, T_i = t, E_i = e\right) \\
&= \frac{f\left(y | \check{X}_i = \check{x}, C_i = \ell\right) f\left(t, e | \check{X}_i = \check{x}, C_i = \ell\right) \; P(C_i = \ell | \tilde{X}_i = \tilde{x})}{\sum_{k=1}^{L} f\left(y | X_i = x, C_i = k\right) f\left(t, e | \check{X}_i = \check{x}, C_i = k\right) \; P(C_i = k | \tilde{X}_i = \tilde{x})}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = \ell)\left(\lambda_0(t)|_{C_i=\ell} e^{X_i^T\delta}\right)^e e^{-\int_0^t \lambda_0(s)|_{C_i=\ell} e^{X_i^T\delta} ds}\ \pi_{i\ell}}{\sum_{k=1}^{L}\ \phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = k)\left(\lambda_0(t)|_{C_i=k} e^{X_i^T\delta}\right)^e e^{-\int_0^t \lambda_0(s)|_{C_i=k} e^{X_i^T\delta} ds}\pi_{ik}} \\
&= \frac{\phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = \ell)\left(\lambda_0(t)|_{C_i=\ell} e^{X_i^T\delta}\right)^e e^{-\Lambda_i(t)|_{C_i=\ell}}\pi_{i\ell}}{\sum_{k=1}^{L}\ \phi_{ik}(y|\check{X}_i = \check{x}, C_i = k)\left(\lambda_0(t)|_{C_i=k} e^{X_i^T\delta}\right)^e \left(e^{-\Lambda_i(t)|_{C_i=\ell}}\right)\pi_{ik}}
\end{aligned}
\tag{3.8}
$$

siendo:

- $\hat{\pi}_{i\ell}$ es la probabilidad a posteriori de que un sujeto i pertenezca a la clase ℓ, dada su información sobre la variable dependiente y tiempos de observación
- $\phi_{i\ell}$ es la función de densidad multivariada del vector de interés Y_i dada la clase latente ℓ
- $\lambda_{i\ell}$ es la función de riesgo de observar el evento o la censura dada la clase latente ℓ
- $\Lambda_{i\ell}$ es la función de riesgo acumulado de observar el evento o la censura dada la clase latente ℓ
- E_i es una variable indicadora que toma el valor de 1 si se observó el evento de interés en el sujeto i, y toma el valor de 0 en caso contrario

La clasificación a posteriori, resulta entonces al comparar las probabilidades de pertenencia a las clases obtenidas para cada sujeto i y seleccionar para este la clase con mayor probabilidad a posteriori.

Capítulo 4

Estimación del modelo

Para la estimación de los parámetros del modelo lineal mixto conjunto de clases latentes utilizaremos el método de máxima verosimilitud. En este trabajo asumiremos un modelo lineal mixto conjunto de clases latentes de la forma dada en la sección 3.4.

Uno de los supuestos cuando se trabaja con el modelo lineal mixto conjunto de clases latentes es que asume independencia entre la variable longitudinal y el proceso de supervivencia, dadas las covariables y las clases latentes, como se mostró en (3.7).

Considerando un número determinado L de clases latentes, el vector de parámetros a estimar estará dado por:

$$\theta = \left(\xi, \beta, \nu_\ell, w_\ell^2, vec\left(\mathbb{H}\right), \sigma_\epsilon^2, \sigma_\omega^2, \rho, \zeta_\ell, \delta, \delta_\ell\right) \tag{4.1}$$

donde,

- ξ es el vector de coeficientes de regresión asociado a las covariables del modelo de pertenencia a las clases.
- β es el vector de coeficientes asociado a las variables de efectos fijos comunes para todas las clases del modelo lineal mixto.
- ν_ℓ es el vector de coeficientes asociado a las variables de efectos fijos específicos por clase ℓ del modelo lineal mixto.
- $vec\left(\mathbb{H}\right)$ es el vector de parámetros que permiten modelar la matriz de varianzas y covarianzas vectorizado de los efectos aleatorios.
- σ_ϵ^2 es la varianza de los errores de medición ϵ_{ij}.
- σ_ω^2 es la varianza del proceso estocástico
- ρ es la correlación entre las observaciones, considerada en el proceso estocástico.
- δ es el vector de coeficientes de regresión comunes asociados a las covariables del modelo de riesgos proporcionales de Cox.
- δ_ℓ es el vector de coeficientes de regresión específicos por clase ℓ, asociados a las covariables del modelo de riesgos proporcionales de Cox.

- ζ_ℓ es el vector de coeficientes de la función de riesgo basal del modelo de riesgos proporcionales de Cox.

La verosimilitud, en base a la data observada del vector aleatorio $(Y_i\ ,\ U_i\ ,\ \check{X}_i\ \ ;\ \ i = 1, 2, ..., n)$, donde:

$$\begin{aligned}\mathbf{Y_i} &= (Y_{i1}, Y_{i2}, ..., Y_{in_i})\\ \mathbf{U_i} &= (T_i, E_i)\\ \check{X}_i &= \left(\tilde{X}_i, X_{1i}, X_{2i}, Z_i, \check{X}_{1i}, \check{X}_{2i}\right)\end{aligned}$$

puede escribirse, en base a (3.7) y (2.16), como:

$$\begin{aligned}&L(\theta|Y_i = y, \check{X}_i = \check{x}, C_i = \ell, T_i = t, E_i = e)\\ &\quad= \prod_{i=1}^{n}\left(\sum_{\ell=1}^{L} P(C_i = \ell|\tilde{X}_i = \tilde{x}) f\left(y|X_i = x, C_i = \ell\right) f\left(t, e|\check{X}_i = \check{x}, C_i = \ell\right)\right)\\ &\quad= \prod_{i=1}^{n}\left(\sum_{\ell=1}^{L} \pi_{i\ell}\ \phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = \ell)\left(\lambda_0(t)|_{C_i=\ell} e^{X_i^T\delta}\right)^e e^{-\int_0^{t_i}\lambda_0(s)|_{C_i=\ell} e^{X_i^T\delta} ds}\right)\\ &\quad= \prod_{i=1}^{n}\left(\sum_{\ell=1}^{L} \pi_{i\ell}\ \phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = \ell)\left(\lambda_0(t)|_{C_i=\ell} e^{X_i^T\delta}\right)^e e^{-\Lambda_i(t)|_{C_i=\ell}}\right)\end{aligned} \tag{4.2}$$

donde $\lambda_i(t)|_{C_i=\ell}$ es la función de riesgo, $\phi_{i\ell}$ es la función de densidad normal multivariada de la variable de interés $\mathbf{Y}_i$, dado el vector de parámetros θ y la clase latente ℓ, es decir $\mathbf{Y_i} \sim N_{n_i}\left(X_{1i}^T\beta + X_{2i}^T\nu_\ell\ ,\ Z_i w_\ell^2 \mathbb{H} Z_i^T + Cov(W_i) + \sigma_\epsilon^2 \mathbb{I}_n\right)$, y donde $e^{-\Lambda_i(t)|_{C_i=\ell}}$ es la función de supervivencia.

La log-verosimilitud tiene entonces la forma:

$$\begin{aligned}&log\, L(\theta|Y_i = y, \check{X}_i = \check{x}, C_i = \ell, T_i = t, E_i = e)\\ &\quad= \sum_{i=1}^{n} log\left(\sum_{\ell=1}^{L} P(C_i = \ell|\tilde{X}_i = \tilde{x}) f\left(y|\check{X}_i = \check{x}, C_i = \ell\right) f\left(t, e|\check{X}_i = \check{x}, C_i = \ell\right)\right)\\ &\quad= \sum_{i=1}^{n} log\left(\sum_{\ell=1}^{L} \pi_{i\ell}\ \phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = \ell)\left(\lambda_0(t)|_{C_i=\ell} e^{X_i^T\delta}\right)^e e^{-\int_0^t \lambda_0(s)|_{C_i=\ell} e^{X_i^T\delta} ds}\right)\\ &\quad= \sum_{i=1}^{n} log\left(\sum_{\ell=1}^{L} \pi_{i\ell}\ \phi_{i\ell}(y|\check{X}_i = \check{x}, C_i = \ell)\left(\lambda_0(t)|_{C_i=\ell} e^{X_i^T\delta}\right)^e e^{-\Lambda_i(t)|_{C_i=\ell}}\right)\end{aligned} \tag{4.3}$$

En la expresión (4.3), tal como se indicó en la sección 3.3, la función de riesgo basal debe ser expresada en base al vector de parámetros ζ_ℓ.

4.1. Algoritmo iterativo de Marquardt

Debido a la complejidad de la log-verosimilitud de nuestro modelo, no es posible encontrar una solución analítica que permita maximizarlo. Por ello, es necesario emplear un procedi-

miento de optimización numérica para encontrar $\hat{\theta}$.

El paquete *lcmm* trabaja con un algoritmo numérico de Newton Raphson empleando un algoritmo iterativo de Marquardtque que la autora muestra en trabajos previos que tiene un buen desempeño (Proust-Lima et al. (2015)).

Mediante este algoritmo, el vector de parámetros se va actualizando iterativamente hasta que llegue a converger según los criterios abajo indicados. La ecuación que se emplea para ello tiene la forma:

$$\theta^{k+1} = \theta^k - \delta\left(\tilde{H}^k\right)^{(-1)} \nabla\left(L\left(\theta^k\right)\right) \tag{4.4}$$

donde δ asegura que la log-verosimilitud vaya incrementándose de iteración en iteración, $\nabla\left(L\left(\theta^k\right)\right)$ es el gradiente de de la log-verosimilitud en la iteración anterior, y $\tilde{H}$ es la matriz Hessiana con la diagonal inflada para asegurar que sea definida positiva, siendo esta:

$$\tilde{H}_{ii} = H_{ii} + \lambda\left[(1-\eta) + \eta\left(tr\left(H\right)\right)\right] \tag{4.5}$$

Aquí H es la matriz Hessiana, es decir:

$$H = \frac{\partial^2 log\left(L\left(\theta\right)\right)}{\gamma\theta^2} \tag{4.6}$$

y H_{ii} son los términos de su diagonal, mientras que λ y η pueden ser fijados como 0.01. Estos parámetros podrían reducirse aún más si $\tilde{H}$ es definida positiva, de lo contrario se incrementan.

El algoritmo de Marquardt considera tres criterios de convergencia:

- En base a la estabilidad del parámetro

$$\sum_{j=1}^{n_\theta}\left({\theta_j}^k - {\theta_j}^{k-1}\right)^2 \leq \varepsilon_a \tag{4.7}$$

- En base a la estabilidad de la log-verosimilitud

$$\left|L^{(k)} - L^{(k-1)}\right| \leq \varepsilon_b \tag{4.8}$$

- En base al tamaño de la las derivadas

$$\frac{\left[\nabla\left(L\left(\theta^k\right)\right)\right]^T\left(H^k\right)^{(-1)}\left[\nabla\left(L\left(\theta^k\right)\right)\right]}{n_\theta} \leq \varepsilon_c \tag{4.9}$$

donde n_θ es la longitud del vector de parámetros. En estos criterios, los valores de $\varepsilon_a, \varepsilon_b, \varepsilon_c$ se fijan por defecto en $\varepsilon_a = \varepsilon_b = \varepsilon_c = 10^{-4}$.

Una estimación de la matriz de varianza-covarianza para las estimaciones por máxima verosimilitud ($\widehat{V(\hat{\theta})}$) está dada por la inversa de la matriz Hessiana evaluada en el estimador de máxima verosimilitud.

El proceso de estimación mencionado anteriormente es realizado por el paquete *lcmm* desarrollado por Proust-Lima et al. (2015). De este modo, no es necesario indicar previamente la forma de la log-verosimilitud que se desea maximizar, ya que el paquete se encarga de construirla en base a los componentes de los submodelos que conforman el modelo lineal mixto conjunto de clases latentes que indicamos mediante la función *JointLCMM*.

Capítulo 5

Estudio de Simulación

La finalidad de realizar este estudio de simulación fue mostrar el desempeño de la función *Jointlcmm* del paquete *lcmm*, mediante la recuperación de los parámetros estimados al construir el modelo lineal mixto conjunto de clases latentes.

5.1. Descripción

Se consideró la base de datos *paquid* incluida en el paquete *lcmm* (Proust-Lima et al., 2017), la cual contiene información sobre una investigación en Francia de pacientes observados durante 20 años aproximadamente y a quienes se les aplicó un test psicológico cada 2 o 3 años registrando también en cada visita si fueron diagnosticados con demencia durante este periodo y la fecha respectiva del diagnóstico. Se realizaron 5 000 simulaciones trabajadas en paralelo con la ayuda del *Proyecto Legión* de la Pontificia Universidad Católica del Perú.

La variable dependiente fue $MMSE$, una variable que indica el puntaje que obtuvo el paciente i en el periodo de observación j. La variable que se empleó en el modelo fue $normMMSE$ la cual es la variable resultante de la estandarización de $MMSE$.

Las covariables consideradas fueron:

- CEP: una variable dicotómica que indica si el paciente recibió educación. Toma el valor de 1 si recibió educación y 0 en caso contrario.
- AGE: una variable que indica la edad del paciente al momento de la visita.
- $AGE65$: una variable resultante de la transformación de AGE, obtenida al restarle 65 y dividirla entre 10.
- $SEXO$: una variable que indica el género del paciente. Toma el valor de 1 si es hombre y 0 si es mujer.

Para el ejercicio de simulación se generó un nuevo conjunto de datos con en base a las covariables de *paquid*. Se establecieron 2 clases latentes. La probabilidad real de pertenencia a las clases para cada paciente fue considerada en base a las probabilidades a posteriori obtenidas al estimar los valores de los parámetros que serían fijados, la distribución porcentual de los pacientes fue del 22 %y 77 % para cada clase latente. Esta distribución porcentual también

se espera recuperar con el presente estudio de simulación.

Las estimaciones obtenidas de los parámetros del modelo construido fueron consideradas como los valores reales de los parámetros necesarios para generar los valores simulados de la variable dependiente. Estos valores reales de los parámetros son los que se buscó recuperar.

5.2. Resultados

El modelo lineal mixto conjunto de clases latentes, del cual se simulan los datos, estará constituido por los siguientes submodelos:

- Probabilidad de pertenencia:

$$\pi_{i\ell} = \frac{e^{\xi_{0\ell}}}{1 + e^{\xi_{0\ell}}} \quad \ell = 1, 2. \tag{5.1}$$

- Modelo lineal mixto por clase:

$$Y_{ij}|_{C_i=\ell} = \beta_1 CEP + \upsilon_{0\ell} + \upsilon_{1\ell}(AGE65) + \alpha_{0i} + \alpha_{1i}(AGE65) + \epsilon_{ij} \tag{5.2}$$

- Modelo de riesgo proporcional por clase:

$$\lambda_i(t)|_{C_i=\ell} = \left(Weibull\left(t; \zeta_1; \zeta_2\right)\right) e^{\delta_1 CEP} \tag{5.3}$$

donde t es el tiempo hasta la observación del evento o de la censura (T_i), y $\alpha_i = (\alpha_{0i}, \alpha_{1i})$, siendo:

$$\alpha_i \sim N_q\left(0\ ,\ H = \begin{pmatrix} s_0 & s_{01} \\ s_{01} & s_1 \end{pmatrix}\right)$$
$$\epsilon_{ij} \sim N(0\ ,\ \sigma_\epsilon^2)$$

Para cada una de las cuatro clase latentes consideradas se simuló una media $\mu_{i\ell} = \beta_1 CEP + \upsilon_{0\ell} + \upsilon_{1\ell}(AGE65) + \alpha_{0i} + \alpha_{1i}(AGE65)$ según los valores de las covariables y los valores de los parámetros mencionados anteriormente. Con estas medias obtenidas, y luego de calcular la matriz de varianzas y covarianzas $\mathbb{V}_{i\ell} = Z_i \mathbb{H} Z_i^T + \sigma_\epsilon^2 \mathbb{I}_n$ también específica para cada clase, se pudo simular los valores de $Y_{ij}|_{C_i=\ell}$.

Luego de obtener 500 valores simulados de los $Y_{ij}|_{C_i=\ell}$ para las distintas clases, se estimó el modelo lineal mixto conjunto de clases latentes mediante la función *Jointlcmm*, guardando las salidas del vector de la distribución porcentual de pacientes en cada clase latente y del vector de parámetros estimados por el modelo. Esto se realizó para cada una de las 500 simulaciones.

Al calcular los valores promedio de la distribución de los pacientes en las 2 clases latentes, obtenidos en cada simulación realizada, se obtuvieron los resultados de la Tabla 5.1. De esta

manera, fue posible comparar la verdadera distribución de los pacientes en cada clases con la distribución promedio de pacientes obtenida mediante la simulación.

Tabla 5.1: Porcentaje de pacientes por clase latente, según el conjunto de datos *paquid* y la simulación

	Datos	**Simulación**
% de pacientes en Clase 1	12.625	12.629
% de pacientes en Clase 2	87.375	87.371

Se pudo apreciar que el modelo *Jointlcmm*, divide a la data en proporciones similares a las esperadas, pero realizando un *Label Switching* (cambio de etiqueta) en algunas estimaciones; esto es, que se obtienen los mismos porcentajes pero las clases no tienen el mismo orden que las reales.

Para poder controlar este cambio de etiquetas de las clases y poder comparar las estimaciones con los valores reales de la clase correspondiente, dentro del bucle del algoritmo para la simulación se desarrollaron unas líneas de código que permitieron reordenar las clases de acuerdo al orden de referencia. Este reordenamiento se realizó para cada vector de estimaciones obtenidos en cada una de las simulaciones, guardando los valores en una matriz que etiquetaba y ordenaba las mismas según el orden de referencia.

El tiempo total aproximado de simulación fue de 6 horas, realizándose la simulación en paralelo. El desempeño del modelo estudiado se pudo evaluar calculando y analizando la media, el sesgo, el sesgo porcentual, la desviación estándar y el error cuadrático medio de cada uno de los parámetros. En la Tabla 5.2, se muestran los indicadores del desempeño del modelo *Jointlcmm*.

Se pudo apreciar que la mayoría de los valores medios de las estimaciones obtenidas fueron muy similares a los valores reales de los parámetros, esto se evidencia en el valor del sesgo porcentual y media de cada uno de ellos. Se observó que las estimaciones de los coeficientes del modelo lineal mixto fueron las que presentaron los menores sesgos porcentuales.

Fue posible notar también que en este tipo de modelos resulta de gran impacto la elección de los valores iniciales para la estimación, es por eso que sugerimos realizar la estimación del modelo en 2 etapas. En la primera etapa calcular los valores iniciales mediante la función *gridsearch* (como se muestra en los códigos adjuntos en Anexos), y en la segunda etapa proceder a estimar el modelo empleando los valores iniciales encontrados en la primera etapa.

Tabla 5.2: Indicadores del desempeño del modelo *Jointlcmm-* (500 simulaciones con Legion)

Parámetros	**Valores Reales**	**Media**	**Desviación Estándar**	**Sesgo**	**Sesgo Porcentual**	**Error cuadrático medio**
ξ_{02}	**1.9055**	1.935	0.157	0.029	1.522	1.306
ζ_{11}	**0.3437**	0.345	0.155	0.001	0.377	0.011
ζ_{21}	**1.6417**	1.657	0.011	0.015	0.931	0.09
ζ_{12}	**0.1834**	0.178	0.081	-0.005	2.945	0.005
ζ_{22}	**4.0921**	7.031	0.001	2.939	71.828	2.957
δ_1	**0.2055**	0.174	0.207	-0.032	15.514	0.14
υ_{01}	**89.8601**	90.295	3.206	0.435	0.484	2.998
υ_{02}	**75.6098**	75.887	1.493	0.277	0.366	1.531
υ_{11}	**-27.3375**	-27.434	1.690	-0.096	0.353	1.763
υ_{12}	**-10.2766**	-10.309	0.700	-0.033	0.316	0.685
β_1	**13.0781**	12.917	1.273	-0.161	1.233	1.161
s_0	**229.7198**	229.412	1.003	-0.308	0.134	25.797
s_{01}	**-97.9317**	-97.319	0.801	0.612	0.625	14.955
s_1	**78.9266**	77.793	0.298	-1.133	1.436	9.75
σ_ϵ	**10.529**	10.551	0.192	0.022	0.206	0.188

Capítulo 6

Aplicación

En este capítulo se ilustrará el modelo lineal mixto conjunto de clases latentes con un conjunto de datos de pacientes diagnosticados con cáncer de Próstata. Se analiza de manera conjunta un proceso longitudinal y un proceso de supervivencia. El proceso longitudinal considera como variable respuesta el valor de PSA (Antígeno Prostático Específico, o por sus siglas en inglés "Prostate-Specific Antigen") en la sangre el cual es una proteína elaborada por la próstata cuya concentración en la sangre puede ser más elevada de lo normal en hombres con diagnóstico de cáncer de próstata, hiperplasia prostática benigna (HPB), o inflamación de la próstata. El proceso de supervivencia considera como evento de interés la existencia o ausencia de una falla bioquímica (FBQ), definiendo como falla bioquímica cuando el valor del PSA en el paciente ha superado los 0.20 nanogramos por cada mililitro de sangre (ng/ml).

6.1. Descripción de los datos

Se trabajó con una base de datos de pacientes diagnosticados con cáncer de próstata, de la Clínica Oncosalud - AUNA, información facilitada por la Dirección Científica y Académica de Oncosalud. Debido a que toda la información se encontraba registrada en historias clínicas digitales y no se contaba con el registro de las variables en una base de datos digital, la data fue recolectada y estructurada en colaboración con la autora de la presente tesis.

La base de datos estuvo conformada por información de una muestra aleatoria de 95 pacientes diagnosticados con cáncer de próstata, los cuales fueron seleccionados mediante un muestreo aleatorio simple de un listado proporcionado por la clínica con pacientes cuya fecha de diagnóstico de cáncer va desde el año 2010 al 2012. El tiempo de observación fue de Enero del 2010 a Noviembre del 2017, y uno de los principales criterios de inclusión fue que todos los pacientes recibieron por tratamiento una intervención quirúrgica denominada *prostatectomía radical* (PR). Otro criterio de inclusión considerado fue que la primera medición de PSA de los pacientes debía ser menor a 1 ng/ml.

Con la información de la base de datos, se construyeron variables, como por ejemplo el primer valor de PSA medido luego de la prostatectomía radical, el tiempo desde la PR hasta cada toma de PSA y el tiempo desde la PR hasta la presencia de una FBQ o de una censura. Debido a que el interés se centra en estudiar la trayectoria del PSA hasta la ocurrencia de una FBQ o de una censura, la información respecto a cada paciente se analizó hasta la observación

de alguno de estos eventos mencionados.

Luego de realizar la limpieza de la data y crear las variables necesarias en base a la información recopilada, se pudo obtener la Figura 6.1, en la cual se aprecian las trayectorias de PSA por paciente hasta la observación de una falla bioquimica o una censura.

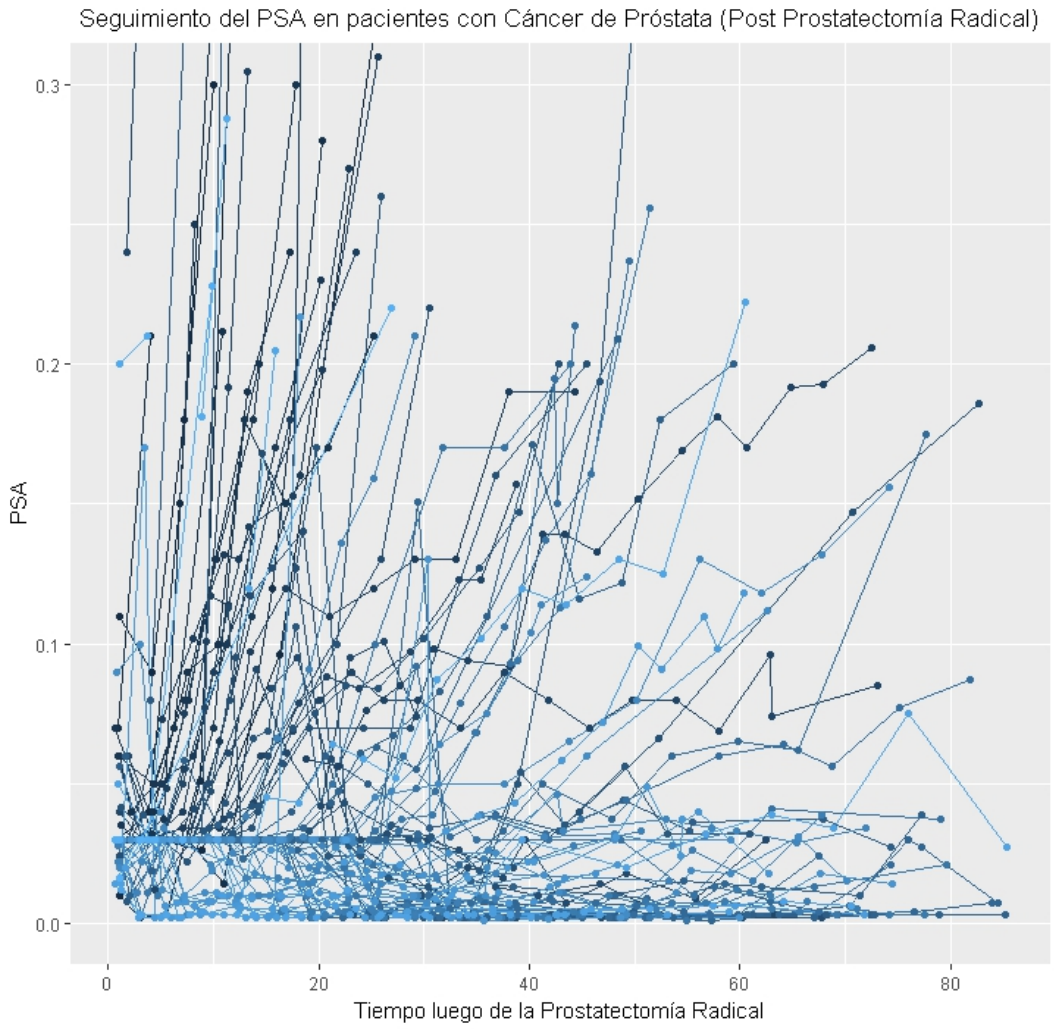

Figura 6.1: Trayectorias de PSA hasta falla bioquímica o censura

A continuación, se describen las variables en estudio, tanto variables originales como variables calculadas:

- *PSAD*: valor del PSA al momento del diagnóstico de cáncer.
- *EC*: es una clasificación subjetiva propia de la opinión del médico tratante que se le da al paciente según la gravedad de la enfermedad, la cual indica el estadío clínico del paciente al momento del diagnóstico de cáncer. Toma valores discretos del 1 al 4, donde 1 representa el primer estadío del cáncer y 4 el estadío de cáncer más avanzado.
- *Gleason*: valor estandarizado del puntaje de Gleason al momento del diagnóstico. El puntaje de Gleason toma valores entre 2 y 10 e indica la probabilidad de diseminación

o expansión del tumor.

- *Edad*: años de vida del paciente al momento de la prostatectomía radical.
- PSA: valor del PSA medido en cada paciente cada 1, 2 o 3 meses consecutivamente luego de la fecha de la prostatectomía radical. El tiempo entre cada toma es diferente (irregular) para cada paciente.
- $PSABase$: valor del primer PSA medido después de la prostatectomía radical.
- $TimePR$: tiempo desde la prostatectomía radical hasta el momento de medición.
- FBQ: presencia o ausencia de una falla bioquímica (evento de interés).
- $TimeFBQ$: tiempo desde la prostatectomía radical hasta la ocurrencia del evento o presencia de una censura.

Al analizar la clasificación subjetiva propia de la opinión del médico tratante (denominada estadío clínico), se pudo observar que del total de pacientes que conforman la muestra en estudio, el 4 % fue diagnosticado en estadío clínico I, el 58 % en estadío clínico II, el 25 % en estadío clínico III y el 13 % en estadío clínico IV. Esta información se comparará posteriormente con la información obtenida de la clasificación aposteriori de los pacientes mediante el modelo en estudio, en las que no se toma en cuenta este diagnóstico.

6.2. Resultados

Las variables de la base de datos con las que se construyó el modelo lineal mixto conjunto de clases latentes se distribuyeron como se muestra a continuación:

- Modelo de pertenencia a la clase:

  Para el modelo de pertenecia a la clase no se consideraron covariables, solo se consideraron interceptos. Aquí, L representa la cantidad total de clases latentes identificadas y $\pi_{i\ell}$ representa la probabilidad de que el paciente i pertenezca a la clase latente ℓ, donde $\ell = 1, 2, ..., L-1$.

$$\pi_{i\ell} = \frac{e^{\xi_{0\ell}}}{1 + e^{\xi_{01}} + e^{\xi_{02}} + ... + e^{\xi_{0(L-1)}}} \tag{6.1}$$

- Modelo lineal mixto:

  En el modelo lineal mixto se consideró a las mediciones del PSA en el tiempo como la variable respuesta (PSA), se incluyeron tres variables asociadas a los efectos fijos generales (PSA al momento del diagnóstico, primer PSA medido luego de la prostatectomía radical y edad del paciente al momento de la prostatectomía radical), una variable de efectos fijos por clase latente (tiempo desde la prostatectomía radical),Un variable de efectos aleatorios con intercepto y pendiente aleatoria asociada a la variable tiempo

desde la prostatectomía radical ($TimePR$) por individuo.

$$PSA_{ij}|_{C_i=\ell} = \beta_1(PSAD)_i + \beta_2(PSABase)_i + \beta_3(Edad)_i + \upsilon_{1\ell}(TimePR)_{ij} + \alpha_{0i} + \alpha_{1i}(TimePR)_{ij} + \epsilon_{ij} \tag{6.2}$$

donde, $i = 1, 2, ..., 95$, $j = 1, 2, ..., n_i$, $\alpha_i \sim N_q(0, \mathbb{H})$, $\mathbb{H}$ es una matriz de varianza-covarianza de efectos aleatorios, y con un vector aleatorio de errores de medición $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$.

- Modelo de riesgo proporcional:

  En el modelo de supervivencia se consideró como variable respuesta al tempo transcurrido desde la prostatectomía radical hasta la ocurrencia de una falla bioquímica o la presencia de una censura ($TimeFBQ$). Se asumió que la función de riesgo basal tenía una distribución Weibull con parámetros ζ_1 y ζ_2. Así también, se consideraron dos variables de efectos fijos en el modelo de regresión (PSA al momento del diagnóstico y valor estandarizado del puntaje de Gleason).

$$\lambda_i(TimeFBQ|_{C_i=\ell}) = (Weibull\,((TimeFBQ); \zeta_1; \zeta_2))\, e^{\delta_1(PSAD)+\delta_2(Gleason)} \tag{6.3}$$

El modelo se construyó en dos etapas. En la primera etapa, se estimaron los valores iniciales con los cuales se construiría el modelo, Proust-Lima et al. (2015) recomienda que en este tipo de modelos es importante ser lo más preciso posible al considerar los valores iniciales. En la segunda etapa, se construyó el modelo lineal mixto conjunto de clases latentes considerando los valores iniciales estimados en la primera etapa.

Por literatura respecto al cáncer de próstata, se tiene conocimiento de la existencia de 4 estadíos clínicos, por lo que se esperaría que existieran 4 clases o subpoblaciones en la data estudiada. Con la finalidad de evaluar si existen menos o más subpoblaciones en la data, se construyeron modelos para 1,2,3,4, y 5 clases latentes. Los valores de la log-verosimilitud, cantidad de parámetros y BIC para cada modelos se aprecia en la Tabla 6.1.

Tabla 6.1: Comparación y selección del mejor modelo *Jointlcmm*

Modelo	**loglik**	**Cantidad de Parámetros**	**BIC**	**Porcentaje de pacientes por clase**				
				Clase1	**Clase2**	**Clase3**	**Clase4**	**Clase5**
Modelo 1	1515.99	12	-2977.33	100.00				
Modelo 2	1520.13	16	-2967.40	89.47	10.53			
Modelo 3	1596.18	20	-3101.29	16.84	72.63	10.52		
Modelo 4	1630.85	24	-3152.41	15.79	20.00	52.68	10.53	
Modelo 5	1630.85	28	-3134.20	15.79	52.63	10.53	21.05	0.00

Al comparar los modelos se pudo identificar que el menor valor de BIC (-3152.41) y el mayor valor de log-verosimilitud (1630.85) se obtuvieron al considerar la existencia de 4 clases

latentes en la data. Por ello, se seleccionó al modelo de 4 clases como el mejor modelo.

Se observó que las etiquetas asignadas por el modelo para cada clase latente no suelen presentarse ordenadas. Con la ayuda de especialistas en el tema de cáncer de próstata, se pudo renombrar las etiquetas de cada clase en base a las características conocidas sobre la evolución de los pacientes pertenecientes a cada estadío clínico, tal como se muestra en la Tabla 6.2

Tabla 6.2: Clases renombradas del modelo *Jointlcmm*

Etiqueta dada por el modelo	**Etiquetas renombradas**	**Porcentaje de pacientes**
class3	Clase 1	53.68 %
class2	Clase 2	20.00 %
class1	Clase 3	15.79 %
class4	Clase 4	10.53 %

El modelo seleccionado de 4 clases latentes *Modelo* 4, estimó 24 parámetros. Los resultados para el modelo de pertenencia a la clase, modelo lineal mixto y modelo de riesgo proporcional se aprecian en la Tabla 6.3.

Del modelo estimado, se puede resaltar que la variable más influyente sobre los valores del PSA en los pacientes diagnosticados con cáncer de próstata es el valor de la primera toma de PSA de un paciente i luego de la prostatectomía radical ("$PSABase$", con parámetro β_2). Es decir, si la disminución del PSA al primer mes fue significativa respecto al PSA de diagnóstico ("$PSAD$"), el valor de los PSA del paciente (Y_{ij}) en el tiempo también disminuirá en promedio significativamente.

Así también, fue posible apreciar que en las clases latentes 3 y 4 el impacto de la variable "$TimePR$" (con parámetro $\nu_{1\ell}$) es mayor sobre el valor del PSA; mientras que en las clases latentes 1 y 2, esta influencia es menor. El valor del PSA aumenta más rápido en la clase latente 3 y 4 que en la clase latente 1 y 2.

En la Figura 6.2 es posible apreciar los valores medios predichos y observados de la variable dependiente PSA, sin considerar clases latentes (izquierda) y considerando 4 clases latentes (derecha). Se observó que al considerar clases latentes, fue posible identificar las diferentes trayectorias que sigue el valor del PSA en el tiempo para cada paciente, la estimación mediante el modelo lineal mixto conjunto de cuatro clases latentes nos permitió realizar una predicción más fina de las trayectorias de PSA, según las características de los pacientes pertenecientes a cada clase.

Al comparar las curvas de supervivencia libre de falla bioquímica (Figura 6.3), sin considerar clases latentes (izquierda) y considerando 4 cada clase (derecha), se pudo apreciar que a el modelo de 4 clases latentes nos permite identificar que en la muestra analizada exis-

Tabla 6.3: Estimadores de máxima verosimilitud del modelo *Jointlcmm*

MODELO DE PERTENENCIA A LA CLASE

	Coeficiente	Desviación Estándar	Estadístico de Wald	p-value
ξ_{01}	1.586	0.351	4.522	0.00001
ξ_{02}	0.664	0.397	1.675	0.09401
ξ_{03}	0.389	0.426	0.914	0.36085

MODELO DE REGRESIÓN LINEAL MIXTO
(variables asociadas a efectos fijos)

	Coeficiente	Desviación Estándar	Estadístico de Wald	p-value
β_1	-0.00160	0.00044	-3.655	0.00026
β_2	1.36922	0.04908	27.900	0.00000
β_3	-0.00031	0.00008	-3.827	0.00013
ν_{11}	0.00000	0.00000	1.112	0.26594
ν_{12}	0.00011	0.00001	15.525	0.00000
ν_{13}	0.00037	0.00002	20.047	0.00000
ν_{14}	0.00099	0.00005	20.662	0.00000

MODELO DE REGRESIÓN LINEAL MIXTO
(variables asociadas a efectos aleatorios)

	α_{0i}	α_{1i}
α_{0i}	$6{,}69 \times 10^{-4}$	$-1{,}69 \times 10^{-7}$
α_{1i}	$-1{,}69 \times 10^{-7}$	$3{,}83 \times 10^{-10}$

MODELO DE REGRESIÓN LINEAL MIXTO
(error estándar residual)

	Coeficiente	Desviación Estándar
σ_ϵ	0.03661	0.00084

MODELO DE RIESGO PROPORCIONAL

	Coeficiente	Desviación Estándar	Estadístico de Wald	p-value
ζ_{11}	0.00000	0.02828	0.000	1.00000
ζ_{21}	1.09212	165.79562	0.007	0.99474
ζ_{12}	0.02533	0.00127	19.990	0.00000
ζ_{22}	1.70481	0.22155	7.695	0.00000
ζ_{13}	0.03951	0.00215	18.408	0.00000
ζ_{23}	1.70028	0.19125	8.890	0.00000
ζ_{14}	0.05614	0.00603	9.310	0.00000
ζ_{24}	1.55753	0.17682	8.809	0.00000
δ_1	0.00053	0.01549	0.034	0.97256
δ_2	0.42935	0.23112	1.858	0.06322

ten cuatro curvas de supervivencia diferente. Tratándose del las mediciones en el tiempo del PSA en pacientes diagnosticados con cáncer de próstata, consideramos que el *Modelo* 4 tiene un mejor ajuste ya que, al identificar que cada clase tiene tiempos diferentes de presentar una falla bioquímica, contribuiría a poder brindarle un tratamiento diferente a cada grupo de pacientes, según las características propias de los pacientes pertenecientes a cada clase latente.

Respecto a las curvas de supervivencia para cada grupo de pacientes clasificados según el modelo de cuatro clases latentes. Las clases latentes 3 y 4, presentan las curvas de supervivencia con mayor decaimiento, indicando que durante el primer y segundo año respectivamente de haber sido intervenidos mediante una prostatectomía radical todos los pacientes de estas subpoblaciones latentes presentaron una falla bioquímica; los pacientes de la clase latente 4 tienen una probabilidad del 0.4 de no presentar una falla bioquímica durante los 2 primeros años posteriores a la intervención quirúrgica, y los pacientes pertenecientes a la clase latente 1 no experimentan el evento de interés.

Al comparar el estadío clínico asignado por los médicos tratantes con las clases latentes identificadas con el modelo lineal mixto conjunto de clases latentes (Tabla 6.4) se observó que gran parte de los pacientes diagnosticados con estadío clínico 2 presentaron una trayectoria de PSA mucho más similar a la de pacientes de la clase latente 1 (37 pacientes, 38.95 %) y algunos presentaron una trayectoria similar a la de pacientes de la clase latente 3; de igual manera se pudo apreciar que algunos pacientes diagnosticados en estadío clínico 3 presentaron una trayectoria de PSA mucho más similar a la de pacientes de la clase latente 2 (7 pacientes, 7.37 %) o clase latente 1 (10 pacientes, 10.53 %). Se observó también que determinados pacientes diagnosticados en estadío clínico 4 presentaron una trayectoria de PSA con mayor pendiente respecto a los pacientes de esta clase, por lo que fueron considerados por el modelo dentro de la clase latente 3, y algunos pacientes diagnosticados en estadío clínico 2 y 3, presentaron características de la clase latente 4 (2 y 1 paciente respectivamente, 2.11 % y 1.05 %).

Tabla 6.4: Pacientes según estadío clínico diagnosticado vs clases latentes

Diagnóstico \Latente	Clase 1	Clase 2	Clase 3	Clase 4
Estadío clínico 1	**4**	0	0	0
Estadío clínico 2	37	**10**	6	2
Estadío clínico 3	10	7	**6**	1
Estadío clínico 4	0	2	3	**7**

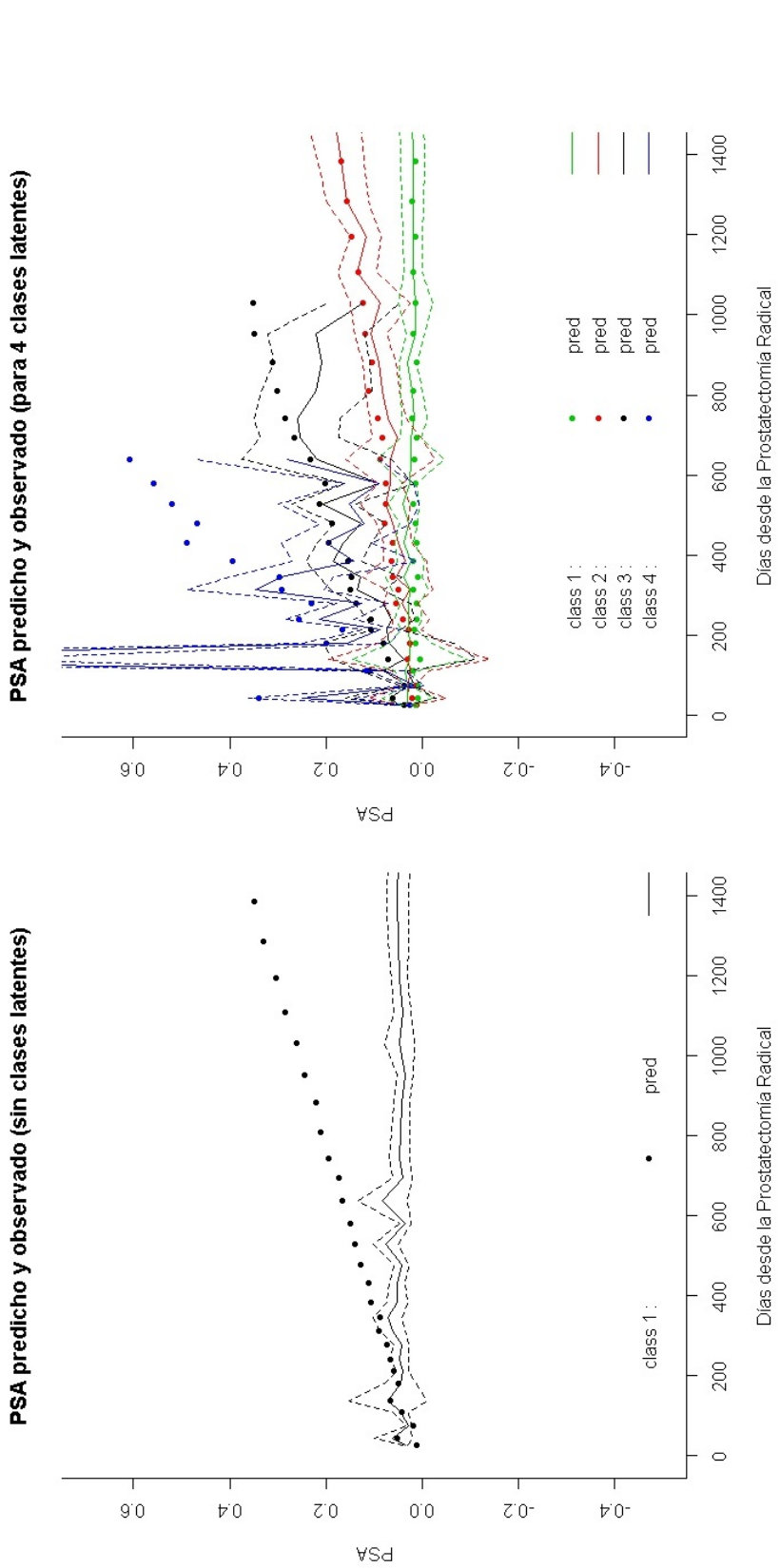

Figura 6.2: Comparación de las curvas de supervivencia libre de falla bioquímica sin considerar clases latentes y considerando cuatro clases latentes.

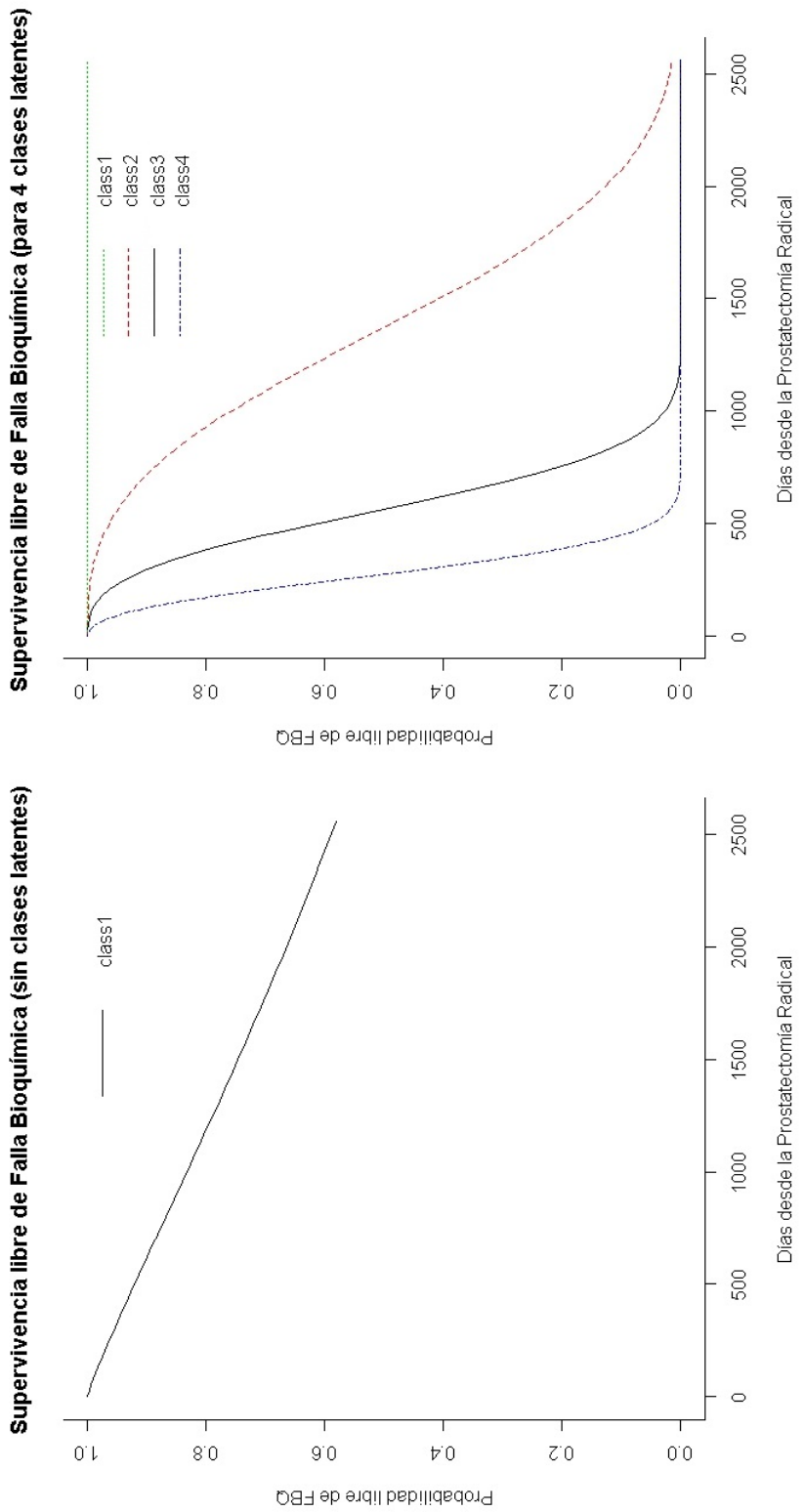

Figura 6.3: Comparación de las trayectorias predichas y observadas sin considerar clases latentes y considerando cuatro clases latentes.

Capítulo 7

Conclusiones y Sugerencias

7.1. Conclusiones

- El modelo lineal mixto conjunto de clases latentes desarrollado por Cécile Proust-Lima, Viviane Philipps y Benoit Liquet en Proust-Lima et al. (2015) es ideal para modelar simultáneamente un proceso longitudinal y un proceso de supervivencia, para cada una de las clases latentes que puedan existir en una determinada población en estudio. En el presente trabajo se logró detallar la forma en la que se realiza la estimación de parámetros de los submodelos para cada clase latente, describiendo la construcción de una verosimilitud conjunta para ambos procesos y para la probabilidad de pertenencia a cada clase.

- Resulta interesante resaltar, que el modelo en estudio considera tiempos irregulares para la diferencia de tiempo entre cada una de las consecutivas mediciones de la variable dependiente de cada uno de los individuos en estudio. Esta característica resulta ser muy conveniente, considerando que en la realidad, las mediciones entre uno y otro individuo no suelen darse necesariamente en los mismos intervalos de tiempo.

- Mediante el desarrollo del estudio de simulación, se pudo evaluar el desempeño del modelo lineal mixto conjunto de clases latentes al intentar recuperar parámetros. Se calculó la media, sesgo, desviación estándar y error cuadrático medio para cada uno de los estimadores. Fue posible apreciar que algunos valores estimados resultaron tener un sesgo porcentual mínimo pero otros, presentaron sesgo porcentual muy alto, como los componentes de la matriz de varianzas y covarianzas (s_0, s_{01} y s_1).

- En este tipo de modelos es muy importante la elección de los valores iniciales para la estimación. Sugerimos realizar la estimación en dos etapas. En la etapa 1 calcular los valores iniciales mediante la función *gridsearch* (que permite estimar n-veces el modelo y selecciona las estimaciones de mayor verosimilitud), y en la etapa 2 estimar el modelo con los valores iniciales encontrados anteriormente.

- La aplicación del modelo lineal mixto conjunto de clases latentes a una base de datos de pacientes diagnosticados con cáncer de próstata permitió apreciar la utilidad y las bondades de este modelo. Debido a la naturaleza de la variable dependiente (PSA), se esperaba encontrar cuatro subpoblaciones en los datos. El modelo permitió identificar

estas cuatro clases latentes, sin embargo se pudo apreciar que algunos pacientes diagnosticados en, por ejemplo, estadío clínico 2, realmente presentaron una trayectoria de sus valores de PSA similares a la de pacientes de la clase latente 1, de igual manera estos pacientes presentaron una curva de supervivencia que se mantenía constante, con una probabilidad muy mínima de presentar una falla biouímica, característica propia también de los pacientes agrupados en la clase latente 1.

- Si bien el modelo en estudio permite identificar clases latentes en la población, las salidas del modelo no etiquetan a estas clases en algún orden determinado. Es necesario entonces renombrar o volver a etiquetar las clases según la expertise del tema. En este caso, se renombró a las clases según la trayectoria del PSA, ya que se conocía de antemano que el aumento en la pendiente de las trayectorias del PSA era directamente proporcional al estadío clínico del paciente en observación.

- Al comparar las estimaciones del modelo sin considerar clases latentes (*Modelo* 1), con el modelo que considera 4 clases latentes (*Modelo* 4), se pudo verificar que el modelo lineal mixto de clases latentes permite calcular estimaciones más finas, haciendo posible identificar las taryectorias de PSA y curvas de supervivencia de los pacientes de cada una de clase latente en la muestra en estudio.

7.2. Sugerencias para investigaciones futuras

- Durante el desarrollo teórico del presente trabajo, se hace referencia a un componente estocástico de error, sin embargo, debido a las características de la base de datos analizada,que cuenta con no muchas observaciones por paciente y no muchos pacientes, no fue necesario el incluirlo en el modelo. Sería interesante extender el estudio de este modelo para bases de datos de más largo aliento que pudiera requerir la inclusión de un componente estocástico. Consideramos que la adición de este componente permitiría describir con mayor exactitud las trayectorias y correlación serial de la variable dependiente en estudio.

- Para desarrollar simulaciones, similares a la presentada en el capítulo 5, se sugiere ejecutar el algoritmo en paralelo, con la finalidad de disminuir el tiempo de ejecución. En nuestro caso, gracias a la ejecución del algoritmo mediante *Legión* (Sistema de cómputo en malla con una capacidad máxima estimada de 10^{12} operaciones matemáticas por segundo, desarrollado por la Pontificia Universidad Católica del Perú, que facilita el desarrollo de proyectos de investigación con procesos que requieren de grandes cálculos), el tiempo de ejecución disminuyó notablemente.

- En el capítulo 6, se pudo apreciar que debido a que la cantidad de pacientes en cada clase latente era muy pequeña, los intervalos de confianza de las trayectorias de PSA resultaron ser muy amplios. Una recomendación para futuras investigaciones respecto a este modelo es intentar modelar bases de datos con la mayor cantidad de individuos posible a fin de obtener estimaciones más precisas.

Apéndice A

Rutinas en R

A.1. Programa en R para la Simulación

```
### MODELO JOINTLCMM PARA 2 CLASES ########################################
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H

#install.packages("lcmm")
library("lcmm")
library("NormPsy")
paquid$normMMSE <- normMMSE(paquid$MMSE)
paquid$age65 <- (paquid$age-65)/10
head(paquid)
paquidS       <- paquid[paquid$agedem > paquid$age_init, ]
paquidS$timeDif <- paquidS$agedem - paquidS$age_init


### DATOS #################################################################
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H

## ID y número de observaciones
dataSinDup <- paquidS[!duplicated(paquidS$ID), ] #Data sin duplicados
ti <- as.numeric(table(paquidS$ID)); table(ti)   #Número de obs: 1-9 oberv
obs <- unlist(sapply(ti, function(x) sort(1:x))) #obs en formato largo

##### CLASES
probPosts <- read.csv(file.path(getwd(), "pprob.csv"),header = T, sep = ",")
probPosts <- ifelse(probPosts==1, 2, 1)
probPosts <- as.matrix(probPosts)
probPost <- rep(probPosts, ti)
paquidS  <- cbind(paquidS, probPost)    #Agrerar probabilidades aposteriori a data

claseC   <- probPosts  ;  table(claseC)
prop.table(table(claseC))
prop.table(table(claseC))*100
```

```
###Número de pacientes por clase
n1 <- table(claseC)[1] ;n1
n2 <- table(claseC)[2] ;n2
n  <- n1+n2            ; n  #Número total de pacientes


##### COVARIABLES
dataSim <- as.data.frame(cbind(paquidS$ID, obs,
paquidS$age65,
paquidS$CEP,
paquidS$male,
paquidS$agedem,
paquidS$age_init,
paquidS$dem,
(paquidS$agedem - paquidS$age_init),
paquidS$probPost))
colnames(dataSim) <- c("ID", "obs", "age65", "CEP",
"male", "agedem", "age_init",
"dem", "timeDif", "clase")
summary(dataSim$timeDif)
N  <- length(dataSim$obs)  ; N

#paRr <- read.csv("pparam2c.csv")
#paRr <- as.matrix(paRr)

### Valores para simular data
#### Valores reales del Modelo de pertenencia a la clase
c01 <- 1.905501

pi1 <- exp(c01)/(1+exp(c01))
pi2 <- 1 - (pi1)


#### Valores reales del Modelo de supervivencia
w1c1 <- 0.3437324  ; w2c1 <- 1.641782
w1c2 <- 0.1834847  ; w2c2 <- 4.092174
CEP  <- 0.2055406

### Valores reales de los parámetros
b01 <- 89.86019 ; b11 <- -27.33755
b02 <- 75.60979 ; b12 <- -10.27664
```

```
b   <- 13.07813

### Valores reales de matriz de var y covar de efectos aleatorios
s0  <- 229.7199
s01 <- -97.93176
s1  <- 78.92665

### Valor real de varianza de errores
sigma <- 10.52889


### FUNCIÓN ##########################################################
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H

Legion <- 10

M <- Legion
#Matriz para guardar las estimaciones
ressNA       <- NULL
Parametros   <- matrix(0, ncol=16 ,nrow=M)
Porcentajes  <- matrix(0, ncol=2  ,nrow=M)
DiagVar      <- matrix(0, ncol=16 ,nrow=M)


m <- 1
print(m)

t <- proc.time() # Inicia el cronómetro
for (m in 1:M) {

rr <- 0
while (rr == 0) {


#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
#### Simulación del tiempo de ocurrencia del evento #H#H#H#H#H#H#H#
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#

unif  <- runif(n, 0, 1)
argu  <- numeric(n)
timeS <- matrix(0, nrow = n, ncol = 2)
we1 <- 1/(w1c1)^2
```

```
we2 <- 1/(w1c2)^2

argu <- 1-(1-unif)^exp(-(CEP*dataSim[unique(dataSim$ID),]$CEP))

i <- 1
for (i in 1:n) {
timeS[i,1] <- qweibull(argu[i], (w2c1)^2, we1)
timeS[i,2] <- qweibull(argu[i], (w2c2)^2, we2)
}
dataSim$timeSim1 <- rep(timeS[,1], ti)
dataSim$timeSim2 <- rep(timeS[,2], ti)


### Seleccionar tiempo de acuerdo a clase de pertenencia
dataSim$timeSim <- ifelse(dataSim$clase==1, dataSim$timeSim1, dataSim$timeSim2)

table(dataSim[unique(dataSim$ID),]$dem)
table(dataSim[unique(dataSim[dataSim$clase==1,]$ID),]$dem)
table(dataSim[unique(dataSim[dataSim$clase==2,]$ID),]$dem)


#H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H#
#### Simulación del tiempo de censura #H##H##H##H##H##H##H##H##H##H##H#
#H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H##H#

arguCensura <- numeric(n)
censuraS    <- matrix(0, nrow = n, ncol = 2)
arguCensura <- 1-(1-unif)

for (i in 1:n) {
censuraS[i,1]  <- qweibull(arguCensura[i], 4, 11)
censuraS[i,2]  <- qweibull(arguCensura[i], 1, 19)
}

dataSim$censuraS1 <- rep(censuraS[,1], ti)
dataSim$censuraS2 <- rep(censuraS[,2], ti)

### Seleccionar time de censura de acuerdo a clase de pertenencia
dataSim$censuraS <- ifelse(dataSim$clase==1, dataSim$censuraS1, dataSim$censuraS2)
```

```
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
#### Generación de variable indicadora del evento #H#H#H#H#H#H#H#H#H#
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#

dataSim$Ee <- ifelse(dataSim$timeSim < dataSim$censuraS, 1, 0)

table(dataSim[unique(dataSim$ID),]$Ee)
table(dataSim[unique(dataSim$ID),]$dem)

table(dataSim[unique(dataSim[dataSim$clase==1,]$ID),]$Ee)
table(dataSim[unique(dataSim[dataSim$clase==1,]$ID),]$dem)

table(dataSim[unique(dataSim[dataSim$clase==2,]$ID),]$Ee)
table(dataSim[unique(dataSim[dataSim$clase==2,]$ID),]$dem)



#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
#### Simulación de efectos aleatorios y errores #H#H#H#H#H#H#H#H#H#
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#

mu  <- c(0,0)
S   <- matrix(c(s0, s01, s01, s1), nrow=2)
library(MASS)
a   <- mvrnorm(n, mu=mu, Sigma=S)
a0  <- a[,1]
a1  <- a[,2]
dataSim$a0  <- rep(a0,ti)
dataSim$a1  <- rep(a1,ti)

### Simulación de errors
dataSim$eij <- rnorm(N,0,sigma)



#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
#### Simulación de Yij según clase latente #H#H#H#H#H#H#H#H#H#H#H#H#
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#

### Pacientes agrupados por clase
dataC1 <- dataSim[dataSim$clase==1,]
dataC2 <- dataSim[dataSim$clase==2,]
length(dataC1$obs)+length(dataC2$obs)
```

```
##### Simulación de Yij - Para CLASE 1 ######################################
dataC1$uij <- (b01) + (b11*dataC1$age65) + (b*dataC1$CEP) +
(dataC1$a0) + dataC1$a1*(dataC1$age65) +
(dataC1$eij)
###Varianza de los Yij para C1
Zi_C1 <- cbind(rep(1,length(dataC1$age65)),(dataC1$age65))
Si_C1 <- sigma*diag(length(dataC1$age65))
Vi_C1 <- Zi_C1%*%S%*%t(Zi_C1) + Si_C1
ssdd1 <- sqrt((diag(Vi_C1)))
dataC1$yij <- apply(mvrnorm(length(dataC1$uij),mu=dataC1$uij, Sigma=Vi_C1),2,mean)


##### Simulación de Yij - Para CLASE 2 ######################################
dataC2$uij <- (b02) + (b12*dataC2$age65) + (b*dataC2$CEP) +
(dataC2$a0) + dataC2$a1*(dataC2$age65) +
(dataC2$eij)
###Varianza de los Yij para C2
Zi_C2 <- cbind(rep(1,length(dataC2$age65)),(dataC2$age65))
Si_C2 <- sigma*diag(length(dataC2$age65))
Vi_C2 <- Zi_C2%*%S%*%t(Zi_C2) + Si_C2
ssdd2 <- sqrt((diag(Vi_C2)))
dataC2$yij <- apply(mvrnorm(length(dataC2$uij),mu=dataC2$uij, Sigma=Vi_C2),2,mean)



##### Data unida
dataSIMULADA <- rbind(dataC1,dataC2)
dataSIMULADA <- dataSIMULADA[order(dataSIMULADA$ID,dataSIMULADA$obs),]



### Algoritmo para detectar errores en convergencia ########################
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H

myTryCatch <- function(expr){
warn <- err <- NULL
value <- withCallingHandlers(
tryCatch(expr, error=function(e){
err <<- e
NULL
```

```
}), warning=function(w){
warn <<- w
invokeRestart("muffleWarning")
})
list(value=value, warning=warn, error=err)
}
```

```
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
#### Modelos con LCMM #H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#H#
```

```
library(lcmm)
```

```
base <- Jointlcmm(yij ~ age65 + CEP,
random = ~ age65,
survival = Surv(timeSim, Ee) ~ CEP ,
hazard = "Weibull", subject = "ID",data = dataSIMULADA, ng=1)
summarytable(base)
base$best ; length(base$best)
```

```
m20 <- gridsearch(rep = 10,maxiter = 15,minit = base,
Jointlcmm(yij ~ age65 + CEP,
mixture = ~ age65,
random = ~ age65,
survival = Surv(timeSim, Ee) ~ CEP ,
hazard = "Weibull", subject = "ID",data = dataSIMULADA, ng=2))
summarytable(m20)

Binit2 <- rep(0,length(base$best)+5)
Binit2[c(2,3,6:7,9,11:15)] <- base$best
Binit2[c(1,4,5,8,10)]  <- m20$best[c(1,4,5,8,10)]

mj20 <- Jointlcmm(fixed = yij ~ age65 + CEP,
mixture = ~ age65,
random = ~ age65,
survival = Surv(timeSim, Ee) ~ CEP ,
hazard = "Weibull", subject = "ID",data = dataSIMULADA, ng=2, B=Binit2)
summarytable(mj20)
```

```
#Estimaciones del mejor modelo

Aa <- summarytable(mj20)[5]
Bb <- summarytable(mj20)[6]

### Condicionar el % en las clases != 0

ifelse(Aa == 0 | as.numeric(mj20$best[15]) < 0 | is.null(myTryCatch(summary(mj20))$error) =
ifelse(Bb == 0 | as.numeric(mj20$best[15]) < 0 | is.null(myTryCatch(summary(mj20))$error) =
rr <- 0, rr <-1))
rr <- rr

}


#Reordenando Clases
Parametros[m,7]  <- mj20$best[6]
Parametros[m,12] <- mj20$best[11]
Parametros[m,13] <- mj20$best[12]
Parametros[m,14] <- mj20$best[13]
Parametros[m,15] <- mj20$best[14]
Parametros[m,16] <- mj20$best[15]


DiagVar[m,7] <- as.numeric(diag(VarCov(mj20))[6])
DiagVar[m,12] <- as.numeric(diag(VarCov(mj20))[11])
DiagVar[m,13] <- as.numeric(diag(VarCov(mj20))[12])
DiagVar[m,14] <- as.numeric(diag(VarCov(mj20))[13])
DiagVar[m,15] <- as.numeric(diag(VarCov(mj20))[14])
DiagVar[m,16] <- as.numeric(diag(VarCov(mj20))[15])

newClass     <- matrix(0,ncol = 3,nrow = 2)
newClass[,2] <- as.numeric(summarytable(mj20)[5:6])
newClass[,3] <- c(1:2)
newClass <- newClass[order(newClass[,2]),]
newClass[,1] <- c(1,2)
newClass

i <- 0
```

```
for (i in 1:2) {
k <- newClass[newClass[,3]==i,1]
Porcentajes[m,k] <- summarytable(mj20)[4+i]
DiagVar[m,k]     <- as.numeric(diag(VarCov(mj20))[4+i])
}

i <- 0
for (i in 1:1) {
k <- newClass[newClass[,3]==i,1]
Parametros[m,k]   <- mj20$best[i]
DiagVar[m,k] <- as.numeric(diag(VarCov(mj20))[i])
}

i <- 0
for (i in 1:2){
k <- newClass[newClass[,3]==i,1] ; k
Parametros[m,3+(k*2-2)]  <- mj20$best[2+(i*2-2)]
Parametros[m,4+(k*2-2)]  <- mj20$best[3+(i*2-2)]
Parametros[m,k+7] <- mj20$best[i+6]
Parametros[m,k+9] <- mj20$best[i+8]

DiagVar[m,3+(k*2-1)] <- as.numeric(diag(VarCov(mj20))[2+(i*2-2)])
DiagVar[m,4+(k*2-1)] <- as.numeric(diag(VarCov(mj20))[3+(i*2-2)])
DiagVar[m,k+7] <- as.numeric(diag(VarCov(mj20))[i+6])
DiagVar[m,k+9] <- as.numeric(diag(VarCov(mj20))[i+8])
}

ressNA <- append(ressNA,list(Parametros[m,],
Porcentajes[m,],
DiagVar[m,]))

print(m)
m <- m +1
print(m)
}

proc.time()-t    # Detiene el cronómetro


save(ressNA, file=file.path(getwd(),"ressNA.rda")) #guardamos en un formato del R.
```

A.2. Programa en R para la Aplicación al conjunto de datos reales

```
### Librerías
#################################################################

#install.packages("devtools")
library(devtools)
#install_version("lcmm", version = "1.7.8", repos = "http://cran.us.r-project.org")
library(lcmm)
#install.packages("data.table")
#install.packages("ggplot2")


### Datos
#################################################################
set.seed(333)
setwd("F:/1 - PUCP/4 Seminario de Tesis II/Últimas versiones/Aplicacionnn/AplicacionF")
library(data.table)
datosCP0 <- fread("dataAplicacionTesisCaP.csv")
table(datosCP0$EC)
prop.table(table(datosCP0$EC))

datosCP           <- datosCP0
datosCP$DNI       <- as.character(datosCP$DNI)
datosCP$Fecha_Nac <- as.character.Date(datosCP$Fecha_Nac)
datosCP$Edad      <- as.numeric(datosCP$Edad)
```

```
datosCP$Fecha_Dx  <- as.character.Date(datosCP$Fecha_Dx)
datosCP$Fecha_PR  <- as.character.Date(datosCP$Fecha_PR)
datosCP$Gleason_N <- as.numeric(datosCP$Gleason_N)

datosCP$Gleason_N <- as.numeric(scale(datosCP$Gleason_N, center= TRUE, scale=TRUE))

dataPR <- datosCP
IDnew  <- seq(1,length(dataPR$Edad),1)
dataPR <- cbind(dataPR[,1],IDnew,dataPR[,-1])

## long format data
dataPRlong <- reshape(dataPR, idvar="IDnew", varying=list(c("Fecha1","Fecha2","Fecha3","Fec
"Fecha11","Fecha12","Fecha13","Fecha14","Fecha15","Fecha16","Fecha17","Fecha18","Fecha19","
"Fecha21","Fecha22","Fecha23","Fecha24","Fecha25","Fecha26","Fecha27","Fecha28","Fecha29","
c("PSA1","PSA2","PSA3","PSA4","PSA5","PSA6","PSA7","PSA8","PSA9","PSA10",
"PSA11","PSA12","PSA13","PSA14","PSA15","PSA16","PSA17","PSA18","PSA19","PSA20",
"PSA21","PSA22","PSA23","PSA24","PSA25","PSA26","PSA27","PSA28","PSA29","PSA30","PSA31","PS
v.names=c("Fecha","PSA"), timevar="time",time=1:33, direction="long")
dataPRlong <- dataPRlong[,-1]

# Tiempos irregulares
dataPRlong  <- dataPRlong[order(dataPRlong$IDnew),]
dataPRlong  <- dataPRlong[is.na(dataPRlong$PSA)==F,]
dataPRlong  <- dataPRlong[dataPRlong$PSA != "",]

ti              <- as.numeric(table(dataPRlong[,1])) ; ti
sum(ti)
dataPRlong$time <- unlist(sapply(ti, function(x) sort(1:x)))

dataPRlong$Fecha <- as.character.Date(dataPRlong$Fecha)
dataPRlong$PSA   <- as.numeric(dataPRlong$PSA)

dataPRlong$Fecha_Nac <- as.Date(dataPRlong$Fecha_Nac, format = "%d/%m/%Y")
dataPRlong$Fecha_Dx  <- as.Date(dataPRlong$Fecha_Dx, format = "%d/%m/%Y")
dataPRlong$Fecha_PR  <- as.Date(dataPRlong$Fecha_PR, format = "%d/%m/%Y")
dataPRlong$Fecha     <- as.Date(dataPRlong$Fecha, format = "%d/%m/%Y")


#Diferencia en Tiempo de toma de PSA:  (t2-t1)
dataPRlong$TimePSA <- numeric(length(dataPRlong$IDnew))
#Tiempo de toma de PSA desde PR
dataPRlong$TimePR <- numeric(length(dataPRlong$IDnew))
```

```
k <- 0
M <- max(dataPRlong$IDnew)
obs <- numeric(M)
obs <- as.vector(colSums(table(dataPRlong$time,dataPRlong$IDnew)))
for(i in 1:M){
for(j in 1:obs[i]){
if(dataPRlong$time[k+j]==1) {
dataPRlong$TimePSA[k+j]  <- dataPRlong$Fecha[k+j] - dataPRlong$Fecha_PR[k+j]
dataPRlong$TimePR[k+j]   <- dataPRlong$TimePSA[k+j]}

if((dataPRlong$time[k+j]==(j)) & (dataPRlong$time[k+j]!=1 )) {
dataPRlong$TimePSA[k+j]  <- dataPRlong$Fecha[k+j] - dataPRlong$Fecha[(k+j-1)]
dataPRlong$TimePR[k+j] <- dataPRlong$TimePSA[k+j] + dataPRlong$TimePR[k+j-1]}
}
k <- k+obs[i]
}
dataPRlong$TimePRm <- dataPRlong$TimePR/30


# Observaciones hasta FBQ
ti <- as.numeric(table(dataPRlong[,1])) ; ti
PRlong <- NULL
for(i in 1:M) {
PRlong <- rbind(PRlong, dataPRlong[dataPRlong$IDnew==i & dataPRlong$time==1,])
j <- 2
k <- 0
for (j in 2:ti[i]) {
if(k == 0){
PRlong <- rbind(PRlong, dataPRlong[dataPRlong$IDnew==i & dataPRlong$time==j,])
}
ifelse (dataPRlong[dataPRlong$IDnew==i & dataPRlong$time==j,16] >= 0.2,
k <- 1 ,
j <- j+1)
}
}

PRlong <- PRlong[order(PRlong$IDnew),]


library(ggplot2)
ggplot(PRlong, aes(x=TimePRm,y=PSA,group=IDnew,
```

```
color=IDnew)) +
geom_line()  +
geom_point() + ggtitle("Seguimiento del PSA en pacientes con Cáncer de Próstata (Post Prost
theme(plot.title = element_text(hjust = 0.5)) +
labs(x="Tiempo luego de la Prostatectomía Radical", y="PSA", size=18) +
coord_cartesian(ylim = c(0,0.3)) +
guides(color=F)


### Variable de Supervivencia
Xsv <- numeric(M)
i <- 1
for(i in 1:M) {
j <- 2
for (j in 2:length(PRlong[PRlong$IDnew==i,]$time)) {
if (PRlong[PRlong$IDnew==i & PRlong$time==j,16]>=0.2) {
Xsv[i] <- 1}
j <- j+1 }
}
ni            <- as.numeric(table(PRlong[,1])) ; ni
PRlong$FBQ <- rep(Xsv,ni)


### Tiempo hasta evento
Teve <- numeric(M)
i <- 1
for(i in 1:M) {
Teve[i] <- as.numeric(PRlong[PRlong$IDnew==i & PRlong$time==ni[i],18])
i <- i+1
}
PRlong$TimeFBQ <- rep(Teve,ni)

#Edad en cada toma de PSA
PRlong$Edad_PSA <- (PRlong$Fecha - PRlong$Fecha_Nac)/365


library(lcmm)

Modelo1 <- Jointlcmm(PSA ~ -1 + TimePR + PSA_Dx + PSA_Base + Edad,
random = ~ TimePR,
survival = Surv(TimeFBQ,FBQ) ~ PSA_Dx + Gleason_N ,
hazard = "Weibull",
```

```
subject = "IDnew",
data = PRlong, ng=1)
Modelo1$best ; length(Modelo1$best)                          #12
summarytable(Modelo1)


Modelo2 <- gridsearch(rep=10, maxiter=15, minit=Modelo1,
Jointlcmm(PSA ~ -1 + TimePR + PSA_Dx + PSA_Base + Edad,
mixture = ~ -1 + TimePR,
random = ~ TimePR,
survival = Surv(TimeFBQ,FBQ) ~ PSA_Dx + Gleason_N ,
hazard = "Weibull",
subject = "IDnew",
data = PRlong, ng=2))
Modelo2$best ; length(Modelo2$best)                          #16
summarytable(Modelo2)


Modelo3 <- gridsearch(rep=10, maxiter=15, minit=Modelo1,
Jointlcmm(PSA ~ -1 + TimePR + PSA_Dx + PSA_Base + Edad,
mixture = ~ -1 + TimePR,
random = ~ TimePR,
survival = Surv(TimeFBQ,FBQ) ~ PSA_Dx + Gleason_N ,
hazard = "Weibull",
subject = "IDnew",
data = PRlong, ng=3))
Modelo3$best ; length(Modelo3$best)                          #20
summarytable(Modelo3)


Modelo4 <- gridsearch(rep=10, maxiter=15, minit=Modelo1,
Jointlcmm(PSA ~ -1 + TimePR + PSA_Dx + PSA_Base + Edad,
mixture = ~ -1 + TimePR,
random = ~ TimePR,
survival = Surv(TimeFBQ,FBQ) ~ PSA_Dx + Gleason_N ,
hazard = "Weibull",
subject = "IDnew",
data = PRlong, ng=4))
Modelo4$best ; length(Modelo4$best)                          #24
summarytable(Modelo4)
```

```
Modelo5 <- gridsearch(rep=10, maxiter=15, minit=Modelo1,
Jointlcmm(PSA ~ -1 + TimePR + PSA_Dx + PSA_Base + Edad,
mixture = ~ -1 + TimePR,
random = ~ TimePR,
survival = Surv(TimeFBQ,FBQ) ~ PSA_Dx + Gleason_N ,
hazard = "Weibull",
subject = "IDnew",
data = PRlong, ng=5))
Modelo5$best ; length(Modelo5$best)                          #28
summarytable(Modelo5)



summarytable(Modelo1, Modelo2, Modelo3, Modelo4, Modelo5)



# class-membership diagnosticada
CM_real       <- NULL
CM_real       <- dataPR[,c(2,10)]
CM_real$ecRec <- rep(0,length(CM_real$EC))
for(i in 1:length(CM_real$EC)){
if (CM_real$EC[i] == "I")   CM_real$ecRec[i] <- 1
if (CM_real$EC[i] == "II")  CM_real$ecRec[i] <- 2
if (CM_real$EC[i] == "III") CM_real$ecRec[i] <- 3
if (CM_real$EC[i] == "IV")  CM_real$ecRec[i] <- 4
}


newClassA4     <- matrix(0,ncol = 3,nrow = 4)
newClassA4[,2] <- as.numeric(summarytable(Modelo4)[5:8])
newClassA4[,1] <- c(1:4)
newClassA4[,3] <- c(4,3,2,1)
newClassA4

# class-membership estimada por el modelo
CM_estimada4        <- NULL
CM_estimada4$Cat    <- Modelo4$pprob[,2]; length(CM_estimada4$Cat)
CM_estimada4$CatRec <- rep(0,length(CM_estimada4$Cat))
for(i in 1:length(CM_estimada4$Cat)){
for(j in 1:length(newClassA4[,1]))
if (CM_estimada4$Cat[i]    == newClassA4[j,1]) {
```

```
#     Amj4$pprob[i,2]       <- newClassA[j,3]
CM_estimada4$CatRec[i] <- newClassA4[j,3] }
}
CM_estimada4 <- as.data.frame(CM_estimada4)

CM4 <- cbind(CM_real,CM_estimada4)
table(CM4$ecRec,CM4$CatRec)
table(datosCP$EC)



### RESULTADOS
########################################
summarytable(Modelo1, Modelo2, Modelo3, Modelo4, Modelo5)


### Estadio clinico diagnosticado VS Clase Latente (4)
CM4 <- cbind(CM_real,CM_estimada4)
table(CM4$ecRec,CM4$CatRec)


### Estadio clinico diagnosticado
table(datosCP$EC)


par(mfrow=c(1,2))
plot(Modelo1,which = "fit" ,var.time = "TimePR",break.times = 33,
bty="l", ylab="PSA",xlab="Días desde la Prostatectomía Radical",
xlim=c(0,1400),ylim=c(-0.5,0.7), main="PSA predicho y observado, según clase latente")
plot(Modelo4,which = "fit" ,var.time = "TimePR",break.times = 33,
bty="l", ylab="PSA",xlab="Días desde la Prostatectomía Radical",
xlim=c(0,1400),ylim=c(-0.5,0.7), main="PSA predicho y observado, según clase latente")
par(mfrow=c(1,1))


par(mfrow=c(1,2))
plot(Modelo1,which = "survival" ,var.time = "TimePR",break.times = 30,
bty="l", ylab="Probabilidad libre de FBQ",
xlab="Días desde la Prostatectomía Radical",
main="Supervivencia libre de Falla Bioquímica")
plot(Modelo4,which = "survival" ,var.time = "TimePR",break.times = 30,
bty="l", ylab="Probabilidad libre de FBQ",
```

```
xlab="Días desde la Prostatectomía Radical",
main="Supervivencia libre de Falla Bioquímica")
par(mfrow=c(1,1))
```

Bibliografía

Commenges, D. y Jacqmin-Gadda, H. (2015). *Dynamical Biostatistical Models*, Vol. 86, CRC Press.

Cox, D. R. (1972). Regression models and life-tables, *Breakthroughs in statistics*, Springer, pp. 527–541.

Cupani, M. (2012). Análisis de ecuaciones estructurales: conceptos, etapas de desarrollo y un ejemplo de aplicación, *REVISTA TESIS Facultad De Psicología* **2**(1): 186–199.

McCulloch, C. E. (2003). *Generalized Linear Mixed Models*, IMS, NSF-CBMS regional conference series in probability and statistics, Volume 7.

Pedrero, V., Cabieses, B. y Bernales, M. (2015). El potencial de las variables latentes en investigación en salud, *Revista médica de Chile* **143**(6): 814–815.

Proust-Lima, C., Philipps, V., Diakite, A. y Liquet, B. (2017). *lcmm: Extended Mixed Models Using Latent Classes and Latent Processes*. R package version: 1.7.7.
URL: *https://cran.r-project.org/package=lcmm*

Proust-Lima, C., Philipps, V. y Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: the r package lcmm, *arXiv preprint arXiv:1503.00890* .

Stirrup, O. T., Babiker, A. G., Carpenter, J. R. y Copas, A. J. (2015). Fractional brownian motion and multivariate-t models for longitudinal biomedical data, with application to cd4 counts in hiv-patients, *Statistics in medicine* .

Taylor, J. M., Cumberland, W. y Sy, J. (1994). A stochastic model for analysis of longitudinal aids data, *Journal of the American Statistical Association* **89**(427): 727–736.

Valdivieso, L. (2007). Likelihood inference in processes of ornstein-uhlenbeck type.