PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELO LINEAL MIXTO DE CLASES LATENTES CON RESPUESTA ORDINAL Y SU APLICACIÓN EN LA MEDICIÓN DE LA RELIGIOSIDAD

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Ivonne Mireille Renteria Sacha

Asesor: Dr. Luis Hilmar Valdivieso Serrano

Miembros del jurado:
Dr. Luis Hilmar Valdivieso Serrano
Dr. Cristian Luis Bayes Rodríguez
Dra. Rocío Paola Maehara Aliaga de Benites

Lima, Agosto del 2019

Dedicatoria

A mi esposo, Jorge, quien me animó a emprender esta aventura académica. Su amor y apoyo fueron vitales en toda la maestr´ıa.

A mis padres, Manuel y Bethy, quienes desde pequeña me motivaron a emprender y dar mi opinión.

A mi hermano José Manuel, quien me enseña día a día a celebrar la diferencia.

Agradecimentos

Mi mayor gratitud y mi más profunda admiración al Dr. Luis Valdivieso, por su constante guía y paciencia, y porque me enseña no solo a saber sino también a sentir y gustar la Estad´ıstica.

Al Dr. Cristian Bayes por sus magn´ıficas clases que me motivaron a emprender R-Ladies Lima.

A mis amigos y compañeros de la maestría, y a mi R-Lady favorita, Vilma Romero.

Resumen

Los modelos lineales mixtos de clases latentes desarrollados por Proust-Lima, Philipps y Liquet (2017) son útiles para analizar el aspecto dinámico y la naturaleza multidimensional de un fenómeno de interés en poblaciones no necesariamente homogéneas. Estos permiten identificar las posibles clases latentes en la población bajo estudio y cómo un conjunto de covariables afecta en cada clase a la variable respuesta de interés. En esta tesis se desarrolla el modelo lineal mixto de clases latentes con variable respuesta latente y variable manifiesta ordinal, a través de sus dos componentes: el sub-modelo estructural y el sub-modelo de medición, que son complementados con un modelo logístico multinomial para analizar la probabilidad de pertenencia a una clase latente. El modelo se aplicó a un conjunto de datos pertenecientes al Estudio Nacional de Juventud y Religión (NSYR por las siglas en inglés "National Study of Youth and Religion"), con el fin de encontrar clases latentes en el constructo religiosidad y describir su evolución. Como resultado, se identificaron tres clases latentes con trayectorias distintas para cada caso.

Palabras-clave: modelos lineales mixtos de clases latentes, clases latentes, religión, juventud.

Abstract

Latent class linear mixed models developed by Proust-Lima, Philipps y Liquet (2017) are useful to analyze the dynamic aspect and the multidimensional nature of a phenomenon of interest in populations not necessarily homogeneous. These allow to identify the possible latent classes in the population under study and how a set of covariates affects the response variable of interest in each class. In this thesis, the latent class linear mixed model with latent response variable and ordinal manifest variable is developed, through its two components: the structural sub-model and the measure sub-model, which are complemented with a multinominal logistic model to analyze the probability of belonging to a latent class. The model was applied to a dataset from the National Study of Youth and Religion (NSYR), in order to find latent classes in the religiosity construct and to describe their evolution. As a result, three latent classes were identified with different trajectories for each case.

Keywords: latent class linear mixed models, latent classes, religion, youth.

Índice general

Índice general

Lista de Abreviaturas

LMM	Modelos lineales Mixtos.
LCMM	Modelos lineales Mixtos de Clases Latentes.
LM	Máxima verosimilitud.
REML	Máxima verosimilitud Restricta.
NSYR	Estudio Nacional de Juventud y Religión
RMSE	Raíz del error cuadrático medio
DUREL	Duke University Religion Index
CESD	Center for Epidemiologic Studies Depression Scale
Irelig	Índice de Religiosidad

Lista de Símbolos

t_{ij}	Tiempo de medición del sujeto i en el momento j.
Λ_{ij}	Variable latente para el sujeto i en el tiempo t_{ij}.
$\mathbf{v}_g$	Vector de parámetros de efectos fijos para la clase g.
$\mathbf{u}_{ig}$	Vector de parámetros de efectos aleatorios para la clase g
s_{ij}	Vector de errores de medición.
H	Función de enlace del sub-modelo de medición.
$\boldsymbol{\eta}$	Vector de parámetros de la función de enlace.
D	Matriz de varianza y covarianzas de los efectos aleatorios.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

En la investigación de ciertos fenómenos se busca analizar el efecto a lo largo del tiempo o en conglomerados, que tienen ciertas variables sobre una variable cualitativa que asume valores que representan categorías de una clasificación. Por ejemplo, uno podría estar interesado en investigar los efectos en el tiempo de las variables sexo, estado civil, nivel educativo e ingresos sobre la ansiedad, la cual es una variable latente que podría aproximarse a través de un instrumento que tenga como categor´ıas a "bajo", "leve", "moderado" y "severo". En estos casos, los modelos lineales mixtos (LMM por sus siglas en inglés) no son adecuados porque la variable de interés no sigue una distribución normal, más aún, ella no es observable sino como indicamos latente. Para considerar variables respuesta de distinta naturaleza, cuantitativa o cualitativa, observable o latente, y también para permitir que la población bajo estudio pueda ser heterogénea, Proust-Lima, Philipps y Liquet (2017) proponen los modelos lineales mixtos para clases latentes al cual referiremos como LCMM por sus siglas en inglés "Latent Class Mixed Model".

Los LCMM combinan la teor´ıa de los modelos lineales mixtos y la teor´ıa de los modelos de clases latentes. Esta última tiene como idea básica encontrar clases o grupos no observados que difieren de manera cualitativa y que representan los patrones de asociación entre las observaciones. Con esta familia de modelos se puede estudiar los perfiles de las trayectorias latentes en poblaciones heterogéneas, tomando en cuenta la correlación entre medidas repetidas para el mismo sujeto y distinguiendo entre los grupos latentes. Esto último implica que el modelamiento de los perfiles de las trayectorias latentes depende de la clase latente a la cual se aplique, así se tiene que los modelos serán clase específico.

Para una mejor comprensión de los LCMM, recordemos que los modelos lineales mixtos extienden los modelos de regresión lineal al incorporar efectos aleatorios. Para ser más específicos y siguiendo a Fitzmaurice (2011), los LMM modelan la respuesta media condicional como una combinación lineal de características de toda la población y características propias de cada individuo. Las primeras son recogidas por los efectos fijos que son parámetros que no varían y sirven para caracterizar la línea de regresión poblacional, mientras que las segundas son recogidas por los efectos aleatorios y como se señala en Galecki y Burzykowski (2013), son parámetros que son variables aleatorias que cambian de un sujeto a otro.
Un modelo lineal mixto tiene la siguiente forma:

$$Y_{ij} = \underset{ij}{\mathbf{X}^T}\boldsymbol{\beta} + \underset{ij}{\mathbf{Z}^T}\mathbf{u}_i + s_{ij}, \tag{1.1}$$

donde $i = 1, \ldots, N$ denota a los sujetos bajo estudio, quienes son observados n_i veces y $j = 1, \ldots, n_i$ corresponde a la ocasión en que se registran las variables para el tiempo t_{ij}. Se hace la distinción entre ocasión y tiempo porque no todos los sujetos del estudio tienen el mismo número de observaciones y además porque los tiempos de medición no necesariamente coinciden. Se define Y_{ij} como la variable aleatoria para el sujeto i en el tiempo t_{ij}. El vector $\mathbf{X}_{ij}$ corresponde a las variables explicativas asociadas con los efectos fijos, el vector $\boldsymbol{\beta}$ corresponde a los coeficientes de regresión de efectos fijos, el vector $\mathbf{Z}_{ij}$ corresponde a las variables explicativas asociadas a los efectos aleatorios, el vector $\mathbf{u}_i$ contiene los efectos aleatorios y por último s_{ij} representa el error de medición que no es explicado por el modelo.

Llegado hasta este punto se ha considerado dos fuentes de variabilidad aleatoria, los efectos aleatorios y el error de medición, no obstante de acuerdo a Diggle y Heagerty (2002) existe una tercera fuente de variabilidad aleatoria en los datos longitudinales: la correlación serial en la secuencia de mediciones tomados dentro de cada sujeto, siendo esta correlación modelada por un proceso estocástico. En un artículo previo Diggle (1988) indicó que la secuencia de registros contiene información que debe ser retenida en el análisis de las trayectorias de los sujetos, como el número de mediciones por sujetos y los tiempos en las que son recogidas; y por consiguiente el autor propone agregar un proceso Gaussiano estacionario, $W_i(t_{ij})$, al modelo lineal mixto. Para ilustrar esta propuesta y justificar la inclusión del proceso estocástico, Diggle usa como ejemplo el estudio de los perfiles de contenido protéico de muestras de leche de vaca y sostiene que un modelaje biológicamente plausible debe incluir un elemento que refleje el proceso bioquímico estocástico que experimenta cada sujeto de estudio.

Al adicionar el proceso estocástico o componente de correlación serial se tiene el siguiente modelo lineal mixto extendido:

$$Y_{ij} = \underset{ij}{\mathbf{X}^T}\boldsymbol{\beta} + \underset{ij}{\mathbf{Z}^T}\mathbf{u}_i + W_i(t_{ij}) + s_{ij}. \tag{1.2}$$

Como indican Taylor et al. (1994) al adicionar la componente W_i, el modelo (1.2) permite modelar la información de las secuencias de observaciones que no están igualmente espaciadas dentro de cada sujeto de estudio. En vista de ello, el modelo extendido puede evaluar si los sujetos mantienen su trayectoria en el tiempo aunque vale aclarar que este modelamiento tendrá sentido si se tiene un número considerable de observaciones en el tiempo. En la presente tesis, dado el número reducido de mediciones por sujeto, 4, no se incluirá un componente estocástico, el cual si podría ser relevante de tenerse más mediciones.

Una limitante importante que tienen los modelos lineales mixtos es que consideran a la variable respuesta Y_i como una realización de un vector aleatorio Gaussiano. Sin embargo, una proporción significativa de datos longitudinales no cumplen con este requerimiento. Se observan estudios en las ciencias sociales que recogen variables respuestas ordinales, nominales, binarias o incluso continuas pero con alguna restricción. Hay que mencionar que en el presente trabajo nuestra variable de interés será una variable respuesta latente aproximada por una variable manfiesta de naturaleza cualitativa, lo cual escapa del marco teórico de los

LMM para el análisis de datos longitudinales.

Recientemente en la literatura han sido propuestos los Modelos Lineales Mixtos de Clases Latentes, los cuales extienden (1.2) no sólo a poblaciones heterogéneas sino también a variables respuesta de tipo latente. En su forma más simple el modelo, al cual llamaremos Modelo Lineal Mixto Latente (LCMM), se compone por un sub-modelo estructural, que explica la variable latente a través de covariables y el tiempo, y por un sub-modelo de medición, que aproxima la variable latente a través de variables observables. El sub-modelo estructural estudia al proceso latente de interés Λ_{ij} para el sujeto i en el tiempo t_{ij} de acuerdo al modelo lineal mixto:

$$\Lambda_{ij} = \underset{ij}{\mathbf{X}^T}\boldsymbol{\beta} + \underset{ij}{\mathbf{Z}^T}\mathbf{u}_i + W_i(t_{ij}). \tag{1.3}$$

De otro lado, el sub-modelo de medición vincula la variable manifiesta u observable Y_{ij} con el proceso latente de interés Λ_{ij} mediante:

$$Y_{ij} = H(\tilde{Y}_{ij}; \boldsymbol{\eta}) = H(\Lambda_{ij} + s_{ij}; \boldsymbol{\eta}), \tag{1.4}$$

siendo H una función de enlace cuya forma depende de la naturaleza de la variable manifiesta, $\tilde{Y}_{ij} = \Lambda_{ij} + s_{ij}$ que denota a un proceso latente con ruido en el tiempo t_{ij} y $\boldsymbol{\eta}$ corresponde a los parámetros asociados a la función de enlace.

El LCMM se complementa con:

- Un sub-modelo log´ıstico multinomial que describe la probabilidad de pertenencia de cada uno de los sujetos a cada una de las tt clases latentes:

$$\pi_{ig} = P(C_i = g|\tilde{\mathbf{X}}_i) = \frac{\exp(\xi_{0g} + \tilde{\mathbf{X}}_i^T\boldsymbol{\xi}_{1g})}{\sum_{l=1}^{tt}\exp(\xi_{0l} + \tilde{\mathbf{X}}_i^T\boldsymbol{\xi}_{1l})}, \tag{1.5}$$

  donde C_i es una variable aleatoria discreta latente que determina la pertenencia del sujeto i a una de las tt clases latentes, ξ_{0g} es el intercepto de la clase latente g y $\boldsymbol{\xi}_{1g}$ es el vector de parámetros clase específico asociado al vector de covariables independientes del tiempo, $\tilde{\mathbf{X}}_i$.

- Un sub-modelo estructural que toma la forma:

$$\Lambda_{ij}|(C_i = g) = \mathbf{X}_{1ij}^T\boldsymbol{\beta} + \mathbf{X}_{2ij}^T\mathbf{v}_g + \mathbf{Z}_{ij}^T\mathbf{u}_{ig}, \tag{1.6}$$

  donde $\mathbf{X}_{1ij}$ es el vector de covariables asociadas a los efectos fijos comunes a toda la población, $\mathbf{X}_{2ij}$ es el vector de covariables asociadas a los efectos fijos específicos $\mathbf{v}_g$ de la clase g, y la distribución de los efectos aleatorios $\mathbf{u}_{ig}$ es ahora clase específica.

En el presente proyecto de tesis, se buscará aplicar el LCMM al campo de las ciencias sociales, de manera puntual, la variable de interés será la religiosidad. Para tal fin, se empleará el paquete "lcmm" Proust-Lima et al. (2017) implementado en el software *R* y se trabajará con un conjunto de datos del Estudio nacional de Juventud y Religión.

1.2. Objetivos

El objetivo general de la tesis es presentar el Modelo Lineal Mixto de Clases Latentes(LCMM), estudiar sus fundamentos y propiedades y aplicarlo a un conjuntos de datos reales de las ciencias sociales. De manera específica:

- Revisar la literatura acerca de los modelos lineales mixtos, los modelos de variables latentes y los modelos lineales mixtos de clases latentes.
- Desarrollar el marco teórico de los modelos lineales mixtos de clase latente con una variable respuesta ordinal politómica.
- Realizar estudios de simulación considerando diferentes escenarios para evaluar el desempeño de los estimadores del modelo LCMM en la recuperación de sus parámetros.
- Aplicar el modelo propuesto a un conjunto de datos reales para estudiar el constructo latente religiosidad.

1.3. Organización del Trabajo

En el Capítulo 2, se presenta una revisión teórica de los modelos lineales mixtos, los modelos de variables latentes y los modelos lineales mixtos de clases latentes. En el Cap´ıtulo 3, se estudia los modelos lineales mixtos de clases latentes con variable respuesta politómica; considerando sus propiedades, la metodología de estimación de los parámetros y su bondad de ajuste. En el Capítulo 4, se presenta un estudio de simulación considerando diferentes escenarios. En el Capítulo 5, se presenta la aplicación del modelo a un conjunto de datos para modelar trayectorias de religiosidad de jovénes americanos desde su adolescencia hasta su adultez. Luego, en el Cap´ıtulo 6 se discuten algunas conclusiones obtenidas en este trabajo. Se analizan las ventajas y desventajas del modelo propuesto, y se plantea sugerencias para trabajos posteriores. Finalmente, en el anexo se presentan los programas utilizados tanto para el estudio de simulación como para la aplicación al conjunto de datos reales. (Apéndice A). Se incluye sólo los códigos de uno de los escenarios de simulación, siendo el otro completamente análogo.

Capítulo 2

Conceptos preliminares

En este capítulo presentaremos una breve revisión teórica de los modelos lineales mixtos, los modelos de variables latentes y los modelos de clases latentes, con la finalidad de entender cabalmente el modelo lineal mixto de clases de latentes, objeto de estudio en esta tesis.

2.1. Modelos Lineales Mixtos

En todos los campos de investigación científica se tiene interés por estudiar cómo cambia un fenómeno en el tiempo. Los modelos lineales mixtos (LMM), que fueron propuestos por Laird y Ware (1982), constituyen un vehículo para tal fin. En estos se asumen que cada observación puede ser descompuesta como:

$$\text{Observación} = \text{efectos fijos} + \text{efectos aleatorios} + \text{error}.$$

En este sentido, los LMM extienden los modelos de regresión lineal al incorporar efectos aleatorios que tomen en cuenta la estructura de correlación entre las medidas repetidas de un grupo de sujetos. Así, la respuesta media condicional es modelada como una combinación lineal de características de toda la población y características propias de cada individuo. Las primeras son recogidas por los efectos fijos que son parámetros que no varían y sirven para caracterizar la línea de regresión poblacional, mientras que las segundas son recogidas por los efectos aleatorios que son específicos a cada sujeto y que son asumidas como realizaciones de una variable aleatoria.

De acuerdo a Oehlert (2012), al añadir los efectos aleatorios, la media poblacional o marginal no es afectada, en cambio sí la covarianza poblacional. Por consiguiente, los LMM sirven también para modelar la estructura de la covarianza y resulta útil cuando el interés principal de la investigación es analizar los perfiles de las trajectorias individuales; es decir, como cada sujeto se desvía de la media poblacional. De ahí que se hace inferencia sobre la varianza de los efectos aleatorios.

El modelo lineal mixto tiene la forma:

$$Y_{ij} = \underset{ij}{\mathbf{X}^T}\boldsymbol{\beta} + \underset{ij}{\mathbf{Z}^T}\mathbf{u}_i + \varepsilon_{ij}, \tag{2.1}$$

donde $i = 1, \ldots, N$ son los sujetos bajo estudio, los cuales son observados n_i veces y $j = 1, \ldots, n_i$ corresponde a la ocasión en que se registran las variables del sujeto i para el tiempo t_{ij}. Se hace la distinción entre ocasión y tiempo porque no todos los sujetos del estudio tienen el mismo número de observaciones y además porque los tiempos de medición no necesariamente

coinciden. Se define Y_{ij} como el valor que toma la variable respuesta para el sujeto i en el tiempo t_{ij}. El vector $\mathbf{X}_{ij}$ corresponde a las variables explicativas asociadas con los efectos fijos, el vector $\boldsymbol{\beta}$ corresponde a los coeficientes de regresión de efectos fijos, el vector $\mathbf{Z}_{ij}$ corresponde a las variables explicativas asociadas a los efectos aleatorios, el vector $\mathbf{u}_i$ contiene los efectos aleatorios y por último el vector s_{ij} representa el error de medición que no es explicado por el modelo.

Antes de continuar, es necesario recalcar cuales son los parámetros de interés y sobre las cuales se hace inferencia. Como se ha mencionado, cada observación se divide en tres partes. La parte de efectos fijos, $\mathbf{X}_{ij}^T\boldsymbol{\beta}$, es una combinación lineal de parámetros desconocidos $\boldsymbol{\beta}$. La parte de efectos aleatorios, $\mathbf{Z}_{ij}^T\mathbf{u}_i$, es una combinación lineal de variables aleatorias $\mathbf{u}_i$ que se asumen con distribución normal con vector de media cero y cuya matriz de varianza-covarianza depende de parámetros de varianza-covarianza, siendo estos últimos parámetros desconocidos de interés. La parte del error s_{ij} se asume que se distribuye normalmente con vector de media cero y matriz de varianza-covarianza dependiente de parámetros desconocidos. En síntesis, en el marco de los LMM los parámetros a ser estimados son los coeficientes de efectos fijos, las varianzas y covarianzas de los efectos aleatorios y la varianza del error.

En su forma matricial un LMM se representa de la siguiente manera:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{s}_i, \qquad i = 1,..,N \tag{2.2}$$

$$\mathbf{u}_i \sim \mathsf{N}_q(\mathbf{o}, \mathbf{D})$$

$$\boldsymbol{s}_i \sim \mathsf{N}_{n_i}(\mathbf{o}, \boldsymbol{\Sigma}),$$

donde $\mathbf{u}_1, \ldots, \mathbf{u}_N$, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ son vectores aleatorios independientes. Así, se tiene que $\mathbf{Y}_i \sim \mathsf{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$, donde

$$\mathbf{V}_i(\boldsymbol{\gamma}) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \boldsymbol{\Sigma},$$

siendo $\boldsymbol{\gamma}$ un vector que contiene a todos los parámetros de varianza y covarianza de las matrices $\mathbf{D}$ y $\boldsymbol{\Sigma}$.

De manera más explícita la composición de cada elemento de la ecuación (2.2) viene dada por:

$$\mathbf{Y}_i := \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ \vdots \\ Y_{in_i} \end{bmatrix}, \quad \mathbf{X}_i := \begin{bmatrix} X_{1i1} & X_{2i1} & \cdots & X_{pi1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1ij} & X_{2ij} & \cdots & X_{pij} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1in_i} & X_{2in_i} & \cdots & X_{pin_i} \end{bmatrix}, \quad \mathbf{Z}_i := \begin{bmatrix} Z_{1i1} & Z_{2i1} & \cdots & Z_{qi1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1ij} & Z_{2ij} & \cdots & Z_{qij} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1in_i} & Z_{2in_i} & \cdots & Z_{qin_i} \end{bmatrix},$$

$$\mathbf{u}_i := \begin{bmatrix} u_{i1} \\ u_{i2} \\ u_{ib} \\ \vdots \\ u_{iq} \end{bmatrix}, \quad \boldsymbol{s}_i := \begin{bmatrix} s_{i1} \\ s_{i2} \\ s_{ib} \\ \vdots \\ s_{in_i} \end{bmatrix}.$$

En las dos subsecciones siguientes desarrollaremos dos métodos de estimación para los parámetros de este modelo.

2.1.1. Estimación

La estimación de los parámetros del modelo se puede realizar mediante el método de máxima verosimilitud o el método de máxima verosimilitud restringida. A continuación se desarrolla cada uno de estos métodos.

El método de estimación de máxima verosimilitud (ML) estima simultáneamente los parámetros de interés $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ mediante la maximización de la función de distribución conjunta de $\mathbf{Y} = [\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_N^T]^T \in \mathsf{R}^M$ donde $M = \sum_{i=1}^{N} n_i$. Cabe tener en cuenta que las n_i medidas repetidas para el sujeto i no pueden ser asumidas como independientes pero sí se asume que los vectores $\mathbf{Y}_i$ son independientes entre si. La función de densidad de cada $\mathbf{Y}_i \in \mathsf{R}^{n_i}$ es

$$f_{\mathbf{Y}_i}(\mathbf{y}_i) = (2\pi)^{-\frac{n_i}{2}} |\mathbf{V}_i(\boldsymbol{\gamma})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i(\boldsymbol{\gamma})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right).$$

De acuerdo con estas consideraciones la forma de la función de verosimilitud es:

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

donde $\mathbf{y}_i$ denota al vector de respuestas observadas para el sujeto i.

La estimación por máxima verosimilitud trae consigo algunos problemas. De acuerdo a Commenges y Jacqmin-Gadda (2016), en varios estudios se ha identificado una subestimación sistemáticamente de los parámetros de las varianzas; es decir, los estimados presentan sesgos negativos significativos, y por otra parte, su ejecución puede resultar compleja en términos computacionales cuando se tiene varios parámetros para los efectos aleatorios.

Un método de estimación alternativo es el de Máxima Verosimilitud Restringida (REML por sus siglas en inglés), que estima los componentes de la varianza para luego estimar los parámetros de efectos fijos por mínimos cuadrados generalizados. La ventaja de este método con respecto al de ML es que los valores estimados de $\boldsymbol{\beta}$ sólo dependen del vector $\boldsymbol{\gamma}$, el cual caracteriza las matrices de varianzas-covarianzas $\mathbf{V}_i$. Esto se logra a través de una nueva variable $\mathbf{S}$ que resulta de una combinación lineal de $\mathbf{Y}$ que elimina los efectos fijos, así esta variable tiene media igual a cero. Para definir la nueva variable se necesita previamente contar con la matriz de proyección:

$$\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{2.4}$$

La nueva variable $\mathbf{S}$ resulta de la combinación siguiente:

$$\mathbf{S} = \mathbf{AY}, \tag{2.5}$$

donde se tiene que $\mathbf{S} \sim N_M(0, \mathbf{H})$ para $\mathbf{H}(\boldsymbol{\gamma}) = \mathbf{A}\,\mathrm{Var}(\mathbf{Y})\,\mathbf{A}^T = \mathbf{A}\,\mathbf{V}(\boldsymbol{\gamma})\,\mathbf{A}^T$.

Siguiendo a Fitzmaurice (2011), se describe brevemente el funcionamiento de REML. En primer lugar, se obtiene los residuos modelados por la parte de efectos fijos, luego se estima las varianzas v´ıa ML usando los residuos, y finalmente, se obtiene los estimados de los parámetros de efectos fijos tomando como correctos los estimados de los componentes de la varianza.

Para obtener, $\boldsymbol{\gamma}_{\mathrm{REML}}$, el estimador de máxima verosimilitud restringida del vector de parámetros que definen $\mathbf{V}_i$, se maximiza la siguiente función de log-verosimilitud:

$$\log L_{\mathrm{REML}}(\boldsymbol{\gamma}) \propto -\frac{1}{2}\Big[\log|\mathbf{H}(\boldsymbol{\gamma})| + \mathbf{Y}^T\mathbf{A}^{-1}\mathbf{H}(\boldsymbol{\gamma})^{-1}\mathbf{AY}\Big]. \tag{2.6}$$

En último lugar, se tiene que el estimador de $\boldsymbol{\beta}$ es el estimador de mínimos cuadrados generalizados:

$$\hat{\boldsymbol{\beta}}_{\mathrm{REML}} = \hat{\boldsymbol{\beta}}_{\mathrm{GLS}}(\hat{\boldsymbol{\gamma}}_{\mathrm{REML}}) = (\mathbf{X}^T\mathbf{H}(\hat{\boldsymbol{\gamma}}_{\mathrm{REML}})^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}(\hat{\boldsymbol{\gamma}}_{\mathrm{REML}})^{-1}\mathbf{Y}. \tag{2.7}$$

2.2. Modelos de variables latentes

Los modelos de variables latentes son una herramienta de análisis multivariado donde algunas de la variables son no observables. Según Collins y Lanza (2013), las variables latentes son variables aleatorias que:

1. No son medibles u observables directamente, pero pueden ser aproximadas de manera indirecta a través de una o más variables manifiestas u observables, y

2. No están sujetas a error, a diferencia de las variables observables.

Ejemplos de variables latentes son por citar la inteligencia, confianza empresarial, religiosidad, preferencia política, etc. La Figura 2.1 muestra la estructura de una variable latente, donde las variables manifiestas u observables $X_1, X_2, \ldots, X_p$ aproximan la variable latente, y los errores $e_1, e_2, \ldots, e_p$ están asociados con sus respectivas variables observables. Un concepto importante a subrayar está contenido en la dirección de las flechas, las cuales indican flujo causal. Por consiguiente, las variables observables son causadas tanto por la variable latente como por el error. En otras palabras, las variables observables miden o aproximan la variable latente, pero no la causan. Naturalmente estas misma variables observables podrían también estar influenciadas por otras variables latentes.

Para la formulación del modelo de variables latentes, sea $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ el vector de variables manifiestas u observables con función de probabilidad y/o distribución conjunta $f(\mathbf{x})$ y sea $\mathbf{Y} = (Y_1, Y_2, \ldots, X_q)^T$ el vector de variables latentes con función de probabilidad y/o distribución conjunta $p(\mathbf{y})$, con $q < p$. Siguiendo a Bartholomew, Knott y Moustaki (2011), el modelo de variables latentes está compuesto de dos partes:

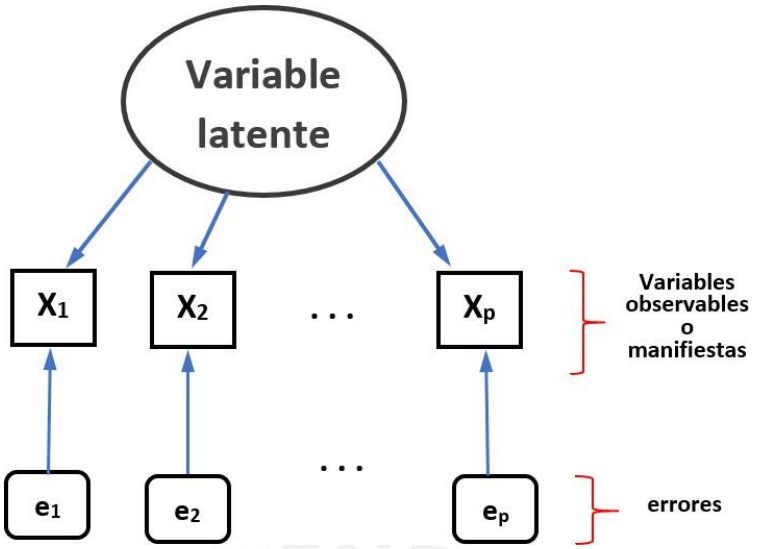

Figura 2.1: Estructura de una variable latente (Collins y Lanza, 2013, p. 45).

1. La función $p(\mathbf{y})$ que se le denomina también como la función de probabibilidad y/o densidad a priori de $\mathbf{Y}$, y

2. El conjunto de funciones probabilidad y/o densidad condicionales $g(\mathbf{x}|\mathbf{y})$ de las variables observables $\mathbf{X}$ dado que las variables latentes $\mathbf{Y}$ toman el valor $\mathbf{y}$.

Puesto que no podemos observar $\mathbf{Y}$, toda inferencia sobre ella deberá basarse en las variables observables $\mathbf{X}$ cuya distribución conjunta la podemos escribir como:

$$f(\mathbf{x}) = \int g(\mathbf{x}|\mathbf{y})p(\mathbf{y})d\mathbf{y}. \tag{2.8}$$

La función $g(\mathbf{x}|\mathbf{y})$ vincula la variable observable $\mathbf{X}$ con la variable latente $\mathbf{Y}$ y se conoce también como el modelo de medición.

Un supuesto crucial en el modelo de variables latentes es la independencia condicional local, de manera que las variables manifiestas u observables serán independientes entre ellas cuando se condicionen respecto a las variables latentes. Como se observa en la Figura 2.1 para $q = 1$, las variables observables o manifiestas sólo se relacionan entre ellas a través de la variable latente; es decir, son independientes entre ellas después de estar condicionadas respecto a la variable latente. Si no estuviera presente la variable latente, las variables observables estarían correlacionadas. El supuesto de independencia condicional se plantea de la siguiente manera:

$$g(\mathbf{x}|\mathbf{y}) = g(x_1, x_2, \ldots, x_p|\mathbf{y}) = \prod_{i=1}^{p} g_i(x_i|\mathbf{y}), \tag{2.9}$$

donde $g_i(x_i|\mathbf{y})$ es la función de densidad o probabilidad condicional de X_i dado que $\mathbf{Y}$ toma el valor $\mathbf{y}$. La distribución conjunta de las variables respuestas observables se obtiene

reemplazando la ecuación (2.9) en la (2.8), y tiene entonces la siguiente forma:

$$f(\mathbf{x}) = \int \prod_{i=1}^{p} g_i(x_i|\mathbf{y}) p(\mathbf{y}) d\mathbf{y}. \tag{2.10}$$

Esta última ecuación es la base de la inferencia en el modelo, y requiere del conocimiento de las distribuciones condicionales $g_i(x_i|\mathbf{y})$ y de la distribución a priori $p(\mathbf{y})$. Otra ecuación que es central en el estudio de estos modelos es la función probabilidad y/o densidad condicional de $\mathbf{Y}$ dado $\mathbf{X}$:

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{y})} \propto p(\mathbf{y})g(\mathbf{x}|\mathbf{y}). \tag{2.11}$$

Asemejando al enfoque bayesiano, la ecuación (2.11) es conocida como la distribución a posteriori y contiene información sobre las variables latentes una vez que se han observado las variables manifiestas.

En el presente trabajo, se considerará el caso particular de una variable latente conceptualizada como categórica. En tal caso a las categorías se les conocen también como clases latentes.

2.3. Modelos de clases latentes

Siguiendo a Valdivieso y Tarazona (2016), los modelos de clases latentes son un caso particular de los modelos de variables latentes donde tanto las variables latentes como las variables manifiestas u observables son de naturaleza categórica. La variable latente Y puede tomar K posibles valores o clases con probabilidades $\eta_j = P(Y = j)$, para $j = 0, 1, \ldots, K-1$, que satisfacen

$$\sum_{j=0}^{K-1} \eta_j = 1.$$

Dado que las variables observables $\mathbf{X}$ son también categóricas, entonces éstas podrían ser binarias, politómicas u ordinales. A continuación, se desarrolla cada uno de estos casos.

2.3.1. Caso: variables observables binarias

Si las p variables del vector $\mathbf{X}$ son dicótomicas y con probabilidad π_{ij} se tendría una respuesta positiva a la variable observable (o item) i dado que la variable latente pertenece a la clase j, entonces la función de probabilidad conjunta de $\mathbf{X}$ vendrá dada por:

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p) = \sum_{j=0}^{K-1} g(\mathbf{x}|j)\eta_j = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i}(1 - \pi_{ij})^{1-x_i}, \tag{2.12}$$

donde la función de probabilidad condicional $g(\mathbf{x}|j)$ de un sujeto con un patrón de respuestas $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ a las p variables observables o items dado que se encuentra en la clase j es un producto de funciones de probabilidad de Bernoulli.

La función de distribución a posteriori dado un vector de respuestas $\mathbf{x}$ de que un sujeto que pertenezca a la clase j viene dada por:

$$p(j|\mathbf{x}) = P(Y = j|\mathbf{X} = \mathbf{x}) = \frac{\eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i}(1 - \pi_{ij})^{1-x_i}}{f(\mathbf{x})}. \tag{2.13}$$

La ecuación (2.13) se suele utilizar como una regla de clasificación de sujetos en clases latentes.

2.3.2. Caso: variables observables nominales o politómicas

Esta es una extensión del caso anterior, donde las variables observables pueden tomar más de 2 categorías. Si c_i es el número de categorías de la variable observable i, podemos definir las siguientes variables binarias para cada categor´ıa $s = 0, \ldots, c_{i-1}$:

$$X_i(s) = \begin{cases} 1, & \text{si la rpta de un sujeto a la variable observable } i \text{ está en la categoría } s, \\ 0, & \text{caso contrario} \end{cases}$$

donde se cumple que $\sum_{s=0}^{c_i-1} X_i(s) = 1$.

La función de probabilidad conjunta del vector de respuestas $\mathbf{x}$ viene dado por

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)}, \tag{2.14}$$

donde $\pi_{ij}(s) = P(X_i(s) = 1 | Y = j)$ es la probabilidad de que un sujeto en la clase j de una respuesta positiva a la categor´ıa s de la variable observable o item i.

La función de distribución a posteriori correspondiente es

$$p(j|\mathbf{x}) = P(Y = j|\mathbf{X} = \mathbf{x}) = \frac{\eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)}}{\sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)}}. \tag{2.15}$$

2.3.3. Caso: variables observables ordinales

Si las variables observables son ordinales, entonces para tomar en cuenta el orden de sus categor´ıas se modela la probabilidad acumulada de responder positivamente en una categor´ıa igual o superior a la categor´ıa s de una variable observable i dado que el sujeto pertenece a la clase j y tiene la siguiente forma:

$$\Pi_{ij}(s) = \pi_{ij}(s) + \pi_{ij}(s+1) + \ldots + \pi_{ij}(c_i - 1).$$

La probabilidad de una respuesta en la categoría s se podrá recuperar en este caso como $\pi_{ij}(s) = \Pi_{ij}(s) - \Pi_{ij}(s+1)$.

La función de probabilidad conjunta de $\mathbf{X}$ viene dada por:

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} \pi_{ij}(s)^{x_i(s)} = \sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} (\Pi_{ij}(s) - \Pi_{ij}(s+1))^{x_i(s)}, \tag{2.16}$$

donde $x_i(s) = 1$ si el sujeto responde positivamente a la categor´ıa s de la variable observable i y $x_i(s) = 0$ en otro caso.

La función de distribución a posteriori viene dada por:

$$p(j|\mathbf{x}) = P(Y = j|\mathbf{X} = \mathbf{x}) = \frac{\eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} (\Pi_{ij}(s) - \Pi_{ij}(s+1))^{x_i(s)}}{\sum_{j=0}^{K-1} \eta_j \prod_{i=1}^{p} \prod_{s=0}^{c_i-1} (\Pi_{ij}(s) - \Pi_{ij}(s+1))^{x_i(s)}}. \tag{2.17}$$

Capítulo 3

Modelo lineal mixto de clases latentes

Los modelos lineales mixtos de clases latentes estudian la presencia de una posible estructura latente en datos longitudinales de una variable respuesta que se ven afectados por un predictor lineal constituido por un conjunto de covariables. Asumen más especificamente que la población está compuesta de subpoblaciones con características específicas que determinan sus perfiles de trayectorias. Una ventaja fundamental de estos modelos es que soportan variables respuesta observables y no observables, siendo estas no necesariamente gaussianas; así ellas brindan una herramienta útil a los estudios longitudinales en las ciencias sociales y del comportamiento.

En la presente tesis se estudiará el modelo lineal mixto de clases latentes considerando una variable manifiesta ordinal que aproxima a una variable de respuesta latente. Más precisamente, la variable latente afecta directamente la probabilidad de que un sujeto dé una respuesta en una categoría específica de la variable observable (Moustaki, 2003).

En estos modelos que los abreviaremos en su nombre por LCMM se asume que cada clase latente está caracterizada a través de los dos sub-modelos siguientes:

1. Un sub-modelo estructural que explica la variable latente a través de covariables y el tiempo.

2. Un sub-modelo de medición que aproxima la variable latente a través de variables observables.

A continuación, exploraremos en detalle cada uno de los sub-modelos para la formulación de los LCMM.

3.1. Modelo Lineal Mixto de Clases Latentes

Siguiendo a Proust, Amieva y Jacqmin-Gadda (2013), supongamos que existen tt clases latentes o valores de una variable categórica no observable, donde cada sujeto i es clasificado en una y sólo una de estás clases según el modelo de regresión logística multinomial siguiente:

$$\pi_{ig} = P(C_i = g|\tilde{\mathbf{X}}_i) = \frac{\exp(\xi_{0g} + \tilde{\mathbf{X}}_i^T \boldsymbol{\xi}_{1g})}{\sum_{l=1}^{tt} \exp(\xi_{0l} + \tilde{\mathbf{X}}_i^T \boldsymbol{\xi}_{1l})}, \tag{3.1}$$

siendo C_i una variable aleatoria discreta latente que indica la clase a la que pertenece el sujeto i, π_{ig} la probabilidad de que el sujeto i pertenezca a la clase latente g, $\tilde{\mathbf{X}}_i$ un vector de covariables independientes del tiempo, ξ_{0g} un intercepto de la clase g y $\boldsymbol{\xi}_{1g}$ un vector de

coeficientes para la clase latente g asociado al vector de covariables $\tilde{\mathbf{X}}_i$.

Seguidamente se considera para el sujeto i un **sub-modelo estructural** en el que su respuesta latente en el tiempo t_{ij}, Λ_{ij}, es explicada a través de un modelo lineal mixto sin error de medición que toma la forma siguiente:

$$\Lambda_{ij}|(C_i = g) = \mathbf{X}_{1ij}^T\boldsymbol{\beta} + \mathbf{X}_{2ij}^T\mathbf{v}_g + \mathbf{Z}_{ij}^T\mathbf{u}_{ig}, \tag{3.2}$$

donde:

$j = 1, 2, \ldots, n_i$ denota a la ocasión en que se registran las variables para el tiempo t_{ij},

$\mathbf{X}_{1ij}$ es un vector de covariables asociadas a los efectos fijos comunes a toda la población,

$\mathbf{X}_{2ij}$ es un vector de covariables asociadas a los efectos fijos para la clase g,

$\mathbf{Z}_{ij}$ es un vector covariables asociadas a los efectos aleatorios para la clase g,

$\boldsymbol{\beta}$ es un vector de parámetros de efectos fijos comunes a toda la población,

$\mathbf{v}_g$ es un vector de parámetros de efectos fijos para la clase g,

$\mathbf{u}_{ig}$ es un vector de efectos aleatorios para la clase g en el cual se asume que $\mathbf{u}_{ig} \sim N_q(\mathrm{o}_g\, \omega^2\mathrm{D})$.

Finalmente, el modelo asume que la variable aleatoria latente anterior, Λ_{ij}, se puede medir o aproximar a través de variables observables Y_{ij} en el tiempo t_{ij} mediante un **sub-modelo de medición** clase dependiente que toma la siguiente forma:

$$Y_{ij}|(C_i = g) = H(\Lambda_{ij} + s_{ij}|C_i = g; \boldsymbol{\eta}) = H(\tilde{Y}_{ij}|C_i = g; \boldsymbol{\eta}), \tag{3.3}$$

donde:

$\tilde{Y}_{ij} = \Lambda_{ij} + s_{ij}$ es un proceso latente con ruido que adiciona a (3.2) un ruido aleatorio s_{ij}.

H es una función de enlace discreta.

$\boldsymbol{\eta}$ es un vector de parámetros asociados a la función de enlace.

s_{ij} es un vector de errores de medición con $s_{ij} \sim N_{n_i}(\mathrm{o}, \mathrm{I}_{n_i})$ que son independientes de los $\mathbf{u}_{ig}$ y los C_i.

Además, con el fin de considerar diferentes tipos de variables respuesta observables, el sub-modelo de medición es formulado de manera flexible y no lineal a través de una apropiada función de enlace. Así, las aproximaciones de la variable latente subyacente pueden ser continuas gaussianas, continuas asimétricas, acotadas, nominales, ordinales, entre otras.

Siguiendo a Bollen (2001), se recomienda considerar lo siguiente para construir el sub-modelo de medición:

1. Establecer el concepto a estudiar.

2. Especificar correctamente la dirección de la relación existente entre las variables observables y la variable latente. Siendo esto crucial para no sesgar los estimados. Es recomendable responder a las siguientes preguntas: ¿La variable latente afecta a las variables observables? o ¿Las variables observables afectan a la variable latente? En el presente trabajo se sigue el paper de Pearce y Foster (2013) donde se considera que la variable latente, religiosidad, afecta a las variables observables diseñadas en el cuestionario del estudio. Con este supuesto, el sub-modelo de medición describe como esta variable latente explica la asociación entre las variables observables (Moustaki, 2003).

3. Establecer si se debe incluir más de una variable observable para aproximar la variable latente. Para nuestra aplicación, en la cual se estudia un concepto multidimensional complejo, se recomienda incluir multiples variables observables.

4. Establecer la naturaleza del dominio de las variables observables y de la variable latente. En el presente trabajo nuestra variable de interés será una latente conceptualizada como categórica ordinal y la variable manifiesta requerida para aproximarla será también de naturaleza categórica.

5. Especificar si más de una variable latente afecta las variables observables.

3.2. Modelo lineal mixto de clases latentes con variable respuesta ordinal

El modelo objetivo de este trabajo estará conformado por los siguientes sub-modelos:

- **Sub-modelo estructural.** Se determina, en primer lugar, la clase de pertenencia del sujeto i a través del siguiente sub-modelo logístico multinomial para la probabilidades de pertenecia:

$$\pi_{ig} = P(C_i = g|\tilde{\mathbf{X}}_i) = \frac{\exp(\xi_{0g} + \tilde{\mathbf{X}}_i^T \boldsymbol{\xi}_{1g})}{\sum_{l=1}^{tt} \exp(\xi_{0l} + \tilde{\mathbf{X}}_i^T \boldsymbol{\xi}_{1l})}, \quad i = 1, 2, \ldots, N \quad \text{y} \quad g = 1, 2, \ldots, tt-1$$

donde C_i es una variable aleatoria discreta latente que indica la clase a la que pertenece el sujeto, π_{ig} es la probabilidad de que el sujeto i pertenezca a la clase latente g, $\tilde{\mathbf{X}}_i$ es un vector de covariables independientes del tiempo de dimensión q_1, ξ_{0g} es un escalar para el intercepto de la clase g y $\boldsymbol{\xi}_{1g}$ es un vector de coeficientes de la clase latente g de dimensión q_1 asociado a las covariables en $\tilde{\mathbf{X}}_i$. Dado que se asume que existen tt clases latentes, entonces para alcanzar la identificabilidad se escogerá a la clase tt como referencia y, por lo tanto, $\xi_{0tt} = 0$ y $\boldsymbol{\xi}_{1tt} = 0$.

Luego de establecida la clase latente del sujeto i, el sub-modelo estructural es propiamente el siguiente modelo lineal mixto:

$$\Lambda_{ij}|(C_i = g) = \mathbf{X}_{1ij}^T \boldsymbol{\beta} + \mathbf{X}_{2ij}^T \mathbf{v}_g + \mathbf{Z}_{ij}^T \mathbf{u}_{ig}, \quad i = 1, 2, \ldots, N \quad \text{y} \quad j = 1, 2, \ldots, n_i$$

donde Λ_{ij} es la respuesta latente del sujeto i en el tiempo t_{ij}, $\mathbf{X}_{1ij}$ es un vector covariables asociadas a los efectos fijos comunes a toda la población de dimensión p_1, $\mathbf{X}_{2ij}$ es un vector de covariables asociadas a los efectos fijos comunes a la clase g de dimensión

p_2, $\mathbf{Z}_{ij}$ es un vector de covariables asociadas a los efectos aleatorios de dimensión q, $\boldsymbol{\beta}$ es un vector de efectos fijos comunes a toda la población de dimensión p_1, $\mathbf{v}_g$ es un vector de efectos fijos para la clase g de dimensión p_2 y $\mathbf{u}_{ig}$ es un vector de efectos aleatorios del sujeto i perteneciente a la clase g de dimensión q.

Cabe mencionar que se podría agregar a este modelo un componente estocástico gaussiano. Esto sin embargo no se incluirá aquí, pues la dimensión de la data longitudinal, n_i, para los sujetos estudiados en nuestra aplicación es muy corta, lo cual imposibilitará estimar de manera efectiva los parámetros de este componente. En caso se tenga una serie de observaciones relativamente larga para cada sujeto, sería interesante el poder incorporar tal componente.

Recordemos que cuando se asumía una población homogénea, el vector de efectos aleatorios de cualquier sujeto i, $\mathbf{u}_i$, se distribuía como:

$$\mathbf{u}_i \sim N_q(0, D).$$

No obstante, ahora que se asume una población heterogénea compuesta por más de una clase latente, se especificará que:

$$\mathbf{u}_{ig} \sim N_q(0, \omega_g^2 D),$$

donde la distribución de los efectos aleatorios u_{ig} es clase específica a través del coeficiente de proporcionalidad ω_g que recoge la intensidad de la variabilidad de cada sujeto dado que pertenece a una clase latente. Por identificabilidad, se restringe el valor de ω_{tt} a 1.

- **Sub-modelo de medición.** Este enlaza la respuesta observada con la respuesta latente de un sujeto i en el tiempo t_{ij} cuando este pertenece a la clase g mediante:

$$Y_{ij}|(C_i = g) = H(\tilde{Y}_{ij}; \boldsymbol{\eta}) = H(\Lambda_{ij} + s_{ij}|C_i = g; \boldsymbol{\eta}), \quad i = 1, 2, ..., N \quad \text{y} \quad j = 1, 2, ..., n_i,$$

donde $s_{ij} \sim N_{n_i}(0, I_{n_i})$ y la forma de la función de enlace, H, permite tomar en cuenta la naturaleza de la variable observable. Cuando las variables observables a considerar son de naturaleza categórica entonces la función H se encarga de asignar probabilidades a la recta real. En nuestro modelo, se considera una variable observable de naturaleza categórica ordinal, por lo que, la función de enlace con la cual trabajaremos será una función tipo umbral de la forma:

$$Y_{ij} = H(\tilde{Y}_{ij}; \boldsymbol{\eta}) = M_0 + A, \quad \text{si} \quad \tilde{Y}_{ij} = \Lambda_{ij} + s_{ij} \in]\eta_A^*, \eta_{A+1}^*], \; A = 0, 1, ..., M - 1,$$

donde M_0 es el valor mínimo o menor categoría de la variable respuesta observable, que sin pérdida de generalidad la podríamos considerar como 0, y $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_{M-1})$ es un vector de umbrales con valores en R que se reparametriza para asegurar que sean

crecientes mediante

$$\eta_A^* = \eta_1 + \sum_{j=2}^{A} \eta_j^2, \ \forall A = 1, 2, \ldots, M-1$$

y se conviene en que $\eta_0^* = -\infty$ y $\eta_M^* = +\infty$.

Los umbrales η_A^* son aquellos valores que dividen la linea latente continua en una serie de regiones correspondientes a categor´ıas ordinales observables Commenges y Jacqmin-Gadda (2016).

La Figura 3.1 muestra la relación que describe el sub-modelo de medición, donde el proceso latente continuo Λ_{ij} subyace tras la variable observable Y_{ij}. As´ı Λ_{ij} representa la verdadera cantidad que es medida a través de una variable observable nominal u ordinal. Se observa que el valor que toma la variable observable Y_{ij} depende del intervalo $]\eta_A^*, \eta_{A+1}^*]$ en el que se ubica la cantidad de interés $\tilde{Y}_{ij}$.

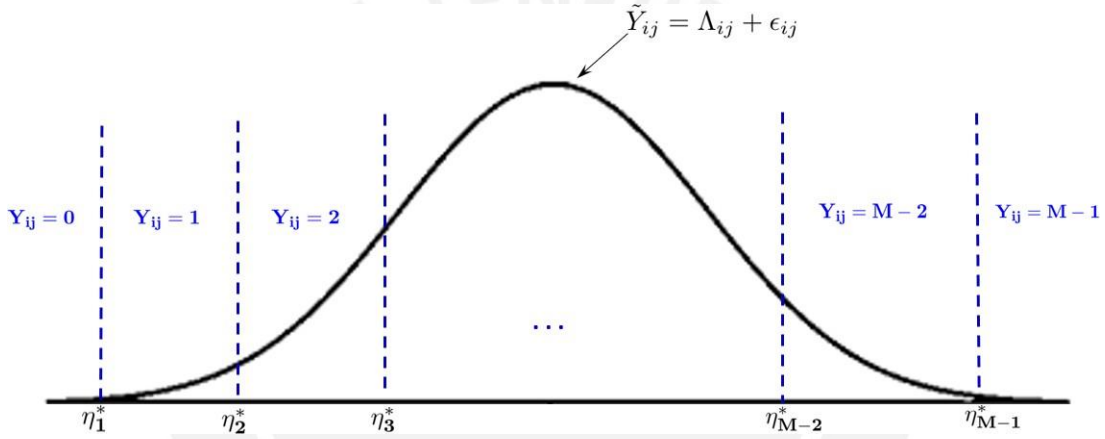

Figura 3.1: Relación entre la variable observable, el proceso latente y los umbrales (Elaboración propia).

De acuerdo al sub-modelo de medición entonces la probabilidad, condicional a los efectos aleatorios, de que un sujeto i clasificado en la clase latente g tenga una respuesta A en el periodo j, toma la forma siguiente:

$$\begin{aligned} P(Y_{ij} = A | C_i = g, \mathbf{u}_{ig} = \mathbf{u}) &= P(\eta_A^* < \tilde{Y}_{ij} \leq \eta_{A+1}^* | C_i = g, \mathbf{u}_{ig} = \mathbf{u}) \\ &= P(\eta_A^* < \Lambda_{ij} + s_{ij} \leq \eta_{A+1}^* | C_i = g, \mathbf{u}_{ig} = \mathbf{u}) \\ &= P(\eta_A^* - \lambda_{ij} < s_{ij} \leq \eta_{A+1}^* - \lambda_{ij}) \\ &= \Phi(\eta_{A+1}^* - \lambda_{ij}) - \Phi(\eta_A^* - \lambda_{ij}), \end{aligned} \tag{3.4}$$

donde $\lambda_{ij} = \mathbf{X}_{1ij}^T \boldsymbol{\beta} + \mathbf{X}_{2ij}^T \mathbf{v}_g + \mathbf{Z}_{ij}^T \mathbf{u}$ y la probabilidad dada en (3.4) resulta ser una diferencia de probabilidades acumuladas normales. Además, como veremos más adelante, esta probabilidad representa la contribución parcial a la verosimilitud de cada observación en el tiempo t_{ij}. En el Cuadro 3.1, se muestra el calculo de las probabilidades condicionales asociados a algunos valores observados.

Dado que s_{ij} y $\mathbf{u}_{ig}$ son variables aleatorias normales mutuamente independientes, en-

Función de Enlace: $Y_{ij} = H(\tilde{Y}_{ij}; \boldsymbol{\eta}) = A$	
A	Probabilidad asociada al efecto aleatorio
0	$P(Y_{ij} = 0 \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\tilde{Y}_{ij} \leq \eta_1^* \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\Lambda_{ij} + s_{ij} \leq \eta_1^* \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(s_{ij} \leq \eta_1^* - \lambda_{ij}) = \Phi(\eta_1^* - \lambda_{ij})$.
1	$P(Y_{ij} = 1 \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\eta_1^* < \tilde{Y}_{ij} \leq \eta_2^* \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\eta_1^* < \Lambda_{ij} + s_{ij} \leq \eta_2^* \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\eta_1^* - \lambda_{ij} < s_{ij} \leq \eta_2^* - \lambda_{ij}) = \Phi(\eta_2^* - \lambda_{ij}) - \Phi(\eta_1^* - \lambda_{ij})$.
.	.
M-1	$P(Y_{ij} = M-1 \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\tilde{Y}_{ij} \geq \eta_{M-1}^* \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(\Lambda_{ij} + s_{ij} \geq \eta_{M-1}^* \mid C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = P(s_{ij} \geq \eta_{M-1}^* - \lambda_{ij}) = 1 - \Phi(\eta_{M-1}^* - \lambda_{ij})$.

Cuadro 3.1: Función de enlace y la probabilidad condicional asociada.

tonces se cumple que cualquier combinación lineal sigue una distribución normal. Así, se tiene la siguiente distribución conjunta para las respuestas latentes de un sujeto i, que lo expresamos a través del vector aleatorio $\tilde{Y}_i = (\tilde{Y}_{i1}, \tilde{Y}_{i2}, \cdots, \tilde{Y}_{in_i})^T$, dada por:

$$\tilde{Y}_i | (C_i = g) \sim N_{n_i}(\mathbf{X}_{1ij}^T \boldsymbol{\beta} + \mathbf{X}_{2ij}^T \mathbf{v}_g, \mathbf{I}_{n_i} + \omega_g^2 \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T).$$

Finalmente, se debe mencionar que el modelo es clase específico, es decir, el modelamiento de las trayectorias latentes depende de la clase a la cual se aplique.

3.3. Verosimilitud

La estimación del vector de parámetros del modelo, $\boldsymbol{\theta}$, compuesto por los distintos parámetros de todos los submodelos se puede realizar simultáneamente usando el método de máxima verosimilitud. El vector de parámetros en mención viene dado por:

$$\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \text{vec}(\mathbf{D}), \boldsymbol{\eta}^T, (\xi_{0g}, \boldsymbol{\xi}_{1g}^T)_{g=1,2,\cdots,tt-1}, (\mathbf{v}_g^T)_{g=1,2,\cdots,tt}, (\omega)_{g=1,2,\cdots,tt-1}, \Sigma_T),$$

donde:

- $\boldsymbol{\beta}$ es un vector de efectos fijos comunes a toda la población. El número de parámetros a estimar es en este caso igual al número de covariables más uno.
- **vec**(D) es un vector fila de parámetros conformado por la entradas de la matriz de varianza-covarianza de los efectos aleatorios.
- $\boldsymbol{\eta}$ es el vector de parámetros de la función de enlace, siendo su número máximo igual al número de umbrales de la variable observable menos uno.
- ξ_{0g} son parámetros correspondientes a los interceptos del modelo logístico multinomial para la clase g. Por identificabilidad, se estima $tt - 1$ parámetros dado que se asume que $\xi_{0tt} = 0$.
- $\boldsymbol{\xi}_{1g}$ es un vector de parámetros clase específicos asociados con las covariables del modelo logístico multinomial. Por identificabilidad, se estima $tt - 1$ parámetros por cada covariable del modelo dado que $\xi_{1tt} = 0$.

$\mathbf{v}_g$ es un vector de efectos fijos comunes a la clase.

ω_g son parámetros correspondientes a los coeficientes de proporcionalidad para la varianzas-covarianzas de los efectos aleatorios en cada clase. Por identificabilidad, se estiman $tt - 1$ parámetros dado que se asume que $\omega_{tt} = 1$

Dado que nuestro modelo contempla datos longitudinales y la verosimilitud es la probabilidad de observar conjuntamente los datos, entonces se debe considerar la función de probabilidad conjunta de los vectores de medidas repetidas $Y_i = (Y_{i1}, Y_{i1}, \cdots, Y_{in_i})^T$. Así la función de verosimilitud tendrá la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} P(Y_i = y_i), \tag{3.5}$$

donde cada $Y_{ij} \in \{0, 1, \ldots, M-1\}$.

Considerando que nuestro modelo tiene una función de enlace discreta e incluye efectos aleatorios, entonces la contribución individual a la verosimilitud debería de condicionarse a los efectos aleatorios así como a la clase latente. Haciendo uso del teorema de probabilidad total y condicionando tanto sobre los efectos aleatorios, que son variables aleatorias continuas, como sobre las clases latentes, que son niveles de la variable latente discreta, entonces:

$$P(Y_i = y_i) = \int \sum_{g=1}^{tt} P(Y_i = y_i | C_i = g, \mathbf{u}_{ig} = \mathbf{u})\, \pi_{ig}\, \varphi_{u_{ig}}(\mathbf{u})\, d\mathbf{u}, \tag{3.6}$$

donde la integral arriba mencionada es multiple de dimensión q. Es necesario recalcar que al condicionar sobre los efectos aleatorios, las medidas repetidas de cada sujeto se vuelven independiente. Asi, es posible introducir (3.6) en (3.5) y obtener la función de verosimilitud siguiente:

$$L(\boldsymbol{\theta}) = \int \sum_{g=1}^{tt} \prod_{i=1}^{N} \prod_{j=1}^{n_i} P(Y_{ij} = y_{ij} | C_i = g, \mathbf{u}_{ig} = \mathbf{u})\, \pi_{ig}\, \varphi_{\mathbf{u}_{ig}}(\mathbf{u})\, d\mathbf{u}. \tag{3.7}$$

Reordenando se tiene que:

$$L(\boldsymbol{\theta}) = \sum_{g=1}^{tt} \prod_{i=1}^{N} \left[\int \prod_{j=1}^{n_i} P(Y_{ij} = y_{ij} | C_i = g, \mathbf{u}_{ig} = \mathbf{u})\, \varphi_{\mathbf{u}_{ig}}(\mathbf{u})\, d\mathbf{u} \right] \pi_{ig}. \tag{3.8}$$

Con respecto a los efectos aleatorios, se ha mencionado que su distribución es clase específica, de modo que podríamos caracterizar sus efectos por $\mathbf{u}_{ig} = \omega_g\, \mathbf{u}_i$ donde ω_g es un coeficiente de proporcionalidad o volatilidad por clase. La función de distribución de $\mathbf{u}_{ig}$ se puede entonces calcular por:

$$P(\mathbf{u}_{ig} \leq x) = P(\omega_g\, \mathbf{u}_i \leq x) = P(\mathbf{u}_i \leq \frac{x}{\omega_g}) = \Phi_{\mathbf{u}_i}(\frac{x}{\omega_g}).$$

$$\frac{d\Phi(\mathbf{u}_i - \boldsymbol{\mu}_g)}{d\mathbf{u}} = \varphi_{\mathbf{u}_i}(\mathbf{u}_i - \boldsymbol{\mu}_g)$$ Así $\varphi_{\mathbf{u}_i}(\mathbf{u}_i - \boldsymbol{\mu}_g) = \varphi_{\mathbf{u}_i}(\mathbf{u} - \boldsymbol{\mu}_g)$ y se cumple que

$$\varphi_{\mathbf{u}_{ig}}(\mathbf{u}) = \varphi_{\mathbf{u}_i}(\mathbf{u} - \boldsymbol{\mu}_g)\,. \tag{3.9}$$

Reemplazando (3.9) en (3.8) se tiene:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{j=1}^{n_i} P(Y_{ij} = y_{ij} | C_i = g, \mathbf{u}_{ig} = \mathbf{u})\, \varphi_{\mathbf{u}_i}(\mathbf{u} - \boldsymbol{\mu}_g)\, d\mathbf{u}\, \pi_{ig}\,. \tag{3.10}$$

Generalizando, por otro lado, la probabilidad calculada para un nivel de la variable observada en la ecuación (3.4), se tiene la siguiente expresión para la probabilidad en (3.10):

$$P(Y_{ij} = y_{ij} | C_i = g, \mathbf{u}_{ig} = \mathbf{u}) = \prod_{l=0}^{M-1} \left[\Phi(\eta_{l+1}^* - \lambda_{ij}) - \Phi(\eta_l^* - \lambda_{ij})\right]^{\mathbb{1}_{y_{ij}=l}}, \tag{3.11}$$

donde recordemos que $\lambda_{ij} = \mathbf{X}_{1ij}^T\boldsymbol{\beta} + \mathbf{X}_{2ij}^T\mathbf{v}_g + \mathbf{Z}_{ij}^T\mathbf{u}$ y $\mathbb{1}$ es una función indicadora que indica la pertenencia de un elemento a un subconjunto, tomando el valor de 1 cuando pertenece al subconjunto y cero cuando no pertenece. Luego, reemplazando (3.11) en (3.10), se obtiene la función de verosimilitud del modelo lineal mixto de clases latentes para una respuesta ordinal. Ella toma la forma:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{g=1}^{G} \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \prod_{l=0}^{M-1} \left[\Phi(\eta_{l+1}^* - \lambda_{ij}) - \Phi(\eta_l^* - \lambda_{ij})\right]^{\mathbb{1}_{y_{ij}=l}} \varphi_{\mathbf{u}_i}(\mathbf{u} - \boldsymbol{\mu}_g)\, d\mathbf{u}\, \pi_{ig} \\ &= \prod_{g=1}^{G} \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \prod_{l=0}^{M-1} \left[\Phi(\eta_{l+1}^* - (\mathbf{X}_{1ij}^T\boldsymbol{\beta} + \mathbf{X}_{2ij}^T\mathbf{v}_g + \mathbf{Z}_{ij}^T\mathbf{u})) \right. \\ &\quad \left. - \Phi(\eta_l - (\mathbf{X}_{1ij}^T\boldsymbol{\beta} + \mathbf{X}_{2ij}^T\mathbf{v}_g + \mathbf{Z}_{ij}^T\mathbf{u}))\right]^{\mathbb{1}_{y_{ij}=l}} \varphi_{\mathbf{u}_i}\left(\frac{\mathbf{u}}{\omega_g}\right) d\mathbf{u}\, \frac{\pi_{ig}}{\omega_g}\,. \end{aligned} \tag{3.12}$$

La maximización de (3.12), sin embargo, se complejiza por la presencia de la integral multiple sobre los efectos aleatorios, para lo cual se requiere integrar numéricamente cada vez que se calcula la contribución individual a la verosimilitud. El paquete “lcmm” implementado en el software R provee un conjunto de funciones para estimar modelos basados en la teoría de los modelos lineales mixtos. En este paquete, la integración numérica se calcula usando el método de cuadratura Gaussiana no adaptativa sobre regiones infinitas. En el caso que se especifique un efecto aleatorio, se integra numéricamente usando la cuadratura Gaussiana univariada con 30 puntos, y si se especifica más de un efecto aleatorio, se integra numéricamente usando la cuadratura Gaussiana multivariada.

3.4. Estimación

Con el objetivo de estimar los parámetros de interés, la función de verosimilitud o log-verosimilitud puede ser maximizada con los algoritmos de la familia EM o con algoritmos tipo Newton-Raphson. En la presente tesis se seguirá el trabajo de Proust-Lima et al. (2017) que reporta una buena tasa de convergencia utilizando un algoritmo número de tipo Newton-Raphson. Con este, el vector de parámetros $\boldsymbol{\theta}$ se actualizará iterativamente hasta alcanzar la convergencia. La ecuación que se emplea para actualizar los estimados de $\boldsymbol{\theta}$ en la iteración $k+1$ tiene la siguiente forma:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^{k} - \delta(\tilde{H}(\boldsymbol{\theta}^{k}))^{-1}\nabla(L(\boldsymbol{\theta}^{k})), \tag{3.13}$$

donde el parámetro δ se establece por defecto igual a 1 pero puede modificarse para asegurar que la verosimilitud vaya mejorando en cada iteración. La matriz $\tilde{H}(\boldsymbol{\theta}^{k})$ es la matriz Hessiana con la diagonal inflacionada en la estimación $\boldsymbol{\theta}^{k}$ que se modifica para asegurar que sea definida positiva y, por lo tanto, invertible. Los elementos inflacionados de la diagonal toman la siguiente forma:

$$\tilde{H}_{ii}(\boldsymbol{\theta}^{k}) = H_{ii}(\boldsymbol{\theta}^{k}) + \lambda\,(1-a)|H_{ii}(\boldsymbol{\theta}^{k})| + a\mathrm{tr}(H(\boldsymbol{\theta}^{k}))\Big), \tag{3.14}$$

donde $H(\boldsymbol{\theta}^{k})$ es la matriz Hessiana original y $H_{ii}(\boldsymbol{\theta}^{k})$ son sus diagonales. Tanto λ y a tienen como valor inicial a 0.01, pudiendo reducirse tal valor si $\tilde{H}(\boldsymbol{\theta}^{k})$ es definida positiva y aumentar si no lo es. $\nabla(L(\boldsymbol{\theta}^{k}))$ es el gradiente de la log-verosimilitud en la iteración anterior. Un dificultad que se presenta cuando se usan algoritmos de la familia Newton-Raphson es el cálculo analítico de las primeras y segundas derivadas. El paquete "lcmm" utiliza diferencias finitas para el cálculo de las derivadas (ver Proust-Lima et al., 2017) y utiliza como criterio de parada que la suma de distancias euclidianas entre estimaciones consecutivas o la diferencia entre logverosimilitudes consecutivas sean tan pequeñas como una tolerancia prefijada.

Finalmente, una ventaja de este algoritmo a utilizar es que brinda una estimación de la matriz de varianza y covarianza de los estimados por máxima verosimilitud $\widehat{V(\boldsymbol{\theta})}$ y está dada por la inversa de la matriz Hessiana.

Capítulo 4

Estudio de Simulación

En este capítulo se presenta un estudio de simulación para evaluar el desempeño de la función “lcmm” del paquete del mismo nombre, a través de su capacidad de recuperar los parámetros utilizados en la simulación de un conjunto de datos del modelo lineal mixto de clases latentes con una variable respuesta latente aproximada por una variable manifiesta ordinal.

Para evaluar este desempeño se calcularon los promedios, sesgos, sesgos porcentuales y raíz del error cuadrático medio (RMSE por sus siglas en inglés “Root Mean Square Error”) de los estimadores de máxima verosimilitud de los parámetros. A continuación se describe la generación de los datos.

4.1. Descripción

Para realizar nuestro estudio de simulación se consideró como contexto la base de datos *paquid* disponible en el paquete *lcmm* (Proust-Lima, Philipps y Liquet, 2017). Este es un estudio cohorte sobre 500 sujetos realizado en Francia durante 20 años para evaluar el envejecimiento cerebral y funcional. La distribución de freciencias del número de mediciones por sujeto en este periodo se muestran en el cuadro 4.1. Entre las medidas repetidas se encuentra el test CESD (siglas en inglés de Center for Epidemiologic Studies Depression Scale), que mide sintomatolog´ıa depresiva, y que tomaremos como referencia para definir nuestra variable manifiesta de respuesta ordinal.

Número de mediciones	Frecuencia
1	76
2	78
3	58
4	56
5	59
6	39
7	34
8	47
9	53

Cuadro 4.1: Distribución de frecuencias de mediciones en la base de datos.

Para evaluar la capacidad de la función *lcmm* en la identificación de clases latentes relevantes en los sujetos en estudio se llevó a cabo 1000 simulaciones trabajadas en

paralelo con el apoyo del *Proyecto Legión* de la Pontificia Universidad Católica del Perú. En el presente estudio de simulación, la trayectoria lineal de la variable latente depresión será explicada por variables como actividad física, sexo, edad y una interacción de estas últimas variables.

Proceso latente subyacente continuo: Depresión, Λ_{ij}.

Variable maniftesta respuesta ordinal: Test de sintomátología depresiva. Con fines comparativos se presenta simulaciones con variables manifiestas de distinto número de niveles, tomando como referencia las 2 versiones más utilizadas del CESD: CESD-10 y CESD-20 (ver Cheng y Chan, 2005). La primera versión, que denotaremos por Y_{1ij}, considera 10 preguntas o items que tiene como posible rango de puntuación o niveles a [0,30] y la segunda versión, que denotaremos por Y_{2ij}, considera 20 preguntas o items que tiene como posible rango de puntuación o niveles a [0,60]. Así, Y_{1ij} e Y_{2ij} tienen los siguientes niveles:

$$Y_{1ij} = \begin{cases} 0, \text{ ningún síntoma de depresión,} \\ \vdots \\ 29, \text{ nivel muy alto de depresión} \end{cases} \qquad Y_{2ij} = \begin{cases} 0, \text{ ningún síntoma de depresión,} \\ \vdots \\ 59, \text{ nivel muy alto de depresión} \end{cases}$$

Covariables:

- $age65 = (age - 65)/10$: variable que se obtiene al centrar la edad del sujeto al momento de la medición alrededor de 65 y reducir la escala.
- sex: variable binaria que toma el valor de 1 cuando el sujeto es hombre y 0 cuando es mujer.
- AF : variable binaria que toma el valor de 1 cuando el sujeto realiza actividad f´ısica regularmente y 0 caso contrario.

4.2. Modelo a simular

Para el proceso de simulación se utilizará el siguiente modelo que está compuesto por 3 sub-modelos.

- **El sub-modelo de regresión logística** se reduce a una probabilidad de pertenencia a una de las 2 clases latentes consideradas pues no se incluyen covariables.

$$\pi_{ig} = P(C_i = g) = \frac{\exp(\xi_{0g})}{1 + \exp(\xi_{01})}, \quad g = 1, 2$$

  El modelo establece una restricción de identificabilidad para este sub-modelo mediante $\xi_{02} = 0$. Por lo tanto, en este sub-modelo solamente se estimará el intercepto de la clase 1, ξ_{01}.

- **El sub-modelo estructural**

$$\Lambda_{ij}|(C_i = g) = \beta_1\, sex + \beta_2\, sex \times age65 + \beta_3\, AF + v_{0g} + v_{1g}\, age65 + u_{0i} + u_{1i}\, age65$$

Para esta especificación el paquete *lcmm* establece una restricción de identificabilidad consistente en una restricción respecto a la locación del proceso latente que implica que el intercepto de la clase 1 es cero, $v_{01} = 0$.

Los efectos aleatorios $\mathbf{u}_i = (u_{0i}, u_{1i})$ se distribuyen de la siguiente manera:

$$\mathbf{u}_{ig} \sim N_2(0, D),$$

siendo

$$D = \begin{pmatrix} varcov1 & varcov2 \\ varcov2 & varcov3 \end{pmatrix}.$$

En este sub-modelo se estimarán los siguientes parámetros:

– Parámetros de efectos fijos comunes a toda la población: β_1, β_2 y β_3

– Parámetros de efectos fijos clase específico: v_{11}, v_{02} y v_{12}

– Parámetros de la matriz de varianza-covarianza de los efectos aleatorios *varcov*1, *varcov*2 y *varcov*3.

- **El sub-modelo de medición** viene dado por un modelo probit acumulado que tiene la siguiente estructura cuando se considera tanto una variable manifiesta de respuesta ordinal con 30 como una con 60 niveles:

$$Y_{1ij} = H(\tilde{Y}_{1ij}; \boldsymbol{\eta}) = \begin{cases} 0, & \text{si } \tilde{Y}_{ij} \leq \eta_1^* = \eta_1, \\ 1, & \text{si } \eta_1^* < \tilde{Y}_{ij} \leq \eta_2^*, \\ \cdot & \\ 29, & \text{si } \tilde{Y}_{ij} > \eta_{29}^*. \end{cases}$$

$$Y_{2ij} = H(\tilde{Y}_{2ij}; \boldsymbol{\eta}) = \begin{cases} 0, & \text{si } \tilde{Y}_{ij} \leq \eta_1^* = \eta_1, \\ 1, & \text{si } \eta_1^* < \tilde{Y}_{ij} \leq \eta_2^*, \\ \cdot & \\ 59, & \text{si } \tilde{Y}_{ij} > \eta_{59}^*. \end{cases}$$

donde $\tilde{Y}_{ij} = \Lambda_{ij} + s_{ij}$, $s_{i.} = \begin{pmatrix} s_{i1} \\ s_{i1} \\ \cdot \\ s_{in_i} \end{pmatrix} \sim N_{n_i}(0, I)$, y $\boldsymbol{\eta}$ es un vector de parámetros de umbrales.

Por lo tanto, en este sub-modelo se estimarán los parámetros que determinan los umbrales, los cuales son comunes para todas las clases.

En resumen, de acuerdo al modelo descrito para el CESD con 30 niveles (Y_{1ij}), se procederá a estimar y recuperar los siguientes 39 parámetros:

$$\xi_{01}, \beta_1, \beta_2, \beta_3, v_{11}, v_{02}, v_{12}, varcov1, varcov2, varcov3,$$

$$\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7, \eta_8, \cdots \text{y } \eta_{29}.$$

De igual manera, para el modelo que emula al CESD con 60 niveles (Y_{2ij}), se procederá a estimar y recuperar los siguientes 69 parámetros:

$$\xi_{01},\ \beta_1,\ \beta_2,\ \beta_3,\ v_{11},\ v_{02},\ v_{12},\ varcov1,\ varcov2,\ varcov3,$$

$$\eta_1,\ \eta_2,\ \eta_3,\ \eta_4,\ \eta_5,\ \eta_6,\ \eta_7,\ \eta_8,\ \cdots \text{ y } \eta_{59}.$$

4.3. Estructura del proceso de simulación

1. Se establece el número de clases latentes: 2 subpoblaciones.
2. Se asignan valores reales a los parámetros que servirán de insumo para construir los sub-modelos y la base de datos simulada. En particular, se asumen grandes varianzas para el intercepto y pendiente aleatoria, con el fin de simular trayectorias individuales variadas.
3. Se simulan los valores de la variable respuesta latente Λ_{ij} para cada clase y con ello los de su variable latente intermedia $\tilde{Y}_{ij}|C_i = g$, siendo ambos continuos.
4. Se simulan las variables manifiestas de respuesta ordinal: $Y_{1ij}|C_i = g$ e $Y_{2ij}|C_i = g$ para cada una de las clases.
5. Se estiman los parámetros del modelo lineal mixto de clases latentes generados usando la función *lcmm*. Siendo más específicos, la variable manifiesta de respuesta ordinal y las covariables permitirán aproximar el proceso latente continuo y así recuperar los valores reales asignados a los parámetros.
6. Debido a que el tiempo aproximado de simulación para modelos que contienen enlaces discretos, efectos aleatorios y clases latentes es considerable, se seguirá la recomendación de Proust-Lima et al. (2017) de estimar un modelo base para una población homogénea (con una sola clase), luego mediante la función *gridsearch* se utilizará 30 vectores aleatorios de valores iniciales de los parámetros y un máximo de 15 iteraciones para escoger la mejor verosimilitud, y finalmente, se estima el modelo propuesto para la simulación con los valores iniciales seleccionados.

4.4. Resultados

En esta sección se evaluará si la distribución de los sujetos en las 2 clases latentes se corresponde con la distribución promedio obtenida en las simulaciones. Se calcularon los valores promedio de la distribución porcentual de los sujetos en las 2 clases y se obtuvieron los resultados resumidos en el Cuadro 4.2 considerando a $Y_{1ij}|C_i = g$ con 30 niveles y a $Y_{2ij}|C_i = g$ con 60 niveles.

Como se puede observar nuestro modelo lineal mixto de clases latente, distribuye a los sujetos en porcentajes muy similares a los establecidos y la precisión en la asignación de cada sujeto en su clase va mejorando consistentemente a medida que se aumentan el número de niveles de la variable manifiesta. El modelo al dividir la población se equivocaría en la clasificación de aproximadamente 1.813 % de los sujetos cuando se tiene una variable

Clases	Datos	Promedio de simulación de $Y_{1ij} \mid C_i = g$ con 30 niveles	Promedio de simulación de $Y_{2ij} \mid C_i = g$ con 60 niveles
1	64.3%	62.487%	64.324 %
2	35.7%	37.513%	35.676 %

Cuadro 4.2: Distribución de sujetos por clase latente con variable manifiesta de distintos niveles.

manifiesta con 30 niveles y este porcentaje se reduce a 0.02415 % con una variable manifiesta de 60 niveles.

Cabe mencionar que en el proceso de simulación, un porcentaje de las simulaciones presentó una alteración en el orden de las clases, *label switching*[1]. Como resultado dos parámetros eran estimados con signo cambiado (intercepto del sub-modelo de regresión logística y el intercepto de la clase 2) y otros dos parámetros veían alterado su orden (el efecto fijo de la clase 1 aparec´ıa como efecto fijo de la clase 2 y viceversa). De manera que se realizaron las simulaciones con el orden correcto de las clases y signos adecuados de los parámetros.

Continuando con la evaluación del modelo, se calculan los promedios de los valores de los parámetros simulados, los sesgos, los sesgos procentuales y el RMSE. Los resultados obtenidos se muestran en los Cuadros 4.3 y 4.4.

Es preciso señalar que este proceso de simulación involucra un recorrido desde un proceso continuo a una variable ordinal para luego desde esta variable ordinal aproximar el proceso continuo latente inicial. Toda la información contenida en el proceso continuo es resumida en una variable ordinal con algunos niveles, es decir, se está produciendo una pérdida de información. Con esta información reducida se busca reproducir el proceso latente, lo cual representa una limitación para recuperar los valores reales de los parámetros en el proceso de simulación. Con el objetivo de evidenciar y minimizar esta pérdida de información, se procedió a simular la variable manifiesta de respuesta ordinal en distintos niveles. Se observa que a medida que se aumenta los niveles de la variable manifiesta, se obtiene estimaciones cada vez más cercanas a sus verdaderos valores. Los sesgos procentuales muestran claramente las mejoras obtenidas.

En particular, las estimaciones de los parámetros del sub-modelo estructural y sub-modelo de medición presentaron los sesgos porcentuales más reducidos. En contraste con los parámetro que componen la matriz de covarianza de los efectos aleatorios que resultaron los más disímiles de manera sostenida. Asi mismo, el cálculo del RSME para cada uno de los parámetros indicaron también una mejora en la precisión de la estimación cuando se aumenta el número de niveles de la variable manifiesta.

[1] El *label switching* como se discutió directamente con la autora del paquete *lcmm*, Cécile Proust-Lima, no es en síun problema, ya que el paquete suele intercambiar las clases cuando se estima sobre un mismo conjunto de datos. Esto puede corregirse internamente al observar los estadísticos resumen e intercambiar los parámetros específicos a cada clase cuando ocurra el cambio. Una forma de hacerlo es considerar sistemáticamente el intercepto o pendiente mas pequeño o la menor probabilidad en la primera clase.

Parámetros	Valor	Promedio real	Sesgo	Sesgo Porcentual	RSME
ξ_{01}	0.65	0.668	0.018	2.809 %	0.005
β_1	-3.93	-3.837	0.093	2.362 %	0.018
β_2	0.19	0.186	-0.004	1.985 %	0.001
β_3	-3.61	-3.593	0.017	0.466 %	0.015
v_{11}	1.23	1.274	0.044	3.589 %	0.006
v_{02}	3.78	3.635	-0.145	3.847 %	0.032
v_{12}	2.75	2.616	-0.134	4.877 %	0.031
*varcov*1	128.84	118.595	-10.245	7.952 %	8.032
*varcov*2	-63.35	-59.296	4.054	6.399 %	3.533
*varcov*3	54.89	48.479	-6.410	11.678 %	5.582
η_1	0.52	0.541	0.021	4.063 %	0.005
η_2	0.95	0.934	-0.016	1.676 %	0.006
η_3	0.77	0.747	-0.023	2.958 %	0.006
η_4	0.50	0.505	0.005	0.916 %	0.004
η_5	0.69	0.680	-0.009	1.439 %	0.006
η_6	0.63	0.636	0.006	0.941 %	0.005
η_7	0.43	0.422	-0.008	1.879 %	0.003
η_8	0.57	0.580	0.010	1.789 %	0.004
η_9	0.56	0.552	-0.008	1.434 %	0.005
η_{10}	0.55	0.560	0.010	1.894 %	0.004
η_{11}	0.36	0.356	-0.005	1.259 %	0.003
η_{12}	0.52	0.507	-0.013	2.453 %	0.003
η_{13}	0.52	0.509	-0.010	1.947 %	0.003
η_{14}	0.34	0.345	0.005	1.480 %	0.002
η_{15}	0.56	0.570	0.010	1.809 %	0.004
η_{16}	0.57	0.563	-0.007	1.225 %	0.003
η_{17}	0.37	0.380	0.010	2.709 %	0.003
η_{18}	0.63	0.616	-0.014	2.224 %	0.004
η_{19}	0.45	0.460	0.010	2.254 %	0.004
η_{20}	0.56	0.571	0.011	1.969 %	0.005
η_{21}	0.35	0.342	-0.008	2.172 %	0.002
η_{22}	0.67	0.683	0.013	1.954 %	0.005
η_{23}	0.54	0.533	-0.007	1.302 %	0.004
η_{24}	0.51	0.499	-0.011	2.228 %	0.005
η_{25}	0.50	0.492	-0.008	1.654 %	0.004
η_{26}	0.60	0.596	-0.004	0.667 %	0.005
η_{27}	0.38	0.384	0.004	1.104 %	0.003
η_{28}	0.65	0.647	-0.003	0.434 %	0.003
η_{29}	0.48	0.472	-0.008	1.650 %	0.004

Cuadro 4.3: Indicadores del desempeño del Modelo Lineal Mixto de Clases Latentes con variable respuesta latente aproximada por una variable manifiesta de respuesta ordinal con 30 niveles.

Parámetros	Valor	Promedio real	Sesgo	Sesgo Porcentual	RSME
ξ_{01}	0.65	0.6541	0.0041	0.6294 %	0.0042
β_1	-3.93	-3.9517	-0.0022	0.5529 %	0.0054
β_2	0.19	0.1915	0.0015	0.7965 %	0.0008
β_3	-3.61	-3.6090	0.0010	0.0286 %	0.0053
ν_{11}	1.23	1.2406	0.0106	0.8594 %	0.0043
ν_{02}	3.78	3.8032	0.0232	0.6127 %	0.0100
ν_{12}	2.75	2.7285	-0.0215	0.7818 %	0.0307
$varcov1$	128.84	124.3075	-4.5325	3.5179 %	5.8299
$varcov2$	-63.35	-66.0750	-2.7250	4.3015 %	2.1357
$varcov3$	54.89	52.7141	-2.1759	3.9641 %	3.7779
η_1	0.21	0.2106	0.0006	0.2905 %	0.0019
η_2	0.47	0.4737	0.0037	0.7894 %	0.0032
η_3	0.39	0.3907	0.0007	0.1821 %	0.0053
η_4	0.56	0.5596	-0.0004	0.0714 %	0.0011
η_5	0.27	0.2718	0.0018	0.6741 %	0.0009
η_6	0.68	0.6790	-0.0010	0.1471 %	0.0037
η_7	0.29	0.2930	0.0030	1.0434 %	0.0022
η_8	0.37	0.3679	-0.0021	0.5719 %	0.0043
η_9	0.55	0.5490	-0.0010	0.1816 %	0.0039
η_{10}	0.44	0.4415	0.0015	0.3432 %	0.0040
η_{11}	0.22	0.2211	0.0011	0.5173 %	0.0024
η_{12}	0.45	0.4524	0.0024	0.5333 %	0.0015
η_{13}	0.63	0.6326	0.0026	0.4127 %	0.0008
η_{14}	0.18	0.1796	-0.0004	0.2278 %	0.0016
η_{15}	0.34	0.3407	0.0007	0.2177 %	0.0021
η_{16}	0.41	0.4084	-0.0016	0.3978 %	0.0007
η_{17}	0.67	0.6721	0.0021	0.3149 %	0.0020
η_{18}	0.51	0.5137	0.0037	0.7255 %	0.0012
η_{19}	0.92	0.9189	-0.0011	0.1196 %	0.0032
η_{20}	0.87	0.8762	0.0062	0.7083 %	0.0037
η_{21}	0.33	0.3288	-0.0012	0.3667 %	0.0004
η_{22}	0.61	0.6122	0.0022	0.3530 %	0.0015
η_{23}	0.28	0.2789	-0.0011	0.3803 %	0.0014
η_{24}	0.43	0.4311	0.0011	0.2512 %	0.0006
η_{25}	1.20	0.1987	-0.0013	0.1121 %	0.0038
η_{26}	0.58	0.5831	0.0031	0.5362 %	0.0006
η_{27}	0.76	0.7649	0.0049	0.6422 %	0.0006
η_{28}	0.90	0.9052	0.0052	0.5778 %	0.0011
η_{29}	0.77	0.7668	-0.0032	0.4143 %	0.0023
η_{30}	0.30	0.3014	0.0014	0.4667 %	0.0013
η_{31}	0.63	0.6341	0.0041	0.6437 %	0.0039
η_{32}	0.71	0.7132	0.0032	0.4507 %	0.0036
η_{33}	0.69	0.6873	-0.0027	0.3928 %	0.0015
η_{34}	0.57	0.5714	0.0014	0.2456 %	0.0016
η_{35}	0.74	0.7364	-0.0036	0.4851 %	0.0016
η_{36}	0.64	0.6435	0.0035	0.5472 %	0.0015
η_{37}	1.01	1.0089	-0.0011	0.1089 %	0.0004
η_{38}	0.80	0.7948	-0.0052	0.6550 %	0.0031
η_{39}	0.93	0.9280	-0.0020	0.2108 %	0.0030
η_{40}	0.86	0.8658	0.0058	0.6745 %	0.0026
η_{41}	1.31	1.3078	-0.0022	0.1657 %	0.0005
η_{42}	1.11	1.1099	-0.0001	0.0090 %	0.0028
η_{43}	0.82	0.8227	0.0027	0.3254 %	0.0006
η_{44}	0.78	0.7768	-0.0032	0.4064 %	0.0001
η_{45}	0.44	0.4406	0.0006	0.1391 %	0.0022
η_{46}	0.53	0.5307	0.0007	0.1359 %	0.0013
η_{47}	0.35	0.3496	-0.0004	0.1057 %	0.0004
η_{48}	0.38	0.3802	0.0002	0.0553 %	0.0003
η_{49}	0.62	0.6202	0.0002	0.0348 %	0.0009
η_{50}	0.49	0.4910	0.0010	0.2068 %	0.0021
η_{51}	0.63	0.6290	-0.0010	0.1621 %	0.0007
η_{52}	0.39	0.3886	-0.0014	0.3526 %	0.0014
η_{53}	0.71	0.7115	0.0015	0.2169 %	0.0007
η_{54}	0.59	0.5931	0.0031	0.5254 %	0.0023
η_{55}	0.94	0.9428	0.0028	0.2931 %	0.0041
η_{56}	0.60	0.5988	-0.0012	0.1932 %	0.0021
η_{57}	0.36	0.3618	0.0018	0.5006 %	0.0014
η_{58}	0.87	0.8728	0.0028	0.3219 %	0.0006
η_{59}	0.39	0.3880	-0.0020	0.5154 %	0.0004

Cuadro 4.4: Indicadores del desempeño del Modelo Lineal Mixto de Clases Latentes con variable respuesta latente aproximada por una variable manifiesta de respuesta ordinal con 60 niveles.

Capítulo 5

Aplicación en la medición de religiosidad

En este capítulo se presenta una aplicación del modelo lineal mixto de clases latentes con variable manifiesta de respuesta ordinal a un conjunto de datos que forma parte del Estudio Nacional de Juventud y Religión (NSYR por sus siglas en inglés "National Study of Youth and Religion"). Este estudio longitudinal fue diseñado para estudiar el comportamiento religioso y espiritual de jóvenes americanos desde su temprana adolescencia hasta su adultez.

Se analizará tanto el aspecto dinámico de la religiosidad a través de un proceso longitudinal como la naturaleza multidimensional de la religiosidad. El proceso longitudinal considera como variable respuesta latente a la religiosidad aproximada a través del índice, Irelig, compuesto por 4 escalas autoreportadas. Este índice está basado en el índice DUREL (por sus siglas en inglés "Duke University Religion Index") que fue diseñado para evaluar las tres principales dimensiones de la religiosidad: cognitiva, conductual y afectiva, en estudios longitudinales. Conforme a la sociología de la religión, la dimensión cognitiva refleja las creencias y visión que se tiene de Dios, la dimensión conductual se divide en la práctica en el ámbito público y en el privado, y la dimensión afectiva explica la conexión entre la vida del sujeto y su religiosidad. De ah´ıque nuestro ´ındice (Irelig) recoge preguntas sobre asistencia al servicio religioso, frecuencia de la oración, relación con Dios e importancia personal de la religión en la toma de decisiones.

5.1. Descripción de los datos

La NSYR [1] es un encuesta teléfonica longitudinal que se tomó en 4 oportunidades durante 10 años a un conjunto de adolescentes americanos en su tránsito a la adultez. Esta encuesta fue diseñada para tener representatividad a nivel nacional de los hogares en los que habiten por lo menos un o una adolescente de habla inglesa o española entre 13 y 17 años para los años 2002 y 2003 en los Estados Unidos.

La técnica de muestreo empleada fue el método de marcación al azar (RDD, por sus siglas en inglés "Random Digit Dialing"), en la cual, el marco muestral está compuesto por todos los números telefónicos posibles y la unidad de muestreo es el número telefónico que está vinculada a un hogar y sus miembros. Se debe agregar que el RDD incluye un muestreo estratificado con 50 estratos correspondientes a los 50 estados.

Además de los números seleccionados por el método RDD, se emplea un sobremuestreo

[1] La base de datos completa de la NSYR fue obtenida gracias a la autorización brindada para la presente tesis por las investigadoras Lisa D. Pearce y Sara Skiles del Center for the Study of Religion & Society de la Universidad de Notre Dame.

de 80 hogares judíos a través de un muestreo aleatorio simple cuyo marco muestral fue un listado de números telefónicos pertenencientes a sujetos con apellidos judíos.

En la primera medición de la NSYR, llevada a cabo entre julio del 2002 y abril del 2003, se entrevistó tanto al adolescente elegido como a un padre/madre o responsable que cohabite con el adolescente. En las siguientes 3 mediciones se intentó volver a contactar y encuestar a todos los adolescentes considerados en la medición 1. A pesar de la dificultad que involucra contactar con participantes que están pasando de la adolescencia a la adultez, se logró una alta tasa de retención, siendo 75 % aproximadamente entre el corte 1 y el corte 4.

Los datos inclu´ıdos en la presente tesis corresponden a un subconjunto de la NSYR, el criterio de inclusión consistió en considerar sólo a los sujetos que hayan respondido a todas las preguntas que miden religiosidad. En el Cuadro 5.1 se presenta la estructura de la base de datos a utilizar.

Corte	Años	Edades	Número de sujetos
1	2002 - 2003	13 - 17	1852
2	2005	16 - 20	1869
3	2007 - 2008	18 - 23	1726
4	2012 -2013	23 - 28	1633

Cuadro 5.1: Estructura de la base de datos de la aplicación.

Con la información de cada una de las mediciones se construyó una nueva base de datos en formato longitudinal y se pudo obtener la Figura 5.1 que muestra las trayectorias de Irelig por sujeto sin distinguir subpoblaciones o clases, de acuerdo a su edad en el momento de la medición.

A continuación, se presenta el conjunto de variables originales y escalas que se utilizaran en la aplicación:

- *Irelig*: Índice compuesto de 4 items o escalas diseñadas para medir religiosidad a través de tres dimensiones (cognitiva, conductual y afectiva), y que puede tomar valores desde 4 a 26.

$$Irelig_{ij} = \begin{cases} 4, \text{ Extremadamente baja religiosidad,} \\ 5, \text{ Muy baja religiosidad,} \\ \vdots \\ 25, \text{ Muy alta religiosidad,} \\ 26, \text{ Extremadamente alta religiosidad.} \end{cases}$$

Las 4 escalas que componen aditivamente el ´ındice *Irelig* son las siguientes:

1. *ASIST*: Escala de 7 niveles que corresponde a la pregunta de cuán frecuentemente Ud. asiste al servicio religioso. Esta medida corresponde a la dimensión conductual y especificamente al ámbito público.
2. *REZAR*: Escala de 7 niveles que corresponde a la pregunta de cuán frecuentemente Ud. reza solo. Esta medida corresponde a la dimensión conductual y especificamente al ámbito privado.

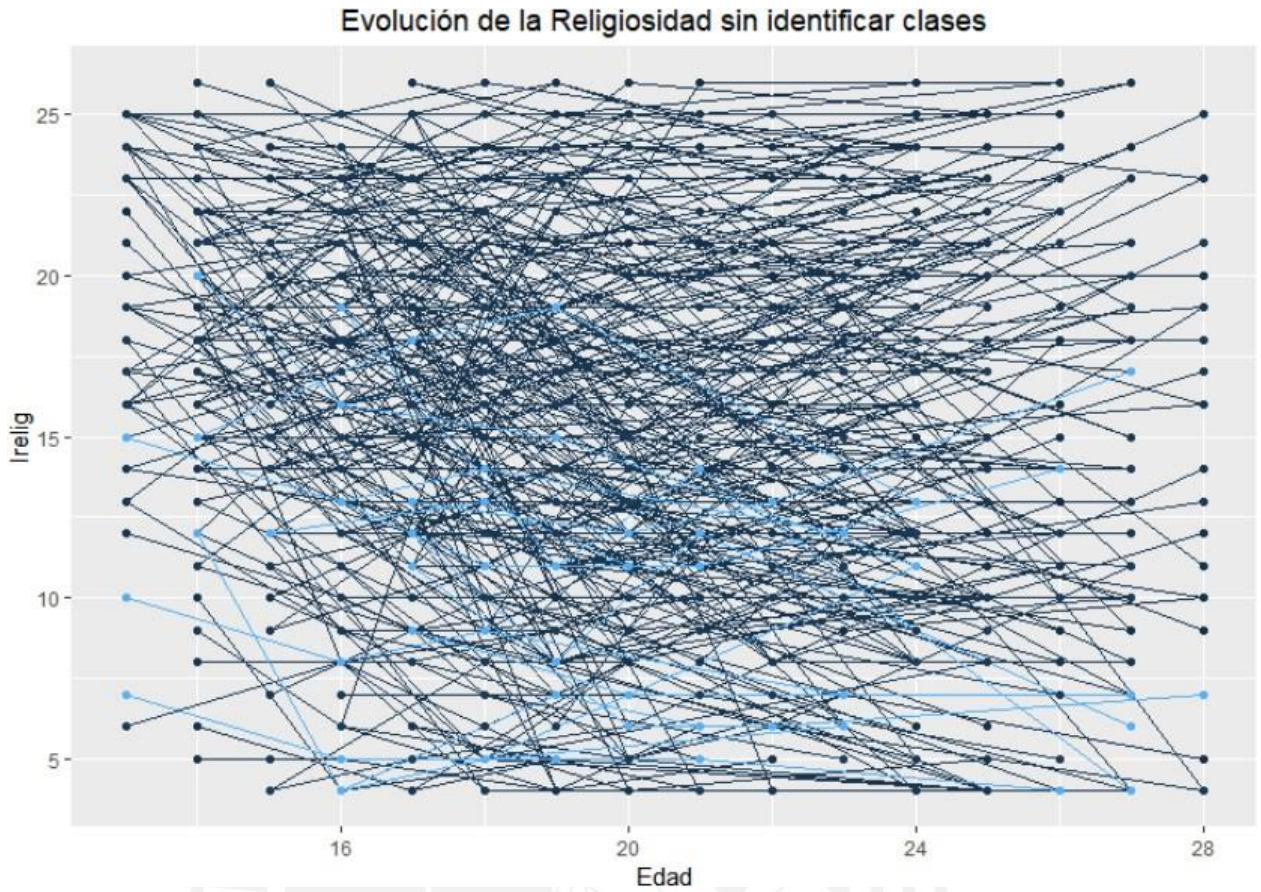

Figura 5.1: Trayectorias de Irelig.

3. *IMPORT* : Escala de 5 niveles que corresponde a la pregunta sobre si Ud. considera importante la religión en la forma cómo vive. Esta medida corresponde a la dimensión afectiva.
4. *CREER*: Escala de 7 niveles que corresponde a la pregunta cuán distante o cercano Ud. se siente de Dios la mayor parte del tiempo. Esta medida corresponde a la dimensión cognitiva.

Estas 4 escalas de religiosidad están ordenadas de menor a mayor grado en escalas de tipo Likert y la distribución de frecuencias de las alternativas por corte para cada escala son mostradas en el Cuadro 5.2.

De otro lado, las covariables de interés en nuestro modelo estarán constituidas por:

- *edad*: variable que indica la edad en años del sujeto al momento de la medición.
- *sex*: variable binaria que toma el valor de 0 cuando el sujeto es hombre y 1 cuando es mujer.
- *religiouspeer*: variable binaria que toma el valor de 0 cuando el sujeto no participa en actividades sociales y recreativas organizadas por grupos religiosos y 1 en caso

contrario. Algunos autores como Hoffmann (2014) consideran esta variable como una aproximación de tener amigos que pertenecen a alguna comunidad religiosa.

- *afilia*: variable binaria que toma el valor de 0 cuando el sujeto no está afiliado a alguna manifestación religiosa y 1 en caso contrario.
- *marihuana*: variable binaria que toma el valor de 0 cuando el sujeto no consume marihuana y 1 en caso contrario.
- *raza*: variable categórica que indica la raza en la que se identifica el sujeto. Toma el valor de 1 cuando el sujeto se identifica como de raza blanca, toma el valor de 2 cuando el sujeto de identifica como afroamericano, toma el valor de 3 para aquellos que se identifican como hispanos y el valor de 4 corresponde a los sujetos que pertenecen a las minorías raciales (asiáticos, isleños, nativos americanos y otros).
- *hijos*: variable binaria que toma el valor de 0 cuando el sujeto no tiene hijos en el corte 4 y 1 en caso contrario.
- *maritalStatus*: variable binaria que toma el valor de 0 si hasta el corte 4 el sujeto nunca ha estado casado y 1 en caso contrario.
- *educDAD*: variable binaria que toma el valor de 0 si en el corte 1 la figura paterna no ha alcanzado estudios universitarios y 1 en caso contrario.
- *educMOM* : variable binaria que toma el valor de 0 si en el corte 1 la figura materna no ha alcanzado estudios universitarios y 1 en caso contrario.
- *educSubject*: variable binaria que toma el valor de 0 si en el corte 4 el sujeto no ha alcanzado estudios universitarios y 1 en caso contrario.
- *parentsIMPORT* : Escala de 6 niveles que corresponde a la pregunta realizada en el corte 1 al padre o madre o tutor sobre si considera importante la religión en la forma como vive.
- *parentsASIST*: Escala de 7 niveles que corresponde a la pregunta realizada en el corte 1 al padre o madre o tutor sobre cuán frecuentemente asiste al servicio religioso.

5.2. Modelo para la aplicación

Las variables anteriormente reseñadas serán utilizadas para explicar el constructo religiosidad a través de un modelo lineal mixto de clases latentes. A continuación se detalla los componentes del modelo:

- **El sub-modelo de regresión logística multinomial** Se utilizará para predecir la probabilidades de pertenencia de un sujeto i a una de las clases latentes g, π_{ig}, dado un conjunto de covariables independientes del tiempo, $\tilde{\mathbf{X}}_{C_i}$.

$$\pi_{ig} = P(C_i = g|\tilde{\mathbf{X}}_{C_i}) = \frac{\exp(\xi_{0g} + \tilde{\mathbf{X}}_{C_i}^T \boldsymbol{\xi}_{1g})}{\sum_{l=1}^{tt} \exp(\xi_{0l} + \tilde{\mathbf{X}}_{C_i}^T \boldsymbol{\xi}_{1l})}, \quad g = 1, 2, \ldots, tt, \tag{5.1}$$

Variable	Nivel	Corte 1	Corte 2	Corte 3	Corte 4
¿Cuán frecuentemente Ud. asiste al servicio religioso?					
	1 - No asiste	320 (17.28 %)	522 (27.93 %)	633 (36.67 %)	768 (47.03 %)
	2 - Pocas veces al año	280 (15.12 %)	322 (17.23 %)	323 (18.71 %)	191 (11.70 %)
	3 - Muchas veces al año	143 (7.72 %)	88 (4.71 %)	103 (5.97 %)	119 (7.29 %)
ASIST	4 - 1 vez al mes	131 (7.07 %)	140 (7.49 %)	124 (7.18 %)	86 (5.27 %)
	5 - 2 a 3 veces al mes	224 (12.10 %)	246 (13.16 %)	183 (10.60 %)	135 (8.27 %)
	6 - 1 vez a la semana	453 (24.46 %)	344 (18.41 %)	232 (13.44 %)	222 (13.59 %)
	7 - más de 1 vez a la semana	301 (16.25 %)	207 (11.08 %)	128 (7.42 %)	112 (6.86 %)
¿Cuán frecuentemente Ud. reza solo?					
	1 - Nunca	268 (14.47 %)	326 (17.44 %)	377 (21.84 %)	297 (18.19 %)
	2 - Menos de 1 vez al mes	154 (8.32 %)	191 (10.22 %)	194 (11.24 %)	219 (13.41 %)
	3 - 1 a 2 veces al mes	243 (13.12 %)	313 (16.75 %)	250 (14.48 %)	164 (10.04 %)
REZAR	4 - Aprox. 1 vez a la semana	214 (11.56 %)	220 (11.77 %)	184 (10.66 %)	110 (6.74 %)
	5 - Algunas veces a la semana	294 (15.87 %)	271 (14.50 %)	214 (12.40 %)	262 (16.04 %)
	6 - Aprox. 1 vez al día	414 (22.35 %)	342 (18.30 %)	268 (15.53 %)	271 (16.60 %)
	7 - varias veces al día	265 (14.31 %)	206 (11.02 %)	239 (13.85 %)	310 (18.98 %)
¿Cuán importante o no importante es la fe religiosa en como Ud. vive su vida?					
	1 - Nada importante	131 (7.07 %)	222 (11.88 %)	239 (13.85 %)	228 (13.96 %)
	2 - No muy importante	229 (12.37 %)	284 (15.20 %)	278 (16.11 %)	166 (10.17 %)
IMPORT	3 - Algo importante	589 (31.80 %)	551 (29.48 %)	508 (29.43 %)	565 (34.60 %)
	4 - Muy importante	537 (29.00 %)	480 (25.68 %)	377 (21.84 %)	321 (19.66 %)
	5 - Extremadamente importante	366 (19.76 %)	332 (17.76 %)	324 (18.77 %)	353 (21.62 %)
¿Cuán distante o cercano Ud. se siente de Dios?					
	1 - No creo en Dios	58 (3.13 %)	94 (5.03 %)	115 (6.66 %)	167 (10.23 %)
	2 - Extremadamente distante	63 (3.40 %)	76 (4.07 %)	92 (5.33 %)	65 (3.98 %)
	3 - Muy distante	85 (4.59 %)	138 (7.38 %)	147 (8.52 %)	102 (6.25 %)
CREER	4 - Algo distante	333 (17.98 %)	411 (21.99 %)	377 (21.84 %)	257 (15.74 %)
	5 - Algo cercano	635 (34.29 %)	646 (34.56 %)	544 (31.52 %)	496 (30.37 %)
	6 - Muy cercano	476 (25.70 %)	374 (20.01 %)	338 (19.58 %)	380 (23.27 %)
	7 - Extremadamente cercano	202 (10.91 %)	130 (6.96 %)	113 (6.55 %)	166 (10.17 %)

Cuadro 5.2: Distribución de frecuencias de las escalas del índice de religiosidad.

donde el vector $\tilde{\mathbf{X}}_{C_i}$ está compuesto por las siguientes covariables: raza, nivel educativo del sujeto y de sus padres, estado marital, hijos y sus correspondientes variables dummy.

- **El sub-modelo estructural** El proceso latente religiosidad se explica para cada sujeto i por un conjunto de variables asociadas a efectos fijos comunes a toda la población (sexo, edad, religiosidad de los padres, pares con v´ınculo religioso y consumo de marihuana), a un conjunto de variable con efectos fijos específicos por clase (edad, pares con vínculo religioso y afiliación) y también se incluye un intercepto aleatorio.

$$\begin{aligned}\Lambda_{ij}|(C_i = g) = {} & \beta_1\, sex_i + \beta_2\, edad_{ij} + \beta_3\, parentsIMPORT_i + \beta_4\, parentsASIST_i + \\ & \beta_5\, religiouspeer_{ij} + \beta_6\, marihuana_{ij} + \\ & v_{0g} + v_{1g}\, edad_{ij} + v_{2g}\, religiouspeer_{ij} + v_{3g}\, afilia_{ij} + \\ & u_{0ig},\end{aligned} \tag{5.2}$$

donde j denota el corte en el que se miden las variables y g al indicador de la clase latente.

Es preciso mencionar que los parámetros estimados de los efectos fijos clase específico para aquellas variables que también tienen efectos fijos poblacionales recogen ambos

efectos. Respecto a los efectos aleatorios, se ha incluido un intercepto aleatorio y también se ha considerado conveniente incluir un variabilidad específica por clase a través de ω_g. De ahí que el intercepto aleatorio tiene la siguiente distribución:

$$u_{0ig} \sim N(0, \omega_g^2 \sigma^2).$$

Habría que decir también que las covariables incluídas tanto en el modelo logístico multinomial como en el modelo estructural han sido sugeridas en la literatura para medir religiosidad Pearce y Foster (2013), Hoffmann (2014) y Pearce y Schorpp (2018).

- **El sub-modelo de medición**. Viene dado para el índice $Irelig$ para el sujeto i y corte j por el modelo Probit acumulado:

$$Y_{ij} = Irelig = H(\tilde{Y}_{ij}, \eta) = \begin{cases} 4, \text{si} & \tilde{Y}_{ij} < \eta_1^* = \eta_1, \\ \vdots & \\ 26, \text{si} & \tilde{Y}_{ij} > \eta_{22}^*, \end{cases}$$

donde $\tilde{Y}_{ij} = \Lambda_{ij} + s_{ij}$ es el proceso de religiosidad subyacente en la clase correspondiente, $s_{i.} \sim N_{n_i}(0, I_{n_i})$ y $\boldsymbol{\eta}$ es un vector que incluye a todos los parámetros de umbrales.

5.3. Resultados

Para el proceso de estimación se sigue la recomendación de Proust-Lima et al. (2017) de estimar en primer lugar un modelo base sobre una población homogénea y luego usar la función *gridsearch* para estimar un conjunto de valores iniciales que servirán de insumo para obtener la mejor verosimilitud. Esta secuencia de pasos disminuye el tiempo de estimación que es muy considerable con enlaces discretos y además mejora las chances de alcanzar un máximo global.

Con el propósito de determinar el número de clases latentes, se planificó estimar modelos considerando 1, 2, 3, 4, 5, 6 y 7 clases latentes. Sin embargo, el modelo especificado es capaz de distinguir 3 clases latentes en la población, aún cuando se le requiere distinguir más clases latentes por problemas de convergencia.

En el Cuadro 5.3 se presentan los valores de la logverosimilitud, AIC y el número de parámetros para cada uno de los modelos estimados que alcanzaron la convergencia. Para elegir el mejor modelo se considera el menor AIC y la mayor verosimilitud, de donde resulta que el mejor modelo identifica 3 clases latentes. Asi mismo, se detalla los porcentajes de sujetos por clase latente para cada uno de los modelos estimados.

Modelo	Nro de clases	Nro de parámetros	loglik	AIC	Porcentaje en cada clase		
					Clase 1	Clase 2	Clase3
Modelo 1	1	30	-17220.21	34500.43	100.00 %		
Modelo 2	2	44	-12918.12	25924.24	35.99%	64.01%	
Modelo 3	3	58	-12872.19	25860.39	18.66%	50.63 %	30.71 %

Cuadro 5.3: Comparación y selección del mejor modelo *lcmm*.

Luego de realizar la comparación de los modelos, se presenta los resultados de la estimación del modelo elegido identificando sus sub-modelos: logístico, estructural y de medición.

Parámetro	Clase	Coeficiente	ODDs	(ODDs-1)x100 %	D.E.	Test de Wald	p-value
Intercepto	1	-1.2475	0.2872	-71.28 %	190.22	-1.744	0.0812
Intercepto	2	0.1122	1.1187	11.87 %	0.3816	0.294	0.7687
Covariable	Clase	Coeficiente	OR	(OR-1)x100 %	D.E.	Test de Wald	p-value
Raza afroamericana	1	-5.9450	0.0026	-99.74 %	19.3368	-0.307	0.7585
Raza afroamericana	2	1.3396	3.8175	281.75 %	0.4626	2.896	0.0038
Raza hispana	1	-0.7242	0.4847	-51.53 %	0.9335	-0.776	0.4379
Raza hispana	2	0.1851	1.2033	20.33 %	0.4228	0.438	0.6616
Raza minorías	1	-0.5041	0.6040	-39.60 %	0.6190	-0.814	0.4154
Raza minorías	2	0.0578	1.0595	5.95 %	0.4818	0.120	0.9045
educDAD	1	0.6089	1.8384	83.84 %	0.4692	1.298	0.1944
educDAD	2	-0.1052	0.9001	-9.99 %	0.2726	-0.386	0.6995
educMOM	1	0.3460	1.4134	41.34 %	0.4516	0.766	0.4435
educMOM	2	-0.6152	0.5405	-45.95 %	0.2724	-2.259	0.0239
educSubject	1	0.5557	1.7432	74.32 %	0.3492	1.591	0.1116
educSubject	2	0.1436	1.1544	15.44 %	0.2711	0.530	0.5963
maritalStatus	1	-1.1531	0.3157	-68.43 %	0.4431	-2.602	0.0093
maritalStatus	2	0.6666	1.9477	94.77 %	0.2622	2.542	0.0110
hijos	1	-0.3796	0.6842	-31.58 %	0.7890	-0.481	0.6305
hijos	2	1.0655	2.9022	190.22 %	0.3436	3.101	0.0019

Cuadro 5.4: Estimadores del sub-modelo log´ıstico multinomial con variable manifiesta *Irelig*.

En los Cuadros 5.4 y 5.7 se muestran los resultados de la estimación del mejor modelo para el índice *Irelig*. A continuación se intentará describir las clases latentes encontradas a través de las covariables, en especial sobre las que resultan significativas en el test de Wald sobre el sub-modelo log´ıstico multinomial:

Clase latente 1 Está caracterizada por incluir con mayor probabilidad a sujetos solteros, desde que el cambio porcentual en el ratio de ODDs (CPRODDs) de pertenecer a la clase 1 en relación a la clase 3 para la variable *maritalStatus* es de -68.43 %. Aunque no significativo al 5 % esta clase se caracteriza también por tener una alta probabilidad de incluir a sujetos de raza blanca que no tienen hijos (CPRODDs = -31.58 %) y que ellos y sus padres tienen estudios superiores.

Clase latente 2 Está caracterizada por incluir con mayor probabilidad a sujetos afroamericanos casados con hijos y madres sin estudios universitarios. En efecto, los CPRODDs son 94.77 % para la variable *maritalStatus*, 190.22 % para la variable *hijos*, -45.95 % para la variable *educMOM* y 281.75 % más alto para raza afroamericana respecto al nivel de referencia, raza blanca. Aunque no significativo al 5 % esta clase se caracteriza también por incluir a sujetos de raza hispana y minorías, y marginalmente ellos mismos mas no su papá cuentan con educación superior.

Clase latente 3 Está caracterizada por incluir con mayor probabilidad a pocos sujetos afroamericanos y a mayor cantidad de sujetos de raza blanca, a sujetos sin hijos y con madre con estudios

universitarios. En efecto, por principio de transitividad, los CPRODDs hallados para las variables que resultaron significativas paras las clases 1 y 2 permiten deducir la descripción de esta clase. Aunque no significativo al 5 % esta clase se caracteriza también por incluir a sujetos con bajo nivel de estudios.

Con el propósito de verificar en la base de datos lo encontrado en la estimación de los parámetros asociados a las covariables que resultaron significativas para predecir la pertenencia a las clases, se presenta en los Cuadros 5.5 y 5.6. Haciendo los cruces entre las clases y las covariables, se confirma lo hallado por nuestro modelo. Como por ejemplo, que en la clase 1 está conformada por 94.72 % por sujetos de raza blanca y por más sujetos solteros que casados. En la clase 2 está el 87.61 % de los sujetos afroamericanos, el 74.66 % de los sujetos casados, el 86.59 % de los sujetos con hijos y el 70.16 % de los sujetos con madres sin estudios universitarios. Respecto a la clase 3 está conformada por 82.80.72 % por sujetos de raza blanca y sólo por 3.21 % de afroamericanos, también en está clase existe mayor cantidad de sujetos sin hijos que sujetos con hijos, y finalmente los sujetos con madres con estudios universitarios triplican a los sujetos con madres sin estudios universitarios. Se comprueban también las otras características auxiliares de las clases.

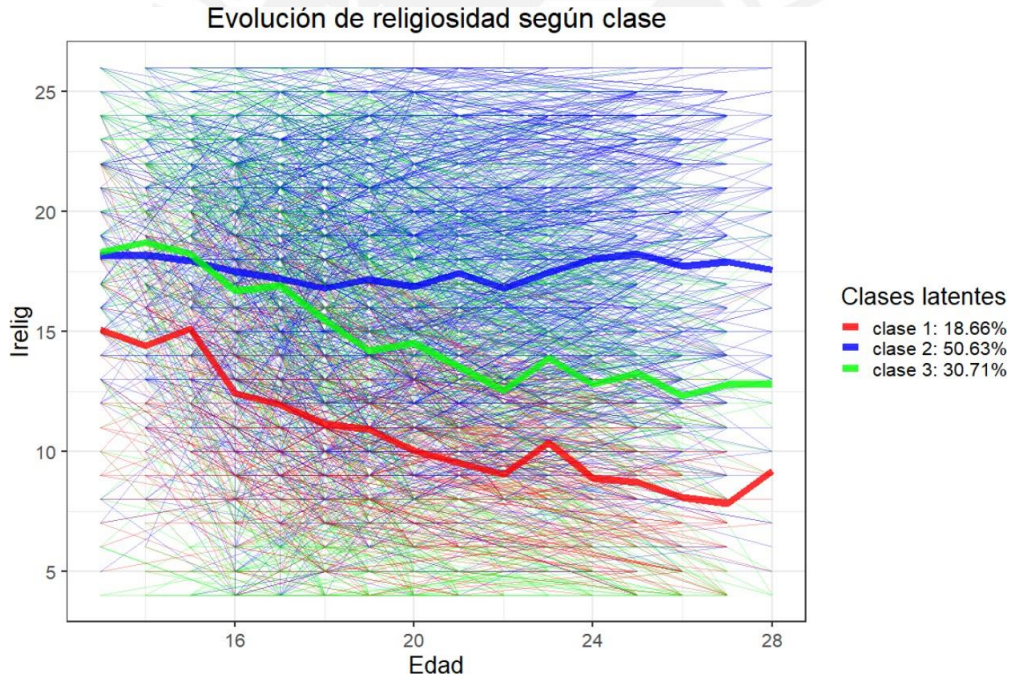

Figura 5.2: Trayectorias de religiosidad promedio según clase para *Irelig*.

En el Cuadro 5.7 se encuentran resumidos los resultados del sub-modelo estructural y sub-modelo de medición, y a continuación se presentan los principales hallazgos:

– Todas las covariables con efecto poblacional son significativas y en particular, las variables *sex* y *parentsIMPORT* resultan tener mayor efecto sobre los valores del ´ındice *Irelig*. La religiosidad se estima por ejemplo a ser 0.4203 desviaciones estándares mayor en mujeres que en hombres bajo la escala latente.

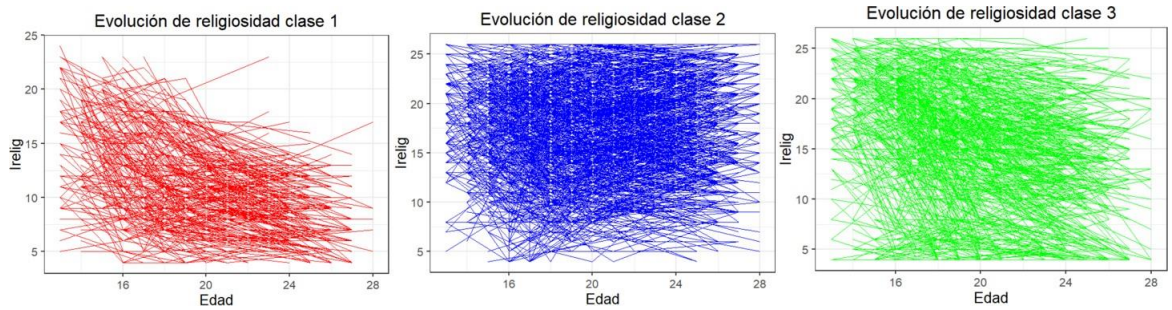

Figura 5.3: Trayectorias de religiosidad separadas por clase.

Clase	Raza				Total
	Blanca	Afroamericana	Hispana	Minor´ıas	
1	251 94.72 % 22.04 %	0 0.00 % 0.00 %	4 1.51 % 4.00 %	10 3.77 % 14.71 %	265 18.66 %
2	527 73.30 % 46.27 %	99 13.77 % 87.61 %	63 8.76 % 63.00 %	30 4.17 % 44.12 %	719 50.63 %
3	361 82.80 % 31.69 %	14 3.21 % 12.39 %	33 7.57 % 33.00 %	28 6.42 % 41.18 %	426 30.70 %
Total	1139	113	100	68	1420
	80.21 %	7.96 %	7.04 %	4.79 %	

Cuadro 5.5: Tabla de contingencia de raza y clase.

– Todas las variables con efecto clase fijo específico también resultan significativas. Se aprecia un efecto positivo (mayor religiosidad) según la afiliación religiosa y la confraternización con pares religiosos, al margen de la clase, aunque con un mayor efecto para la clase 3 y un menor efecto sobre la clase 1. En general, el efecto de la variable afiliación religiosa resulta ser el más importante para entender la religiosidad en las tres clases, superando largamente el impacto de las otras covariables. La segunda variable con mayor impacto es la religiosidad de su entorno social, sobrepasando la influencia de la religiosidad de los padres, recogida por *parentsIMPORT* y *parentsASIST* , lo que indica que para predecir la religiosidad resulta más importante la influencia de la religiosidad de los amigos que la de los padres durante el tránsito de la adolescencia a la adultez. Respecto al efecto de la edad sobre la religiosidad, se puede apreciar un efecto positivo pero muy bajo en la clase 2, lo cual indica que su trayectoria promedio de religiosidad es un poco más estable en el tiempo. Mientras en las clases 1 y 3 se observa trayectorias de religiosidad de cambio hacia una menor escala de religiosidad, siendo más pronunciado el declinamiento en la clase 1, como se puede apreciar tanto de manera conjunta en la Figura 5.2 como separadamente por clase en la Figura 5.3.

– Respecto al efecto aleatorio, la varianza estimada del intercepto aleatorio (1.1918) es levemente mayor respecto a la varianza del error de medición que se asume igual a

Variable	Nivel	Clase 1	Clase 2	Clase 3
educSubject	0 - Sin estudios Universitarios	70 (9.20 %)	454 (59.66 %)	237 (31.14 %)
educSubject	1 - Con estudios Universitarios	195 (29.59 %)	265 (40.21 %)	199 (30.20 %)
educDAD	0 - Sin estudios Universitarios	21 (4.52 %)	312 (67.10 %)	132 (28.39 %)
educDAD	1 - Con estudios Universitarios	244 (25.55 %)	407 (42.62 %)	304 (31.83 %)
eduMOM	0 - Sin estudios Universitarios	27 (6.29 %)	301 (70.16 %)	101 (23.54 %)
eduMOM	1 - Con estudios Universitarios	238 (23.78 %)	428 (42.76 %)	335 (33.47 %)
MaritalStatus	0 - Soltero	246 (29.29 %)	286 (34.05 %)	308 (36.67 %)
MaritalStatus	1 - Casado	19 (3.28 %)	433 (74.66 %)	128 (22.07 %)
hijos	0 - No tiene hijos	260 (24.48 %)	409 (38.51 %)	393 (37.01 %)
hijos	1 - Tiene hijos	5 (1.40 %)	310 (86.59 %)	43 (12.01 %)

Cuadro 5.6: Distribución de sujetos según variables predictoras de las clases.

1, por lo tanto la correlación entre las medidas repetidas de los individuos será baja. En nuestro modelamiento se incluyó un coeficiente, ω_g, para considerar una variabilidad clase específica en los efectos aleatorios. Para su interpretación se calcularon los coeficientes de correlación intraclase de la siguiente manera:

Coeficiente de correlación intraclase para los sujetos pertenecientes a la clase latente 1:

$$Corr(Y_{ij1}, Y_{il1}) = \frac{\omega_1^2 \sigma_{u_0}^2}{1 + \omega_1^2 \sigma_{u_0}^2} = 0.423877.$$

Coeficiente de correlación intraclase para los sujetos pertenecientes a la clase latente 2:

$$Corr(Y_{ij2}, Y_{il2}) = \frac{\omega_2^2 \sigma_{u_0}^2}{1 + \omega_2^2 \sigma_{u_0}^2} = 0.526906.$$

Coeficiente de correlación intraclase para los sujetos pertenecientes a la clase latente 3:

$$Corr(Y_{ij3}, Y_{il3}) = \frac{\omega_3^2 \sigma_{u_0}^2}{1 + \omega_3^2 \sigma_{u_0}^2} = 0.543758.$$

Conforme a estos cálculos, la correlación entre las medidas repetidas de los individuos que pertenecen a la clase 3 es ligeramente mayor respecto a la correlación que exhiben las medidas repetidas individuales en las otras clases. En otras palabras, las distintas mediciones de religiosidad para un mismo sujeto estarían ligeramente más vinculadas, si el sujeto pertenece a la clase latente 3.

Parámetros asociados a efectos poblacionales					
Covariables		Coeficiente	Desviación estándar	Test de Wald	p-value
sex		0.4203	0.0712	5.907	0.0000
marihuana		-0.1775	0.0421	-4.222	0.0000
parentsIMPORT		0.3248	0.0354	9.189	0.0000
parentsASIST		0.2141	0.0220	9.744	0.0000
Parámetros asociados a efectos clase específtco					
Parámetro	Clase	Coeficiente	Desviación estándar	Test de Wald	p-value
Intercepto	2	-2.7085	0.5089	-5.322	0.0000
Intercepto	3	-1.7979	0.3928	-4.577	0.0000
Covariable	Clase	Coeficiente	Desviación estándar	Test de Wald	p-value
edad	1	-0.1626	0.0143	-11.372	0.0000
edad	2	0.0243	0.0087	2.792	0.0052
edad	3	-0.1300	0.0164	-7.922	0.0000
afilia	1	0.6776	0.2019	3.357	0.0008
afilia	2	1.5484	0.1104	14.021	0.0000
afilia	3	3.2693	0.2065	15.834	0.0000
religiouspeer	1	0.3351	0.1588	2.110	0.0348
religiouspeer	2	0.6220	0.0794	7.831	0.0000
religiouspeer	3	0.9217	0.1237	7.452	0.0000
Parámetros asociados a los efectos aleatorios					
Varianza del intercepto aleatorio			1.1918		
Coeficiente proporcional de la varianza del efecto aleatorio			Clase	Coeficiente	Desviación estándar
	ω_1		1	0.7857	0.1191
	ω_2		2	0.9667	0.1062
Parámetros asociadas a la función de enlace					
Umbral		Coeficiente	Desviación estándar	Test de Wald	p-value
η_1		-2.9855	0.2992	-9.978	0.0000
η_2		0.8219	0.0331	24.815	0.0000
η_3		0.6564	0.0291	22.596	0.0000
η_4		0.6200	0.0265	23.379	0.0000
η_5		0.6776	0.0242	27.951	0.0000
η_6		0.6471	0.0223	28.995	0.0000
η_7		0.6014	0.0210	28.615	0.0000
η_8		0.6157	0.0196	31.351	0.0000
η_9		0.6101	0.0184	33.199	0.0000
η_{10}		0.5759	0.0176	32.745	0.0000
η_{11}		0.5558	0.0172	32.388	0.0000
η_{12}		0.5478	0.0169	32.495	0.0000
η_{13}		0.5484	0.0167	32.786	0.0000
η_{14}		0.5363	0.0166	32.244	0.0000
η_{15}		0.5431	0.0167	32.533	0.0000
η_{16}		0.5927	0.0168	35.275	0.0000
η_{17}		0.6109	0.0173	35.366	0.0000
η_{18}		0.6052	0.0183	33.109	0.0000
η_{19}		0.6280	0.0195	32.227	0.0000
η_{20}		0.6646	0.0211	31.493	0.0000
η_{21}		0.8013	0.0243	32.936	0.0000
η_{22}		0.9582	0.0335	28.611	0.0000

Cuadro 5.7: Estimadores del sub-modelo de regresión lineal mixto con variable manifiesta *Irelig*.

Como complemento al análisis previo, describiremos ahora por clase las 4 escalas que com-

ponen nuestro índice *Irelig*: $IMPORT$, $CREER$, $REZAR$ y $ASISTIR$, según la variable edad.

- **La clase latente 1.** Los sujetos en esta clase muestran trayectorias de declinamiento pronunciado en 2 de las 4 escalas tal como se observa en la Figura 5.4. As´ı, al final del periodo de estudio, la mayoria de los sujetos rezarán menos de 1 vez al mes y asistirán muy pocas veces al año al servicio religioso e incluso algunos dejarán de asistir. En las otras 2 escalas, en general, caen sólo un nivel durante todo el periodo. De modo que en la adultez se sentirán algo lejanos de Dios y su fe religiosa no será muy importante en su vida. Esta clase representa el 18.66 % de nuestra muestra.

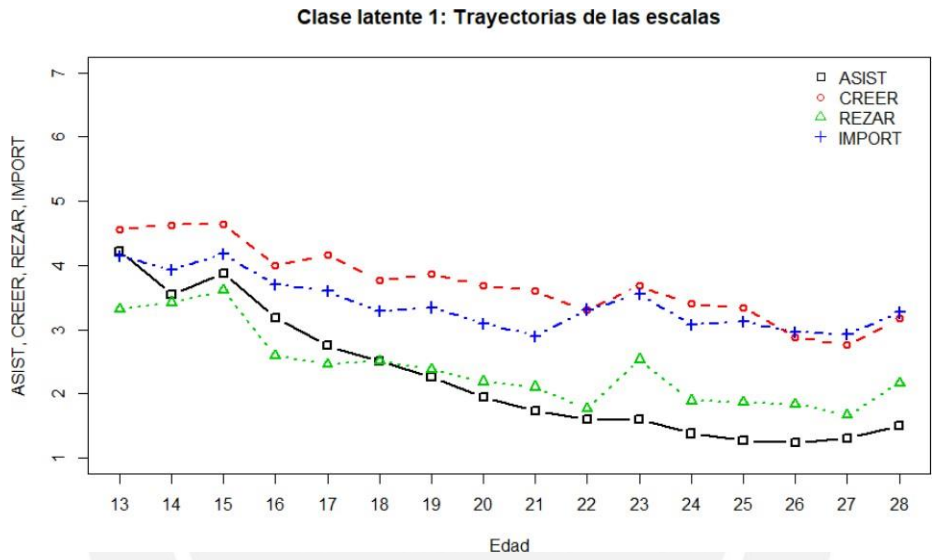

Figura 5.4: Trayectorias de las escalas de la variable manifiesta *Irelig* para la clase 1.

- **La clase latente 2.** Los sujetos en esta clase muestran niveles constantemente altos en las 4 escalas tal como se observa en la Figura 5.5. En general, se sienten muy cercanos a Dios, rezan varias veces a la semana, dismimuye levemente su asistencia al servicio religioso y su fe es considerada importante en su vida. Esta clase representa el 50.63 % de nuestra muestra.

- **La clase latente 3.** Los sujetos en esta clase muestran una trayectoria de declinamiento pronunciado sólo en la escala de asistencia al servicio religioso tal como se observa en la Figura 5.6. En las demás escalas bajan desde un nivel intermedio alto hasta un nivel medio durante todo el periodo de observación. En general, estos sujetos asisten pocas veces al año al servicio religioso, se sienten algo cercanos a Dios, rezan 1 vez a la semana y su fe es algo importante en la forma como toman las decisiones en su vida. Esta clase representa el 30.71 % de nuestra muestra.

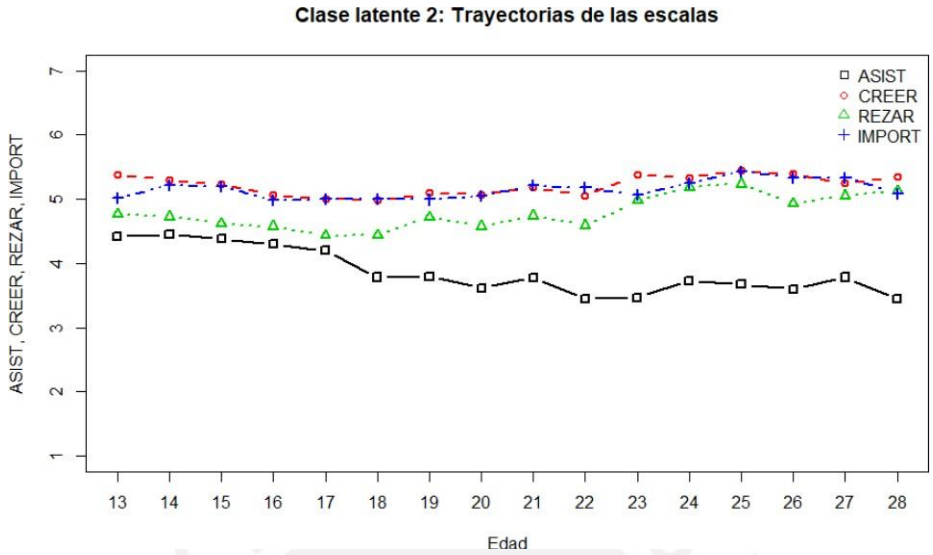

Figura 5.5: Trayectorias de las escalas de la variable manifiesta *Irelig* para la clase 2.

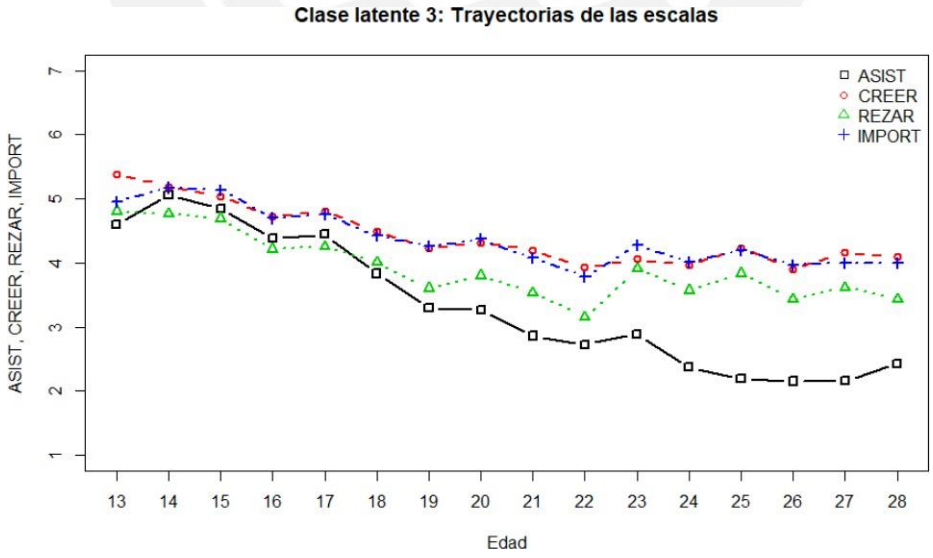

Figura 5.6: Trayectorias de las escalas de la variable manifiesta *Irelig* para la clase 3.

Capítulo 6

Conclusiones y Sugerencias

6.1. Conclusiones

- Los modelos lineales mixtos extendidos desarrollados por Proust-Lima, Philipps y Liquet (2017) constituyen una importante herramienta metodológica para estudiar cómo cambia un fenómeno en el tiempo cuando se tiene una variable respuesta ordinal, nominal, binaria, continua con restricción o incluso latente. En la presente tesis se consideró una variable respuesta latente y una variable manifiesta de tipo ordinal para estudiar el constructo religiosidad.

- Con respecto al estudio de simulación, cuando la función de enlace es de tipo umbral resulta aconsejable que la variable respuesta manifiesta tenga más de 20 niveles para obtener buenas estimaciones de los parámetros. A mayor número de niveles de la variable manifiesta, mejores estimaciones.

- Otra cuestión a tomar en cuenta respecto al estudio de simulación es que si se tiene un modelo que incluye efectos aleatorios y una función de enlace que es de tipo umbral, el tiempo de estimación crecerá exponencialmente porque es necesario integrar numéricamente sobre los efectos aleatorios en cada cálculo de la contribución individual a la función de verosimilitud.

- Cuando se trabaja con los LCMM es necesario realizar las estimaciones bajo varios valores iniciales dada la posible múltiple presencia de óptimos locales en la verosimilitud. Es por ello que se recomienda utilizar la función *gridsearch* para seleccionar aquellos valores iniciales que ayudarán tanto a obtener la mejor verosimilitud como a disminuir el tiempo de estimación. Recordemos que los valores iniciales se obtienen estimando el modelo sobre una población homogénea.

- Las ventajas de nuestro modelo se apreciaron en la aplicación a la base de datos de religiosidad. Se logró identificar 3 clases latentes en los datos. Los sujetos que conforman cada clase latente pudieron ser descritos, al igual que sus respectivas trayectorias de los valores de *Irelig*, a través de las variables que resultaron significativas. Se encontró una clase con trayectoria de religiosidad cuya pérdida de religiosidad a largo del tiempo fue muy baja y 2 clases cuyas trayectorias de religiosidad mostraron un decaimiento más acentuado, siendo más pronunciado en una de ellas. Además, el modelo utilizado permite diferenciar los efectos de las covariables tanto en magnitud como en dirección.

6.2. Sugerencias para investigaciones futuras

- Se recomienda estudiar la versión multivariada de nuestro modelo, es decir, incluir en el modelamiento multiples variables manifiestas. Esto permite que el modelamiento considere que el proceso latente puede afectar a varias variables manifiestas y no sólo a una medida global, como se analizó en nuestro caso con el índice *Irelig*. El paquete "lcmm" aún no implementa la versión multivariada cuando se utiliza enlaces discretos. Una aproximación a lo que podría ocurrir bajo este análisis fue presentada descriptivamente al final de los resultados de nuestra aplicación.

- El tiempo promedio de cada estimación en el estudio de simulación con la variable manifiesta CESD-10, que incluyó a un intercepto y una pendiente aleatoria para el submodelo estructural, fue de aproximadamente 24 horas en una PC con Procesador i7 de 2.8 GHz y 16 GB de memoria RAM. Una de las desventajas de la estimación clásica es el tiempo de estimación, que como se aprecia aquí es bastante considerable. Este incluso podría incrementarse más de incluirse otros efectos aleatorios por el problema de la cuadratura. En tal sentido, ser´ıa interesante explorar la posibilidad de utilizar una metodología Bayesiana en la estimación de los parámetros de este modelo.

Apéndice A

Rutinas en R

A.1. Programa en R para la simulación

Se presenta aquí el código en R para la generación de los datos y la estimación del modelo con la variable manifiesta CESD-10. Los códigos con la variable manifiesta CESD-20 son similares.

```
#install.packages(c("lcmm","MASS"))
library(lcmm) ; library(MASS)
#### #### ####   GENERACIÓN ESTRUCTURA DE LA BASE DE DATOS    #### #### ####
#########################################################################
bd  <- paquid
ni  <- as.numeric(table(bd$ID)) #número de observaciones por sujeto
# Establecimiento de las 2250 observaciones en formato largo
obs <- unlist(sapply(ni, function(x) sort(1:x)))
bd  <- cbind(bd,obs)
M   <- length(bd$obs) ;M #Número total de observaciones

#### #### ####                CLASES LATENTES                #### #### ####
#########################################################################
# Inclusión de la clasificación asociada a las probabilidades a posteriori
clase <- read.csv(file.path(getwd(), "clasifica.csv"), header = T, sep = ",")
table(clase)
prop.table(table(clase))*100 # porcentaje de sujetos por clase
# Distribución de sujetos por clase
n1 <- table(clase)[1] ; n1 #Clase1
n2 <- table(clase)[2] ; n2 #Clase2
N <- n1 + n2 ; N # Número total de sujetos en el estudio
# Inclusión de la clase latente en la base de datos
clase <- as.matrix(clase)
CiT <- rep(clase, ni)
bd <- cbind(bd, CiT)
#### #### ####                   COVARIABLES                  #### #### ####
#########################################################################
bd$age65 <- (bd$age-65)/10 #Creación de la variable age65

bd_sim <- as.data.frame(cbind(bd$ID,obs,
bd$age65,
bd$male,
bd$CiT,
bd$CEP))
# Asignación de nombres a las columnas
colnames(bd_sim) <- c("ID","obs","age65","male","clase","AF")
str(bd_sim)
bd_sim$ID <- as.integer(bd_sim$ID)
```

```
bd_sim$male <- as.integer(bd_sim$male)
bd_sim$AF <- as.integer(bd_sim$AF)
#### #### ## ASIGNACIÓN DE VALORES A LOS PARÁMETROS DEL MODELO ### #### ####
# ##############################################################################
### Asignación de valores reales del modelo de pertenecia
X01 <- 0.65 # intercepto de la clase 1

### Asignación de valores reales del modelo estructural
# Efectos fijos comunes a toda la  población
b1  <-  -3.93 #sex
b2  <-   0.19 #male*age65
b3  <-  -3.61 #AF
# Efectos fijos clase 1
v11 <-  1.23 #Parámetro asociado a la variable age 65
# Efectos fijos clase 2
v02 <-  3.78  #Intercepto
v12 <-  2.75 #Parámetro asociado a la variable age 65
### Asignación de valores reales del modelo de medición
eta1Ast <- 0.52 #Primer umbral
eta2    <- 0.95
eta2Ast <- eta1Ast+(eta2^2) #Segundo umbral
eta3    <- 0.77
eta3Ast <- eta2Ast+(eta3^2)#Tercer umbral
eta4    <- 0.5
eta4Ast <- eta3Ast+(eta4^2)#Cuarto umbral
eta5    <- 0.69
eta5Ast <- eta4Ast+(eta5^2)#Quinto umbral
eta6    <- 0.63
eta6Ast <- eta5Ast+(eta6^2)#Sexto umbral
eta7    <- 0.43
eta7Ast <- eta6Ast+(eta7^2)#Septimo umbral
eta8    <- 0.57
eta8Ast <- eta7Ast+(eta8^2)#Octavo umbral
eta9    <- 0.56
eta9Ast <- eta8Ast+(eta9^2)#Noveno umbral
eta10    <- 0.55
eta10Ast <- eta9Ast+(eta10^2)#Decimo umbral
eta11    <- 0.36
eta11Ast <- eta10Ast+(eta11^2)#Umbral 11
eta12    <- 0.52
eta12Ast <- eta11Ast+(eta12^2)#Umbral 12
eta13    <- 0.52
eta13Ast <- eta12Ast+(eta13^2)#Umbral 13
eta14    <- 0.34
eta14Ast <- eta13Ast+(eta14^2)#Umbral 14
eta15    <- 0.56
eta15Ast <- eta14Ast+(eta15^2)#Umbral 15
eta16    <- 0.57
eta16Ast <- eta15Ast+(eta16^2)#Umbral 16
eta17    <- 0.37
eta17Ast <- eta16Ast+(eta17^2)#Umbral 17
eta18    <- 0.63
eta18Ast <- eta17Ast+(eta18^2)#Umbral 18
eta19    <- 0.45
eta19Ast <- eta18Ast+(eta19^2)#Umbral 19
eta20    <- 0.56
eta20Ast <- eta19Ast+(eta20^2)#Umbral 20
eta21    <- 0.35
eta21Ast <- eta20Ast+(eta21^2)#Umbral 21
eta22    <- 0.67
```

```
eta22Ast <- eta21Ast+(eta22^2)#Umbral 22
eta23    <- 0.54
eta23Ast <- eta22Ast+(eta23^2)#Umbral 23
eta24    <- 0.51
eta24Ast <- eta23Ast+(eta24^2)#Umbral 24
eta25    <- 0.5
eta25Ast <- eta24Ast+(eta25^2)#Umbral 25
eta26    <- 0.6
eta26Ast <- eta25Ast+(eta26^2)#Umbral 26
eta27    <- 0.38
eta27Ast <- eta26Ast+(eta27^2)#Umbral 27
eta28    <- 0.65
eta28Ast <- eta27Ast+(eta28^2)#Umbral 28
eta29    <- 0.48
eta29Ast <- eta28Ast+(eta29^2)#Umbral 29

### Asignación de valores reales de la var-cov de los efectos aleatorios
varcov1   <-  128.84
varcov2   <-  -63.35
varcov3   <-   54.89
### Asignación de valores reales de la varianza de los errores de medición
Sigma <- 1
#### Vector de los valores reales asignados a los 39 parámetros
vr <- c( X01 , v02 , b1 , v11 , v12 , b3 , b2 ,
varcov1, varcov2, varcov3, eta1Ast, eta2, eta3, eta4,eta5,eta6,eta7,eta8,eta9,
eta10, eta11, eta12, eta13, eta14,eta15,eta16,eta17,eta18,eta19,
eta20,eta21,eta22,eta23,eta24,eta25,eta26,eta27,eta28,eta29)

#### #### #### ####            SIMULACIÓN          #### #### #### ####
##############################################################################
S  <- 1 # número de simulaciones
# # Matriz para almacenar todas las simulaciones
resultados   <- NULL
Parametros   <- matrix(0,nrow = S,ncol = 39)
Porcentajes <- matrix(0,nrow = S,ncol = 2)
VarPar       <- matrix(0,nrow = S,ncol = 39)
# Calcula el tiempo que tomará la simulación
t <- proc.time()
#
for(s in 1: S){

####  Simulación de los efectos aleatorios
u   <- mvrnorm(N, mu = c(0,0), Sigma = matrix(c(varcov1,varcov2,varcov2,varcov3),nrow
    = 2))
bdsimx = bd_sim
u0  <- u[,1]
u1  <- u[,2]
bdsimx$u0 <- rep(u0,ni)
bdsimx$u1 <- rep(u1,ni)

#### Simulación de los errores de medición
bdsimx$eij <- rnorm(M, 0, 1)

#### Separación de la base de datos según clase
datosC1  <- bdsimx[bdsimx$clase==1, ]
datosC2  <- bdsimx[bdsimx$clase==2, ]

####  Simulación de sub-modelo estructural según clase
# Simulación de la variable latente Lambda CLASE 1
datosC1$Lamba_ij <- (b1*datosC1$male)+(b2*(datosC1$male)*(datosC1$age65))+(b3*datosC1$
```

```
    AF)+
(v11*datosC1$age65) +
(datosC1$u0) + (datosC1$u1*(datosC1$age65))
# Simulación de la variable latente Lambda CLASE 2
datosC2$Lamba_ij <- (b1*datosC2$male)+(b2*(datosC2$male)*(datosC2$age65))+(b3*datosC2$
    AF)+
(v02) + (v12*datosC2$age65) +
(datosC2$u0) + (datosC2$u1*(datosC2$age65))

####  Simulación de la variable latente intermedia según clase
# Simulación del Ytilde_ij (variable latentes con ruido)- Clase 1
datosC1$Ytilde_ij <- datosC1$Lamba_ij + (datosC1$eij)
# Simulación del Ytilde_ij (variable latentes con ruido)- Clase 2
datosC2$Ytilde_ij <- datosC2$Lamba_ij + (datosC2$eij)

####  Simulación sub-modelo de medición según clase
# Simulación de la variable manifiesta ordinal Y - CLASE 1
datosC1$Y <- rep(1,length(datosC1$Ytilde_ij))
datosC1$Y[datosC1$Ytilde_ij <= eta1Ast] <- 0
datosC1$Y[eta2Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta3Ast]    <- 2
datosC1$Y[eta3Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta4Ast]    <- 3
datosC1$Y[eta4Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta5Ast]    <- 4
datosC1$Y[eta5Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta6Ast]    <- 5
datosC1$Y[eta6Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta7Ast]    <- 6
datosC1$Y[eta7Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta8Ast]    <- 7
datosC1$Y[eta8Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta9Ast]    <- 8
datosC1$Y[eta9Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta10Ast]   <- 9
datosC1$Y[eta10Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta11Ast] <- 10
datosC1$Y[eta11Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta12Ast] <- 11
datosC1$Y[eta12Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta13Ast] <- 12
datosC1$Y[eta13Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta14Ast] <- 13
datosC1$Y[eta14Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta15Ast] <- 14
datosC1$Y[eta15Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta16Ast] <- 15
datosC1$Y[eta16Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta17Ast] <- 16
datosC1$Y[eta17Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta18Ast] <- 17
datosC1$Y[eta18Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta19Ast] <- 18
datosC1$Y[eta19Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta20Ast] <- 19
datosC1$Y[eta20Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta21Ast] <- 20
datosC1$Y[eta21Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta22Ast] <- 21
datosC1$Y[eta22Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta23Ast] <- 22
datosC1$Y[eta23Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta24Ast] <- 23
datosC1$Y[eta24Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta25Ast] <- 24
datosC1$Y[eta25Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta26Ast] <- 25
datosC1$Y[eta26Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta27Ast] <- 26
datosC1$Y[eta27Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta28Ast] <- 27
datosC1$Y[eta28Ast<datosC1$Ytilde_ij & datosC1$Ytilde_ij<=eta29Ast] <- 28
datosC1$Y[datosC1$Ytilde_ij > eta29Ast]  <- 29

# Simulación de la variable manifiesta ordinal Y - CLASE 2
datosC2$Y <- rep(1,length(datosC2$Ytilde_ij))
datosC2$Y[datosC2$Ytilde_ij <= eta1Ast] <- 0
datosC2$Y[eta2Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta3Ast]    <- 2
datosC2$Y[eta3Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta4Ast]    <- 3
datosC2$Y[eta4Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta5Ast]    <- 4
datosC2$Y[eta5Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta6Ast]    <- 5
datosC2$Y[eta6Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta7Ast]    <- 6
datosC2$Y[eta7Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta8Ast]    <- 7
datosC2$Y[eta8Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta9Ast]    <- 8
datosC2$Y[eta9Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta10Ast]   <- 9
datosC2$Y[eta10Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta11Ast] <- 10
```

```
datosC2$Y[eta11Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta12Ast] <- 11
datosC2$Y[eta12Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta13Ast] <- 12
datosC2$Y[eta13Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta14Ast] <- 13
datosC2$Y[eta14Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta15Ast] <- 14
datosC2$Y[eta15Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta16Ast] <- 15
datosC2$Y[eta16Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta17Ast] <- 16
datosC2$Y[eta17Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta18Ast] <- 17
datosC2$Y[eta18Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta19Ast] <- 18
datosC2$Y[eta19Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta20Ast] <- 19
datosC2$Y[eta20Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta21Ast] <- 20
datosC2$Y[eta21Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta22Ast] <- 21
datosC2$Y[eta22Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta23Ast] <- 22
datosC2$Y[eta23Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta24Ast] <- 23
datosC2$Y[eta24Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta25Ast] <- 24
datosC2$Y[eta25Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta26Ast] <- 25
datosC2$Y[eta26Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta27Ast] <- 26
datosC2$Y[eta27Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta28Ast] <- 27
datosC2$Y[eta28Ast<datosC2$Ytilde_ij & datosC2$Ytilde_ij<=eta29Ast] <- 28
datosC2$Y[datosC2$Ytilde_ij > eta29Ast]  <- 29

####  Base de datos simulados
datos_simulados <- rbind(datosC1,datosC2) #
ordenar los datos
datos_simulados <- datos_simulados[order(datos_simulados$ID, datos_simulados$obs),]

#### #### #### ####          ESTIMACIÓN CON LCMM          #### #### #### ####
# ##############################################################################

## Estimar el modelo base para una pobl. homogénea
mod_base <- lcmm(Y~male+(male*age65)+AF,
random = ~ age65,
subject = "ID",
data = datos_simulados,
link = 'thresholds',
ng =1)

# Uso de gridsearch para encontrar un conjunto de valores iniciales ## que
maximizen la convergencia a un máximo global
mgrid <- gridsearch(rep = 30, maxiter = 15, minit = mod_base,
lcmm(Y~male+(male*age65)+AF,
mixture = ~age65,
random = ~ age65,
subject = "ID",
data  =  datos_simulados,
link = 'thresholds',
ng =2))

# ###  ####  ####  ####        REORDENANDO CLASES          #### #### #### ####
CLA1 <- summarytable(mgrid)[5]
CLA2 <- summarytable(mgrid)[6]

if( CLA1 < CLA2 ) {

Parametros[s,1]  <- -1* mgrid$best[1]
Parametros[s,2]  <-  mgrid$best[2]
Parametros[s,3]  <-  mgrid$best[3]
Parametros[s,4]  <-  mgrid$best[5]
Parametros[s,5]  <-  mgrid$best[4]
Parametros[s,6]  <-  mgrid$best[6]
Parametros[s,7]  <-  mgrid$best[7]
```

```
Parametros[s,8]  <-  mgrid$best[8]
Parametros[s,9]  <-  mgrid$best[9]
Parametros[s,10] <- mgrid$best[10]
Parametros[s,11] <- mgrid$best[11]
Parametros[s,12] <- mgrid$best[12]
Parametros[s,13] <- mgrid$best[13]
Parametros[s,14] <- mgrid$best[14]
Parametros[s,15] <- mgrid$best[15]
Parametros[s,16] <- mgrid$best[16]
Parametros[s,17] <- mgrid$best[17]
Parametros[s,18] <- mgrid$best[18]
Parametros[s,19] <- mgrid$best[19]
Parametros[s,20] <- mgrid$best[20]
Parametros[s,21] <- mgrid$best[21]
Parametros[s,22] <- mgrid$best[22]
Parametros[s,23] <- mgrid$best[23]
Parametros[s,24] <- mgrid$best[24]
Parametros[s,25] <- mgrid$best[25]
Parametros[s,26] <- mgrid$best[26]
Parametros[s,27] <- mgrid$best[27]
Parametros[s,28] <- mgrid$best[28]
Parametros[s,29] <- mgrid$best[29]
Parametros[s,30] <- mgrid$best[30]
Parametros[s,31] <- mgrid$best[31]
Parametros[s,32] <- mgrid$best[32]
Parametros[s,33] <- mgrid$best[33]
Parametros[s,34] <- mgrid$best[34]
Parametros[s,35] <- mgrid$best[35]
Parametros[s,36] <- mgrid$best[36]
Parametros[s,37] <- mgrid$best[37]
Parametros[s,38] <- mgrid$best[38]
Parametros[s,39] <- mgrid$best[39]

#
Porcentajes[s,] <- c( summarytable(mgrid)[6] , summarytable(mgrid)[5]) #
VarPar[s,1]      <- as.numeric(diag(VarCov(mgrid))[1])
VarPar[s,2]      <- as.numeric(diag(VarCov(mgrid))[2])
VarPar[s,3]      <- as.numeric(diag(VarCov(mgrid))[3])
VarPar[s,4]      <- as.numeric(diag(VarCov(mgrid))[5])
VarPar[s,5]      <- as.numeric(diag(VarCov(mgrid))[4])
VarPar[s,6]      <- as.numeric(diag(VarCov(mgrid))[6])
VarPar[s,7]      <- as.numeric(diag(VarCov(mgrid))[7])
VarPar[s,8]      <- as.numeric(diag(VarCov(mgrid))[8])
VarPar[s,9]      <- as.numeric(diag(VarCov(mgrid))[9])
VarPar[s,10]     <- as.numeric(diag(VarCov(mgrid))[10])
VarPar[s,11]     <- as.numeric(diag(VarCov(mgrid))[11])
VarPar[s,12]     <- as.numeric(diag(VarCov(mgrid))[12])
VarPar[s,13]     <- as.numeric(diag(VarCov(mgrid))[13])
VarPar[s,14]     <- as.numeric(diag(VarCov(mgrid))[14])
VarPar[s,15]     <- as.numeric(diag(VarCov(mgrid))[15])
VarPar[s,16]     <- as.numeric(diag(VarCov(mgrid))[16])
VarPar[s,17]     <- as.numeric(diag(VarCov(mgrid))[17])
VarPar[s,18]     <- as.numeric(diag(VarCov(mgrid))[18])
VarPar[s,19]     <- as.numeric(diag(VarCov(mgrid))[19])
VarPar[s,20]     <- as.numeric(diag(VarCov(mgrid))[20])
VarPar[s,21]     <- as.numeric(diag(VarCov(mgrid))[21])
VarPar[s,22]     <- as.numeric(diag(VarCov(mgrid))[22])
VarPar[s,23]     <- as.numeric(diag(VarCov(mgrid))[23])
VarPar[s,24]     <- as.numeric(diag(VarCov(mgrid))[24])
```

```
VarPar[s,25]     <- as.numeric(diag(VarCov(mgrid))[25])
VarPar[s,26]     <- as.numeric(diag(VarCov(mgrid))[26])
VarPar[s,27]     <- as.numeric(diag(VarCov(mgrid))[27])
VarPar[s,28]     <- as.numeric(diag(VarCov(mgrid))[28])
VarPar[s,29]     <- as.numeric(diag(VarCov(mgrid))[29])
VarPar[s,30]     <- as.numeric(diag(VarCov(mgrid))[30])
VarPar[s,31]     <- as.numeric(diag(VarCov(mgrid))[31])
VarPar[s,32]     <- as.numeric(diag(VarCov(mgrid))[32])
VarPar[s,33]     <- as.numeric(diag(VarCov(mgrid))[33])
VarPar[s,34]     <- as.numeric(diag(VarCov(mgrid))[34])
VarPar[s,35]     <- as.numeric(diag(VarCov(mgrid))[35])
VarPar[s,36]     <- as.numeric(diag(VarCov(mgrid))[36])
VarPar[s,37]     <- as.numeric(diag(VarCov(mgrid))[37])
VarPar[s,38]     <- as.numeric(diag(VarCov(mgrid))[38])
VarPar[s,39]     <- as.numeric(diag(VarCov(mgrid))[39])

} else{

Parametros[s,]  <- mgrid$best[c(1:39)]
Porcentajes[s,] <- summarytable(mgrid)[5:6]
VarPar[s,]      <- as.numeric(diag(VarCov(mgrid)))
}


## ALMACENAMIENTO DE LOS RESULTADOS
resultados <- append(resultados,list(Parametros[s,],
Porcentajes[s,],
VarPar[s,]))
}
# Detiene el cronómetro
proc.time()-t

#Guardar los resultadosen en formato R
save(resultados, file=file.path(getwd(),"res_sim.rda"))
```

:

A.2. Programa en R para la aplicación a la base de datos sobre religión

```
########  Base de datos
load("DATOS_V3.Rda")

######## Librerías
library(lcmm) ; library(ggplot2) ; library(devtools) ; library(tidyverse)

################################################################################ #
################                ESTIMACIÓN DEL MODELO                ################
set. seed (1234)
########       Modelo con 1 clase latente
mod1_Irelig_v1 <- lcmm(Irelig ~ Edad + sex + marihuana + afilia + religiouspeer +
parentsIMPORT + parentsASIST,
random = ~ 1,
subject = "ID",
data = DATOS _ V3,
link = ' thresholds ', ng
=1)
mod1 _ Irelig _ v1 $ best
length ( mod1 _ Irelig _ v1 $ best)
summarytable ( mod1 _ Irelig _ v1 )
summary ( mod1 _ Irelig _ v1 )

########       Modelo con 2 clases latentes
mod2_Irelig_v1 <- gridsearch(rep = 30, maxiter = 15, minit = mod1_Irelig_v1,
lcmm(Irelig ~ Edad + sex + marihuana + afilia + religiouspeer +
parentsIMPORT + parentsASIST,
mixture = ~ Edad + religiouspeer + afilia ,
random = ~ 1,
subject = "ID",
data = DATOS _ V3,
link = ' thresholds ',
classmb = ~ raza +
educDAD + educMOM + educSubject + maritalStatus + hijos, ng =2,
nwg = TRUE ))
mod2 _ Irelig _ v1 $ best
length ( mod2 _ Irelig _ v1 $ best)
summarytable ( mod2 _ Irelig _ v1 )
summary ( mod2 _ Irelig _ v1 )

########        Modelo con 3 clases latentes
mod3_Irelig_v1 <- gridsearch(rep = 30, maxiter = 15, minit = mod1_Irelig_v1,
lcmm(Irelig ~ Edad + sex + marihuana + afilia + religiouspeer +
parentsIMPORT + parentsASIST,
mixture = ~ Edad + religiouspeer + afilia ,
random = ~ 1,
subject = "ID",
data = DATOS _ V3,
link = ' thresholds ',
classmb = ~ raza +
educDAD + educMOM + educSubject + maritalStatus + hijos, ng =3,
nwg = TRUE ))
mod3 _ Irelig _ v1 $ best
length ( mod3 _ Irelig _ v1 $ best)
summarytable ( mod3 _ Irelig _ v1 )
summary ( mod3 _ Irelig _ v1 )
```

```
########        Modelo con 4 clases latentes
mod4_Irelig_v1 <- gridsearch(rep = 30, maxiter = 15, minit = mod1_Irelig_v1,
lcmm(Irelig ~ Edad + sex + marihuana + afilia + religiouspeer +
parentsIMPORT + parentsASIST,
mixture = ~ Edad + religiouspeer + afilia ,
random = ~ 1,
subject = "ID",
data = DATOS_V3,
link = 'thresholds',
classmb = ~ raza +
educDAD + educMOM + educSubject + maritalStatus + hijos, ng =4,
nwg = TRUE ))
mod4_Irelig_v1$best
length(mod4_Irelig_v1$best)
summarytable(mod4_Irelig_v1)
summary(mod4_Irelig_v1)
### Obs: No alcanzó la convergencia

########            Comparación de modelos
summarytable(mod1_Irelig_v1, mod2_Irelig_v1, mod3_Irelig_v1)
summary(mod1_Irelig_v1)
summary(mod2_Irelig_v1)
summary(mod3_Irelig_v1)

########        ODDS y OR
mod3_Irelig_v1$best
round(mod3_Irelig_v1$best, digits = 4)
exp(mod3_Irelig_v1$best)
round(exp(mod3_Irelig_v1$best), digits =4)
(round(exp(mod3_Irelig_v1$best), digits =4) -1)* 100

########   Plot clase vs 4 escalas - Edad
## Clasificación 3 clases latentes
clasificaSubj  <- mod3_Irelig_v1$pprob[,2]
class(clasificaSubj)
clasificaSubj <- as.matrix(clasificaSubj)
ID_SUBClass <- as.numeric(mod3_Irelig_v1$pprob[,1])
BDfinal <- subset(DATOS_V3, DATOS_V3$ID %in% ID_SUBClass)
length(summary(as.factor(BDfinal$ID),maxsum=50000))
ni_C <- as.numeric(table(BDfinal$ID))
CiT_C <- rep(clasificaSubj, ni_C)
BDfinal$clase  <- CiT_C

########  Plot de media de datos agrupados y medidas repetidas
gd <- BDfinal %>%
group_by(clase, Edad) %>%
summarise(ASIST  = mean(as.numeric(ASIST),  na.rm = TRUE),
REZAR  = mean(as.numeric(REZAR),  na.rm = TRUE),
IMPORT = mean(as.numeric(IMPORT), na.rm = TRUE),
CREER  = mean(as.numeric(CREER),  na.rm = TRUE),
Irelig  = mean(as.numeric(Irelig),  na.rm = TRUE))

ggplot(BDfinal, aes(x= Edad, y=Irelig, color=as.factor(clase)))+
geom_line(aes(group=ID), alpha= .3)+
geom_line(data = gd, alpha= .8, size=3)+
theme_bw()+
labs(
title="Evolución de religiosidad según clase",
```

```
x = " Edad ",
y = " Irelig ",
color= NULL
) +
labs(colour= "Clases latentes") +
scale_color_manual(labels = c("clase 1: 18.66%", "clase 2: 50.63%","clase 3: 30.71%"),
    values = c("red", "blue", "green")) + theme_bw(base_size=20) +
theme(plot.title = element_text(hjust = 0.5))

## BD sujetos de la clase 1
BDclase1 <- BDfinal[BDfinal$clase == '1',]

## Evolución de religiosidad clase 1
ggplot(data = BDclase1, aes(x = Edad, y = Irelig, group = ID)) + geom_line()+
scale_color_manual(values = c("red")) + theme_bw(base_size=20) +
theme( plot. title = element_ text( hjust = 0.5))+
labs(
title=" Evolución de religiosidad clase 1",
x = " Edad ",
y = " Irelig ",
color= NULL
) + geom_line( color="red", size=0.25)

## BD sujetos de la clase 2
BDclase2 <- BDfinal[BDfinal$clase == '2',]

## Evolución de religiosidad clase 2
ggplot(data = BDclase2, aes(x = Edad, y = Irelig, group = ID)) + geom_line()+
scale_color_manual(values = c("blue")) + theme_bw(base_size=20) +
theme( plot. title = element_ text( hjust = 0.5))+
labs(
title=" Evolución de religiosidad clase 2",
x = " Edad ",
y = " Irelig ",
color= NULL
) + geom_line( color="blue", size=0.25)


## BD sujetos de la clase 3
BDclase3 <- BDfinal[BDfinal$clase == '3',]

## Evolución de religiosidad clase 3
ggplot(data = BDclase3, aes(x = Edad, y = Irelig, group = ID)) + geom_line()+
scale_color_manual(values = c("green")) + theme_bw(base_size=20) +
theme( plot. title = element_ text( hjust = 0.5))+
labs(
title=" Evolución de religiosidad clase 3",
x = " Edad ",
y = " Irelig ",
color= NULL
) + geom_line( color="green", size=0.25)


####   Trayectoria latente 1 - Consistemente alta religiosidad
age <- c(13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,
    28)

##  ##  ##  ##  ##  ##     CLASE 1

## Serie promedio de ASIST  ----- CLASE 1
```

```
serieAc1 <- gd%>% filter(clase == 1)
serie_Ac1 <- serieAc1$ASIST
serie_ Ac1 <- serie_ Ac1 [!( is.na( serie_ Ac1 ))]

## serie promedio de CREER  ----- CLASE 1
serieCc1 <- gd%>% filter(clase == 1)
serie_Cc1 <- serieCc1$CREER
serie_ Cc1 <- serie_ Cc1 [!( is.na( serie_ Cc1 ))]

## Serie promedio de  REZAR ----- CLASE 1
serieRc1 <- gd%>% filter(clase == 1)
serie_Rc1 <- serieRc1$REZAR
serie_ Rc1 <- serie_ Rc1 [!( is.na( serie_ Rc1 ))]

## serie promedio de IMPORT ----- CLASE 1
serieIc1 <- gd %>% filter( clase == 1) serie_
Ic1 <- serieIc1$IMPORT
serie_Ic1 <- serie_Ic1[!(is.na(serie_Ic1))]
serie_ Ic1 <- serie_ Ic1 * 7 /5

## serie promedio de IRELIG ----- CLASE 1
serieIRc1 <- gd%>% filter(clase == 1)
serie_IRc1 <- serieIRc1$Irelig
serie_IRc1 <- serie_IRc1[!(is.na(serie_IRc1))]

##  ##  ##  ##  ##  ##     CLASE 2

## Serie promedio de ASIST  ----- CLASE 2
serieAc2 <- gd%>% filter(clase == 2)
serie_Ac2 <- serieAc2$ASIST
serie_ Ac2 <- serie_ Ac2 [!( is.na( serie_ Ac2 ))]

## serie promedio de CREER  ----- CLASE 2
serieCc2 <- gd%>% filter(clase == 2)
serie_Cc2 <- serieCc2$CREER
serie_ Cc2 <- serie_ Cc2 [!( is.na( serie_ Cc2 ))]

## Serie promedio de  REZAR ----- CLASE 2
serieRc2 <- gd%>% filter(clase == 2)
serie_Rc2 <- serieRc2$REZAR
serie_ Rc2 <- serie_ Rc2 [!( is.na( serie_ Rc2 ))]

## serie promedio de IMPORT ----- CLASE 2
serieIc2 <- gd %>% filter( clase == 2) serie_
Ic2 <- serieIc2$IMPORT
serie_Ic2 <- serie_Ic2 [!(is.na(serie_Ic2 ))]
# Reescalar
serie_ Ic2 <- serie_ Ic2 * 7 / 5

## serie promedio de IRELIG ----- CLASE 2
serieIRc2 <- gd%>% filter(clase == 2)
serie_IRc2 <- serieIRc2$Irelig
serie_IRc2 <- serie_IRc2[!(is.na(serie_IRc2))]

##  ##  ##  ##  ##  ##     CLASE 3

## Serie promedio de ASIST  ----- CLASE 3
serieAc3 <- gd%>% filter(clase == 3)
serie_Ac3 <- serieAc3$ASIST
serie_ Ac3 <- serie_ Ac3 [!( is.na( serie_ Ac3 ))]
```

```
## serie promedio de CREER  ----- CLASE 3
serieCc3 <- gd%>% filter(clase == 3)
serie_Cc3 <- serieCc3$CREER
serie_ Cc3 <- serie_ Cc3 [!( is.na( serie_ Cc3 ))]

## Serie promedio de  REZAR ----- CLASE 3
serieRc3 <- gd%>% filter(clase == 3)
serie_Rc3 <- serieRc3$REZAR
serie_ Rc3 <- serie_ Rc3 [!( is.na( serie_ Rc3 ))]

## serie promedio de IMPORT ----- CLASE 3
serieIc3 <- gd %>% filter( clase == 3) serie_
Ic3 <- serieIc3$IMPORT
serie_Ic3 <- serie_Ic3[!(is.na(serie_Ic3))]
serie_ Ic3 <- serie_ Ic3 * 7 /5

## serie promedio de IRELIG ----- CLASE 3
serieIRc3 <- gd%>% filter(clase == 3)
serie_IRc3 <- serieIRc3$Irelig
serie_IRc3 <- serie_IRc3[!(is.na(serie_IRc3))]

############################## Plot 4 escalas por clase  ###################3

matplot(age,cbind(serie_Ac1,serie_Cc1, serie_Rc1, serie_Ic1),type="b", lwd = 2, pch
    =0:3,
xlab="Edad",ylab="ASIST, CREER, REZAR, IMPORT",
main = "Clase latente 1: Trayectorias de las escalas",
ylim = c(1,7), xaxt="n", xlim = c(13,28))
axis(1 , at = 1:28)
legend("topright", legend=c("ASIST","CREER", "REZAR", "IMPORT"), col=seq_len(4),cex
    =0.975, pch=0:3,  xpd=TRUE, bty="n")
### Trayectorias promedio clase 2
matplot(age,cbind(serie_Ac2,serie_Cc2, serie_Rc2, serie_Ic2),type="b", lwd = 2, pch
    =0:3,
xlab="Edad",ylab="ASIST, CREER, REZAR, IMPORT",
main = "Clase latente 2: Trayectorias de las escalas",
ylim = c(1,7), xaxt="n", xlim = c(13,28))
axis(1 , at = 1:28)
legend("topright", legend=c("ASIST","CREER", "REZAR", "IMPORT"), col=seq_len(4),cex
    =0.975, pch=0:3,  xpd=TRUE, bty="n")

### Trayectorias promedio clase 3
matplot(age,cbind(serie_Ac3,serie_Cc3, serie_Rc3, serie_Ic3),type="b", lwd = 2, pch
    =0:3,
xlab="Edad",ylab="ASIST, CREER, REZAR, IMPORT",
main = "Clase latente 3: Trayectorias de las escalas",
ylim = c(1,7), xaxt="n", xlim = c(13,28))
axis(1 , at = 1:28)
legend("topright", legend=c("ASIST","CREER", "REZAR", "IMPORT"), col=seq_len(4),cex
    =0.975, pch=0:3,  xpd=TRUE, bty="n")
```

:

Bibliografía

Bartholomew, D., Knott, M. y Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, Wiley.

Bollen, K. (2001). *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier Health Sciences.

Cheng, S.-T. y Chan, A. C. M. (2005). The center for epidemiologic studies depression scale in older chinese: thresholds for long and short forms, *International Journal of Geriatric Psychiatry* **20**(5): 465–470.
URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gps.1314

Collins, L. y Lanza, S. (2013). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, John Wiley & Sons.

Commenges, D. y Jacqmin-Gadda, H. (2016). *Dynamical Biostatistical Models*, Taylor & Francis Group, LLC.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements, *Journal of Biometrics and Biostatistics* .

Diggle, P.J. y Heagerty, P.J. (2002). *Analysis of longitudinal data*, Oxford University Press.

Fitzmaurice, G. (2011). *Applied Longitudinal Analysis*, John Wiley & Sons.

Galecki, A. y Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*, Springer.

Hoffmann, J. P. (2014). Religiousness, social networks, moral schemas, and marijuana use: A dynamic dual-process model of culture and behavior, *Social Forces* **93**(1): 181–208.
URL: https://doi.org/10.1093/sf/sou053

Laird, N. y Ware, J. (1982). Random-Effects Models for Longitudinal Data, *Journal of Biometrics & Biostatistics* .

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, *British Journal of Mathematical and Statistical Psychology* .

Oehlert, G. (2012). *A few words about REML*.

Pearce, L. y Foster, M. (2013). A Person-Centered Examination of Adolescent Religiosity Using Latent Class Analysis, *Journal for the Scientific Study of Religion* .

Pearce, L. y Schorpp, K. (2018). Religious pathways from adolescence to adulthood, *Journal for the Scientific Study of Religion* .

Proust, C., Amieva, H. y Jacqmin-Gadda, H. (2013). Analysis of multivariate mixed longitudinal data: A flexible latent process approach, *British Journal of Mathematical and Statistical Psychology* .

Proust-Lima, C., Philipps, V. y Liquet, B. (2017). Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm, *Journal of Statistical Software* **78**(2): 1–56.

Taylor, J., Cumberland, W. y Sy, J. (1994). A Stochastic Model for Analysis of Longitudinal AIDS Data, *American Statistical Association* .

Valdivieso, L. y Tarazona, E. (2016). Notas de clase del curso de modelos de variables latentes, Maestr´ıa en Estad´ıstica. PUCP.