Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística

Anteproyecto de Tesina

Incorporación de covariables que varían en el tiempo a un modelo mixto

Autor: **Esteban Cometto**

Directora: Noelia Castellana

Codirectora: Cecilia Rapelli

15 de junio de 2023

Índice

1. Introducción

Los datos longitudinales están conformados por mediciones repetidas sobre una unidad, las cuales pueden surgir por ser medidas en diferentes momentos o condiciones. Su principal objetivo es estudiar los cambios en el tiempo y los factores que influencian el cambio.

Los modelos mixtos permiten ajustar datos con estas características, donde la respuesta se modela por una parte sistemática que está compuesta por una combinación de características poblacionales que son compartidas por todas las unidades (efectos fijos), y una parte aleatoria que está constituida por efectos específicos de cada unidad (efectos aleatorios) y por el error aleatorio, las cuales reflejan las múltiples fuentes de heterogeneidad y correlación entre y dentro de las unidades.

En estos modelos pueden incorporarse covariables. Las mismas se pueden clasificar en 2 categorías: covariables no variables en el tiempo (CNVT) y covariables variables en el tiempo (CVT). La naturaleza diferente de estas covariables conduce a considerar distintos enfoques para cada una de ellas en el análisis.

Las CNVT son variables independientes que no tienen variación intra-unidad, es decir que el valor de la covariable no cambia para una unidad determinada en el estudio longitudinal. Este tipo de covariables se pueden utilizar para realizar comparaciones entre poblaciones y describir diferentes tendencias en el tiempo.

Las CVT son variables independientes que contienen ambas variaciones, intra y entre unidad, es decir que el valor de la covariable cambia para una unidad determinada a lo largo del tiempo y además puede cambiar para diferentes unidades. Este tipo de covariables tienen los mismos usos que las CNVT, y además permiten describir la relación dinámica entre la CVT y la respuesta. Sin embargo, esta relación puede estar confundida por valores anteriores y/o posteriores de la covariable y en consecuencia esto puede conducir a inferencias engañosas sobre los parámetros del modelo. Esta tesina realiza una introducción a la problemática de incorporar covariables que varían en el tiempo en modelos mixtos para datos longitudinales, presentando diferentes definiciones de las mismas y enfoques metodológicos.

Estos conceptos se aplican a un conjunto de datos que surge del programa de atención y control de pacientes hipertensos de Fundación ECLA llevado a cabo en Rosario durante el período 2014-2019. Este estudio observacional realizó un seguimiento de un grupo de pacientes hipertensos registrando en cada visita el tratamiento farmacológico dado al paciente, los valores de la tensión arterial sistólica (TAS) y la adherencia a dicho tratamiento entre otras características. Uno de los objetivos que persiguió este estudio fue evaluar si la adherencia al tratamiento influye en los valores de la TAS a lo largo del seguimiento. Como la variable adherencia es una CVT, se presentarán diferentes enfoques para incluirla en un modelo longitudinal mixto que pueda explicar el cambio en la tensión arterial sistólica media a lo largo del tiempo.

2. Objetivos

2.1. Objetivo Principal

Presentar diferentes propuestas metodológicas para la incorporación de covariables que varían con el tiempo en modelos mixtos para datos longitudinales.

2.2. Objetivos Específicos

- Definir los tipos de covariables existentes.
- Describir propuestas de incorporación de covariables que varían en el tiempo en los modelos mixtos.
- Aplicar los conceptos vistos al programa de atención y control de pacientes hipertensos de Fundación ECLA.

3. Datos Longitudinales

Los datos longitudinales están conformados por mediciones repetidas de una misma variable realizadas a la misma unidad en diferentes momentos o condiciones experimentales.

Dado que las mediciones repetidas son obtenidas de la misma unidad, los datos longitudinales están agrupados. Las observaciones dentro de un mismo agrupamiento generalmente están correlacionadas positivamente. Por lo tanto, los supuestos usuales de independencia y homogeneidad de variancias no son válidos

Existen tres fuentes potenciales de variabilidad que influyen sobre la correlación entre medidas repetidas:

- *Heterogeneidad entre las unidades:* Refleja la propensión natural de las unidades a responder. Las unidades tienen diferentes reacciones frente a los mismos estímulos.
- *Variación biológica intra-unidad:* Se espera que la secuencia de medidas repetidas de una unidad tenga un comportamiento determinado, que produce que las mediciones más cercanas sean más parecidas entre sí que las más alejadas.
- *Error de medición:* Errores aleatorios asociados al proceso de medición.

Estas tres fuentes de variación pueden clasificarse en *"variabilidad entre unidades"* (heterogeneidad entre unidades) y *"variabilidad intra unidades"* (variación biológica intra-unidad y error de medición)

Dado que estas fuentes de variabilidad introducen correlación entre las mediciones repetidas, no se pueden utilizar las técnicas habituales, ya que llevarían a inferencias incorrectas sobre los parámetros del modelo.

Con el fin de simplificar la notación, se asumirá que los tiempos de medición son los mismos para todas las unidades y que no hay datos faltantes.

Sean Y_{ij} el valor de la variable respuesta y X_{ij} las covariables, medidas sobre la unidad i en la ocasión j, se obtiene una muestra de N unidades cada una con n mediciones repetidas de la variable en estudio, observadas en los tiempos $t_1, t_2, ..., t_n$, siendo entonces el número total de observaciones $N^* = Nn$, con $i = 1, ..., N; j = 1, ..., n$.

Se asume que Y_{ij} y X_{ij} son simultáneamente medidas. Esto quiere decir que en un análisis de corte transversal, Y_{ij} y X_{ij} se correlacionan directamente. Sin embargo, para un análisis longitudinal se debe asumir que existe un orden pre-establecido: $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), ..., (X_{in}, Y_{in})$

4. Covariables en datos longitudinales

En los estudios longitudinales, las variables independientes pueden ser clasificadas en dos categorías: CNVT y CVT. La diferencia entre ellas puede conducir a diferentes enfoques de análisis así como también a diferentes conclusiones.

Tanto las CNVT como las CVT se pueden utilizar para realizar comparaciones entre poblaciones y describir diferentes tendencias a lo largo del tiempo. Sin embargo, sólo las CVT permiten describir una relación dinámica entre la covariable y la variable respuesta.

4.1. Covariables fijas en el tiempo

Las CNVT son variables independientes que no presentan variación intra-unidad, es decir, los valores de estas covariables no cambian a lo largo del estudio para una unidad en particular.

Éstas covariables pueden ser fijas por naturaleza (por ejemplo, el sexo biológico de una persona o el grupo de tratamiento) o pueden ser covariables basales (es decir, medidas al inicio del estudio). Las covariables basales son fijas por definición pero pueden ser variables en el tiempo por naturaleza, por ejemplo, la edad varia en el tiempo pero la edad basal es fija.

4.2. Covariables variables en el tiempo

Las CVT son variables independientes que incluyen tanto la variación intra-unidad como la variación entre-unidad. Esto significa que, para una unidad en particular, el valor de la covariable cambia a través del tiempo y puede cambiar también entre diferentes unidades. Por ejemplo, el valor del colesterol o la condición de fumador (si/no).

A continuación se describen diferentes tipos de CVT.

4.2.1. Covariables estocásticas y no estocásticas

Las CVT pueden clasificarse en estocásticas y no estocásticas. Las CVT no estocásticas son covariables que varían sistemáticamente a través del tiempo pero son fijas por diseño del estudio o bien su valor puede predecirse. En cambio, las CVT estocásticas son covariables que varían aleatoriamente a través del tiempo, es decir, los valores en cualquier ocasión no pueden ser estimados ya que son gobernados por un mecanismo aleatorio. Ejemplos de las primeras son: tiempo desde la visita basal o edad. Ejemplos de las segundas son: valor del colesterol, ingesta de alcohol (si/no), ingesta de grasas, etc.

4.2.2. Covariables exógenas y endógenas

Otra clasificación de las CVT es en exógenas y endógenas.

Covariables exógenas

Una CVT se define como exógena, respecto a la variable respuesta, si el valor de la covariable en un determinado momento es condicionalmente independiente de todos los valores precedentes de la variable respuesta (Diggle et al., 2002). Formalmente, para la unidad i en la ocasión j:

$$f(X_{ij}|X_{i1},...,X_{ij-1},Y_{i1},...,Y_{ij}) = f(X_{ij}|X_{i1},...,X_{ij-1}) \tag{4.2.1}$$

Y en consecuencia:

$$E(Y_{ij}|X_{i1},...,X_{in}) = E(Y_{ij}|X_{i1},...,X_{ij}) \tag{4.2.2}$$

Esta definición implica que la respuesta en cualquier momento puede depender de los valores previos de la variable respuesta y de la CVT, pero será independiente de todos los demás valores de la covariable. Por ejemplo, en un estudio longitudinal en donde se evalúa si el la cantidad de actividad física (variable explicativa) está asociada al nivel de glucosa en sangre (variable respuesta), es de esperar que la cantidad de actividad física en una determinada ocasión, afecte a niveles posteriores de nivel de glucosa en sangre, pero no se espera que dependa de los niveles glucosa en sangre observados previamente en el sujeto afecten a la cantidad de actividad física.

Es posible examinar empíricamente la suposición de que una CVT es exógena al considerar modelos de regresión para la dependencia de X_{ij} en $Y_{i1},...,Y_{ij-1}$ (o en alguna función conocida de $Y_{i1},...,Y_{ij-1}$) y $X_{i1},...,X_{ij-1}$ (o en alguna función conocida de $X_{i1},...,X_{ij-1}$). La ausencia de cualquier relación entre X_{ij} y $Y_{i1},...,Y_{ij-1}$, dado el perfil de la covariable anterior $X_{i1},...,X_{ij-1}$, proporciona soporte para la validez de la suposición de que la CVT es exógena.

Cuando se puede asumir que las CVT son exógenas con respecto a la variable respuesta, se puede dar una interpretación causal a los parámetros de regresión.

Covariables endógenas

Una CVT que no es exógena se define como endógena. Una variable endógena es una variable estocásticamente relacionada con otros factores medidos en el estudio. Esta también puede definirse como una variable generada por un proceso estocástico relacionado con el individuo en estudio. En otras palabras, las CVT endógenas están asociadas con un efecto individual y, a menudo, pueden explicarse por otras variables en el estudio. Cuando el proceso estocástico de una CVT endógena puede ser (al menos parcialmente) explicado por la variable respuesta, se dice que hay *feedback* entre la respuesta y la CVT endógena. Por ejemplo, cuando se evalúa si la cantidad de actividad física está asociada al nivel de glicemia. El nivel de actividad física en un determinado momento puede estar (o no) asociado a niveles previos y también puede estar asociado a valores previos de glicemia (un paciente con valor de glicemia alto en una visita puede decidir aumentar su nivel de actividad fística para ver si este valor se reduce).

5. Modelo lineal mixto

Los modelos lineales mixtos se utilizan habitualmente para analizar los datos longitudinales, debido a que permiten modelar las distintas fuentes de variabilidad presentes en los mismos.

En estos modelos, la respuesta media se modela como una combinación de características poblacionales que son comunes a todos los individuos (efectos fijos) y efectos específicos de la unidad que son únicos de

ella (efectos aleatorios).

El modelo lineal mixto para la unidad i se puede expresar en forma matricial como:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i; \quad i = 1, ..., N;$$

Donde:

- $\boldsymbol{Y}_i$: Vector de la variable respuesta de la i-ésima unidad, de dimensión $(n \times 1)$, siendo $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ...Y_{in})'$
- $\boldsymbol{X}_i$: Matriz de diseño de la i-ésima unidad, que caracteriza la parte sistemática de la respuesta, de dimensión $(n \times p)$
- $\boldsymbol{\beta}$: Vector de parámetros de dimensión $(p \times 1)$
- $\boldsymbol{Z}_i$: Matriz de diseño de la i-ésima unidad, que caracteriza la parte aleatoria de la respuesta, de dimensión $(n \times k)$
- $\boldsymbol{b}_i$: Vector de efectos aleatorios de la i-ésima unidad, de dimensión $(k \times 1)$
- $\boldsymbol{\varepsilon}_i$: Vector de errores aleatorios de la i-ésima unidad, de dimensión $(n \times 1)$

Se supone que $\boldsymbol{\varepsilon}_i$ y $\boldsymbol{b}_i$ son independientes.

$$\boldsymbol{\varepsilon}_i \sim N_n(0, \boldsymbol{R}_i) \quad \boldsymbol{b}_i \sim N_k(0, \boldsymbol{D})$$

$\boldsymbol{D}$ y $\boldsymbol{R}_i$ son las matrices de variancias y covariancias de los vectores $\boldsymbol{b}_i$ y $\boldsymbol{\varepsilon}_i$ respectivamente. A partir de este modelo se obtiene:

- $E(\boldsymbol{Y}_i/\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ (media condicional o específica de la i-ésima unidad)
- $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ (media marginal)
- $Cov(\boldsymbol{Y}_i/\boldsymbol{b}_i) = \boldsymbol{R}_i$ (variancia condicional)
- $Cov(\boldsymbol{Y}_i) = \boldsymbol{Z}_i\boldsymbol{D}_i\boldsymbol{Z}_i' + \boldsymbol{R}_i = \boldsymbol{\Sigma}_i$ (variancia marginal)

Generalmente, la matriz $\boldsymbol{D}$ adopta una estructura de covariancia arbitraria, mientras que la matriz $\boldsymbol{R}_i$ adopta otra estructura que modela apropiadamente la variabilidad intra individuo.

5.1. Estimación de los parámetros del modelo

Bajo el supuesto de que $\boldsymbol{\varepsilon}_i$ y $\boldsymbol{b}_i$ se distribuyen normalmente se pueden usar métodos de estimación basados en la teoría de máxima verosimilitud, cuya idea es asignar a los parámetros el valor más probable en base a los datos que fueron observados. Se usarán para estimar los parámetros de la parte media y los de las estructuras de covariancia los métodos de máxima verosimilitud (ML) y máxima verosimilitud restringida (REML) respectivamente

5.1.1. Método de máxima verosimilitud (ML)

Bajo el supuesto de que $\boldsymbol{Y}_i \sim N_n(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ y las $\boldsymbol{Y}_i$ son independientes entre sí, se obtiene la siguiente función de log-verosimilitud:

$$l = -\frac{1}{2}\sum_{i=1}^{N} n\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{N}[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})] \tag{5.1.1}$$

Siendo $\boldsymbol{\Sigma}_i$ función del vector $\boldsymbol{\theta}$ que contiene los parámetros de covariancia.

Los estimadores de $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ son los valores que maximizan esta expresión. Cuando $\boldsymbol{\theta}$ es desconocido (lo que generalmente sucede) se obtiene una ecuación no lineal, por lo que no se puede obtener una expresión explícita de $\hat{\boldsymbol{\theta}}$. Para encontrar su solución se recurre a métodos numéricos. El estimador del vector $\boldsymbol{\beta}$ resulta:

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{X}_i)^{-1}\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{Y}_i$$

El estimador $\hat{\boldsymbol{\beta}}$ resulta insesgado de $\boldsymbol{\beta}$. Cuando $\boldsymbol{\theta}$ es desconocido no se puede calcular de manera exacta la matriz de covariancias de $\hat{\boldsymbol{\beta}}$. Si el número de unidades es grande se puede demostrar que asintóticamente (Fitzmaurice et al., 2004):

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \boldsymbol{V}_{\boldsymbol{\beta}}) \quad \textit{donde} \quad \boldsymbol{V}_{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{X}_i)^{-1}$$

5.1.2. Método de máxima verosimilitud restringida (REML)

El inconveniente que posee el método de ML es que los parámetros de covariancia resultan sesgados. Es decir, a pesar de que $\hat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$, no pasa lo mismo con $\hat{\boldsymbol{\theta}}$. Si el tamaño de muestra es chico, los parámetros que representan las variancias van a ser demasiado pequeños, dando así una visión muy optimista de la variabilidad de las mediciones, es decir, se subestiman los parámetros de covariancia. El sesgo se debe a que en la estimación ML de $\boldsymbol{\theta}$ no se tiene en cuenta que $\boldsymbol{\beta}$ es estimado a partir de los datos.

Distintos autores proponen el método de REML para estimar los parámetros del modelo. Este método es una modificación del método de máxima verosimilitud, en el que la parte de los datos usada para estimar $\boldsymbol{\beta}$ está separada de aquella usada para estimar los parámetros de $\boldsymbol{\Sigma}_i$. La función de log-verosimilitud restringida que se propone es:

$$\boldsymbol{l}^* = -\frac{1}{2}\sum_{i=1}^{N} n\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{N}[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})] - -\frac{1}{2}\ln|\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}_i^{-1}}\boldsymbol{X}_i| \tag{5.1.2}$$

Maximizando esta funcion con respecto a $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ se obtiene:

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{X}_i)^{-1}\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{Y}_i$$

Donde $\hat{\boldsymbol{\Sigma}}_i$ es el estimador REML de $\boldsymbol{\Sigma}_i$ (Fitzmaurice et al., 2004).

5.1.3. Problemas con la estimación

Pepe y Anderson (1994) mostraron que las ecuaciones 5.1.1 y 5.1.2 llegan a cero solo si se cumple con el supuesto de independencia condicional:

$$E[Y_{ij}|X_{ij}] = E[Y_{ij}|X_{ij}, j = 1, ..., n] \tag{5.1.3}$$

Con las CNVT, esta suposición se mantiene necesariamente ya que $X_{ij} = X_{ik}$ para todas las ocasiones $k \neq j$. Con las CVT estocásticas, que se fijan por diseño del estudio (por ejemplo, indicador de grupo de tratamiento en una prueba cruzada), la suposición también se cumple ya que los valores de las covariables en cualquier ocasión se determinan a priori por diseño del estudio y de manera completamente no relacionado con la respuesta longitudinal. Sin embargo, cuando una covariable es variable en el tiempo no estocástica, puede que no necesariamente se mantenga.

En general, cuando (5.1.3) no se cumple, los valores precedentes y/o posteriores de la CVT confunden la relación entre Y_{ij} y X_{ij}, esto puede llevar a estimaciones sesgadas de los parámetros del modelo.

Frente a este escenario, Pepe y Anderson (1994) recomendaron plantear el modelo longitudinal marginal y realizar las estimaciones mediante GEE (ecuaciones de estimación generalizadas) con estructura de correlación independiente ya que este es siempre consistente. La estructura de correlación independiente generalmente tiene una alta eficiencia para la estimación de los coeficientes asociados a CNVT. Sin embargo, para las CVT, Fitzmaurice (1995) muestra que esta estructura puede resultar en una pérdida sustancial de eficiencia para la estimación de los coeficientes asociados a las CVT y proporciona un ejemplo en el que la elección de dicha estructura tiene una eficiencia del 60 % en relación con la estructura de correlación verdadera.

Luego, Lai y Small (2007) y Lalonde (2014) definieron 4 nuevos tipos de CVT:

- Tipo I: una CVT es de Tipo I si no hay relación entre la CVT y la variable respuesta en diferentes ocasiones. Las variables que involucran cambios predecibles a lo largo del tiempo, como la edad o el tiempo de observación, generalmente se tratan como Tipo I.

- Tipo II: una CVT es de Tipo II si no está asociada a valores anteriores de la variable respuesta, pero la variable respuesta si puede estar asociada a valores previos de la CVT. Un ejemplo de este tipo de CVT puede ser la medicación para la presión arterial con respecto a la variable respuesta presión arterial, ya que valores acumulados de la medicación en el tiempo se espera que tengan un impacto en la presión arterial en cualquier ocasión.

- Tipo III: para éste tipo, no hay suposición de independencia entre la respuesta y los valores de la CVT en diferentes ocasiones. Por lo tanto, una CVT de Tipo III puede implicar un ciclo de feedback entre

la CVT y la respuesta, en el que los valores de la covariable pueden verse afectados por los valores anteriores de la respuesta. Un ejemplo de este tipo es la CVT medicación para la presión arterial con la respuesta infarto de miocardio. Mientras que es esperado que la medicación impacte en la probabilidad de infarto, un evento de infarto puede provocar un cambio en la medicación para la presión arterial.

- Tipo IV: para una CVT de Tipo IV, la covariable puede estar asociada a valores previos de la variable respuesta, pero la respuesta no está asociada a valores previos de la CVT. Un ejemplo es la CVT presión arterial con la variable respuesta peso. Si bien existe una asociación entre el peso y la presión arterial, la dirección del efecto parece ser que el peso afecta la presión arterial, pero es poco probable que ocurra lo contrario.

Además, propusieron utilizar el "Método generalizado de los momentos" (GMM) (Hansen, 1982). Éste método puede ser utilizado para tratar a cada CVT de manera diferente, dependiendo del tipo de cada una, y evita problemas con la estimación de ecuaciones construidas a partir de componentes no independientes.

En conclusión, si la CVT es exógena puede introducirse tanto a un modelo estimado con GMM como a un modelo lineal mixto de manera tradicional. Sin embargo, si la CVT es endógena, no puede introducirse al modelo lineal mixto y debe definirse uno de sus 4 tipos para utilizar el GMM.

6. Formas de introducir una CVT al modelo

Si al evaluar el tipo de la CVT de la manera vista en 4.2.2 resulta ser exógena, se puede introducir en el modelo lineal mixto sin consideraciones adicionales. Esto se debe a que no habrá problemas con la estimación de los parámetros, ya que se cumple el supuesto de independencia condicional.

A continuación, se presentan distintas maneras de introducir la CVT exógena al modelo lineal mixto. De manera de ejemplo, se tomará un caso en el que la variable respuesta Y_{ij} y la CVT X_{ij} son la tensión arterial y el IMC, respectivamente, del paciente i en la ocasión j.

6.1. Convertirla en CNVT

Una solución rápida al problema de las CVT, sea exógena o endógena, es transformarla en una CNVT, esto se puede lograr resumiendo la información de la misma mediante alguna función como el promedio de los valores de cada individuo y dejarlo fijo a través del tiempo. También podria usarse su valor máximo, mínimo o cualquier transformación que resulte de interés en el estudio. El problema de este enfoque es que se pierde información, dado que se usa una covariable más simple que no refleja la relación dinámica entre la covariable y la respuesta en el tiempo.

En el ejemplo mencionado anteriormente, se podría calcular el IMC promedio de cada uno de los individuos y el modelo resultaría:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 \overline{X}_i + \beta_2 t_j + b_{1i} t_j + \varepsilon_{ij}$$

El coeficiente β_1 se interpreta como el cambio en la tensión arterial por cada cambio unitario en el IMC promedio del individuo.

6.2. Covariable variable en el tiempo

Dado que la CVT es exógena, se puede incorporar al modelo sin ninguna transformación. El modelo resultante es,

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 X_{ij} + \beta_2 t_j + b_{1i} t_j + \varepsilon_{ij}$$

Como se mencionó anteriormente, las CVT contienen efecto intra-unidad y entre-unidad, sin embargo el modelo lineal mixto va a tener solo un coeficiente β_1. Si bien el modelo es válido, para obtener una interpretación clara del efecto de la CVT se necesita dividir su efecto en dos coeficientes, como se explicará en la sección 6.5 (Hoffman, 2015).

6.3. Covariable rezagada

En algunas aplicaciones hay justificacion previa para considerar la covariable en el rezago k momentos antes de la medición de la respuesta. Por ejemplo, el efecto del IMC sobre la tensión arterial probablemente no sea inmediato, por lo que podría interesarnos su valor en la ocasión anterior ($k = 1$). Lo más común es que se desconozca el valor k apropiado y se consideren varias opciones diferentes. El modelo lineal mixto se definiría de la siguiente manera:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_{1k} X_{ij-k} + \beta_2 t_j + b_{1i} t_j + \varepsilon_{ij}$$

En este modelo, el coeficiente β_{1k} depende explícitamente de la elección del rezago k.

6.4. Funcion de las covariables rezagadas

Una alternativa cuando se quiere utilizar la información de las covariables rezagadas es resumir a través de una función la información de éstas en una sola covariable. Un ejemplo puede ser el valor promedio o acumulado hasta la ocasión actual. Sin embargo, la elección de está funcion dependerá del tipo de problema a analizar. Cabe destacar que, al igual que con toda medida resumen, al usar este tipo de covariables se pierde parte de la información. En el ejemplo mencionado, podría ser de interés calcular el IMC promedio hasta la ocasión j, resultando el modelo:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 \overline{X}_{ij} + \beta_2 t_j + b_{1i} t_j + \varepsilon_{ij}$$

Donde $\overline{X}_{ij}$ es el IMC promedio calculado hasta la ocasión j

6.5. Dividiendo efecto entre-unidad y efecto intra-unidad

Otra forma de incorporar la CVT es dividiendo el efecto en dos componentes que reflejen la variación intra-unidad y la variación entre-unidad. Entonces, el coeficiente del modelo que representa a la covariable se puede descomponer en dos:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_W (X_{ij} - \overline{X}_i) + \beta_B \overline{X}_i + \beta_2 t_j + b_{1i} t_j + \varepsilon_{ij}$$

Donde, $\overline{X}_i$ representa el promedio de todos los valores observados en el tiempo de la CVT para la unidad i, β_W representa el cambio esperado en la media de la variable respuesta asociado con cambios de la CVT dentro de la unidad y β_B representa el cambio esperado en la media de la variable respuesta asociado con cambios de la CVT entre unidades

En el ejemplo del IMC y la tensión arterial, β_W se interpreta como el cambio en la tensión arterial por cada diferencia de una unidad en el IMC de la ocasión j comparado con el IMC promedio de la persona, manteniendo constante el IMC de las ocasiones restantes. Por otro lado, β_B se interpreta como el cambio

en la tensión arterial por cada cambio de una unidad en el IMC promedio de la unidad, controlando por el valor del IMC de la ocasión j.

Estos dos efectos permiten medir entonces si la tensión arterial aumenta cuando una persona tiene un IMC mayor que el promedio de la población, y también permiten medir el cambio en la tensión arterial si dicha persona tiene IMC mayor que su promedio habitual.

7. Aplicación

A partir del programa de atención y control de pacientes hipertensos iniciado en el año 2014 en Rosario se obtienen datos de 560 pacientes de entre 30 y 86 años, con una edad media de 58.84 y desvío estándar de 9.87 años, de los cuales un 49.28 % son hombres. A todos estos pacientes se les indicó un tratamiento antihipertensivo al incio del seguimiento (visita basal). Durante 7 meses se agendaron visitas mensuales en donde se registraron, entre otras características, el valor de la TAS y la adherencia al tratamiento. Esta última variable surge de la evaluación del cuestionario Morisky (Morisky et al., 1986) que arroja como resultado: adhiere al tratamiento, no adhiere al tratamiento. Al ser evaluada en todas las visitas mensuales, esta variable dicotómica es una CVT que captura como fue la adherencia durante el periodo desde la visita previa hasta la visita actual. Como para la visita basal no se cuenta con información de adherencia, se toman los datos desde el primer mes de tratamiento.

Uno de los objetivos que persiguió este estudio fue evaluar si la adherencia influye en los valores de la TAS a lo largo del seguimiento.

Para dar respuesta a este interrogante se propuso ajustar un modelo longitudinal de efectos mixtos considerando a la TAS como variable respuesta y la adherencia, sexo y edad como variables explicativas.

Para todas las decisiones de esta sección se utilizará un nivel de significación del 5 %.

7.1. Nomenclatura

A continuación se describen las variables originales que se encuentran en el dataset y sus distintas transformaciones transformaciones.

Siendo $i = 1, ..., 560$ y $j = 1, ..., 7$ se obtienen:

- TAS_{ij}: tension arterial sistólica (mmHg) del paciente i en el mes j.
- $\overline{TAS}_i$: tension arterial sistólica (mmHg) promedio del paciente i a lo largo del seguimiento ($\sum_{k=0}^{n} \frac{TAS_{ik}}{n}$).
- $\overline{TAS}_{ij}$: tension arterial sistólica (mmHg) promedio del paciente i hasta el mes j ($\sum_{k=0}^{j} \frac{TAS_{ik}}{j}$).
- $sexo_i$: sexo del paciente i medido en la ocasión basal (mes 0).
- $edad_i$: edad del paciente i medido en la ocasión basal (mes 0).
- mes_j: meses transcurridos desde el inicio del tratamiento hasta la ocasión j.
- $adherencia_{ij}$: adherencia al tratamiento del paciente i en el mes j (variable dicotómica: =1 si adhiere, =0 si no adhiere).
- $\overline{adherencia_i}$: proporción de visitas en las que el paciente i adhirió al tratamiento a lo largo del seguimiento ($\sum_{k=0}^{n} \frac{adherencia_{ik}}{n}$).

- $\overline{adherencia}_{ij}$: proporción de visitas en las que el paciente i adhiere al tratamiento hasta el mes j ($\sum_{k=0}^{j} \frac{adherencia_{ik}}{j}$).
- $adherencia\ perfecta_i$: variable indicadora, $= 1$ si el paciente i adhirió al tratamiento todos los meses, $= 0$ en otro caso.

7.2. Análisis descriptivo

En esta sección se presentaran diversos gráficos con el fin de describir la población en estudio.

En la figura 7.2.1 se puede observar el grupo de pacientes inició el tratamiento con una TAS promedio de casi 133 mmHg, la cual fue disminuyendo levemente de manera aproximadamente lineal hasta promedio de poco más de 130 mmHg al final del tratamiento.

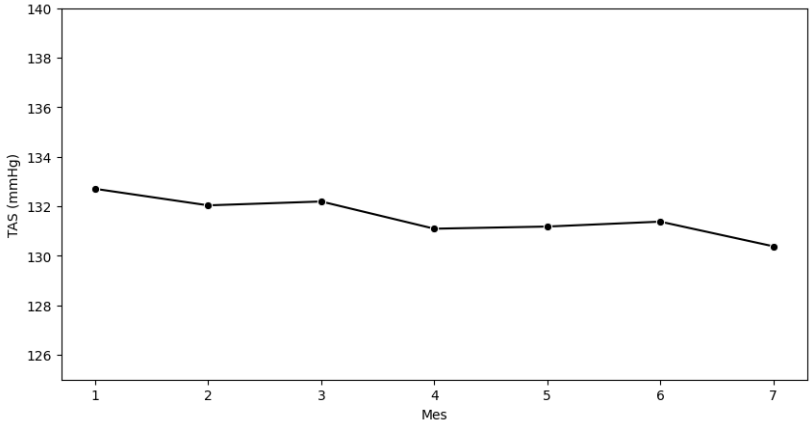

Figura 7.2.1: Evolución de la TAS promedio a lo largo del tratamiento

Otro gráfico que resulta de interés es observar la evolución de la TAS en cada mes sobre cada grupo de las covariables fijas, el resultado es presentado en la figura 7.2.2. Todos los perfiles presentan (en general) una pendiente decreciente de la TAS promedio a través del tiempo, siendo esta menor en cuanto a la ordenada al origen en el grupo de pacientes de mayor edad y de sexo femenino. La covariable edad es continua, pero a fines de poder observar los perfiles promedio de los pacientes se dividió en dos grupos según menores o mayores o iguales 59 años (mediana de edad de los pacientes).

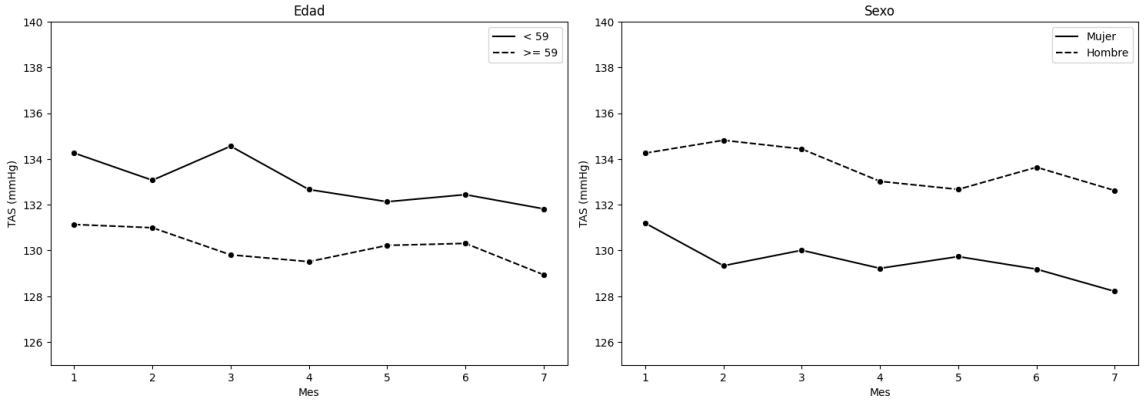

Figura 7.2.2: Evolución de la TAS promedio a lo largo del tratamiento según sexo y edad

Cuando se cuenta con una CVT, cada individuo puede presentar distintos valores de la covariable en cada ocasión. Por lo tanto, para poder visualizar el efecto de la adherencia sobre la TAS según perfiles promedio, es necesario convertir la CVT en una CNVT para que cada paciente pertenezca únicamente a un solo grupo. Para eso, se dividieron distintos perfiles de pacientes según la cantidad de meses que adhirieron al tratamiento, formando los grupos: 3 meses o menos, entre 4 y 6 meses o todos los meses. Se puede observar que para el grupo de pacientes que adhirieron al tratamiento 3 meses o menos la TAS presenta una pendiente creciente a lo largo del estudio. Para los otros 2 grupos, la TAS disminuye a lo largo del tratamiento, siendo menor para el grupo de pacientes que adhirieron la totalidad de los meses.

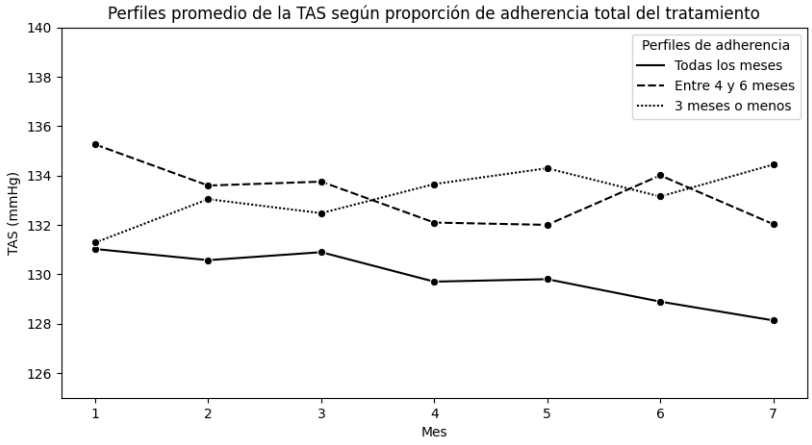

Figura 7.2.3: TAS a través del tiempo según perfiles de adherencia al tratamiento

7.3. Evaluación de la exogeneidad

Para evaluar la exogeneidad de la variable adherencia es necesario ajustar un modelo para cada ocasión en el que se considere a la variable adherencia como variable respuesta y como variables explicativas a la TAS y la adherencia en ocasiones previas. Cómo la CVT adherencia es una variable dicotómica, se ajustan modelos de regresión logística.

Para ajustar estos modelos se utilizarán como covariables a la adherencia y la TAS en el mes anterior,

ya que se asume que las mediciones más cercanas entre sí están más correlacionadas y tambien se utilizarán la adherencia y la TAS promedio desde el inicio hasta 2 meses antes, de esta manera se puede utilizar toda la información del estudio. El modelo resulta:

$$logit(adherencia_{ij}) = \beta_0 + \beta_1\ adherencia_{ij-1} + \beta_2\ TAS_{ij-1} + \beta_3\ \overline{adherencia}_{ij-2} + \beta_4\ \overline{TAS}_{ij-2} + \varepsilon_{ij}$$

Siendo

$$\varepsilon_{ij} \sim N(0, Var(\varepsilon_{ij}))$$

En la tabla 7.3.1 se presentan los valores de los coeficientes para cada covariable y en paréntesis el p-value asociado a cada uno. Como se puede notar, en ninguna ocasión la adherencia depende de valores anteriores de la TAS, por lo tanto puede considerarse como una covariable exógena.

Tabla 7.3.1: Estimación de coeficientes de los modelos logit y sus respectivos p-value

mes (j)	$adherencia_{ij-1}$	$\overline{adherencia}_{ij-2}$	TAS_{ij-1}	$\overline{TAS}_{ij-2}$
2	1,9302 ($<$ 0,001)	–	0,0057 (0,45)	–
3	2,3047 ($<$ 0,001)	0,5683 (0,044)	$-$0,0088 (0,343)	0,0075 (0,419)
4	1,9689 ($<$ 0,001)	1,0734 (0,002)	0,0138 (0,17)	$-$0,017 (0,138)
5	2,2945 ($<$ 0,001)	1,0617 (0,007)	0,0092 (0,441)	$-$0,0141 (0,307)
6	2,2741 ($<$ 0,001)	1,0698 (0,015)	$-$0,0008 (0,938)	$<$ 0,0001 (0,996)
7	2,5812 ($<$ 0,001)	1,4609 (0,003)	$-$0,0005 (0,966)	$-$0,0072 (0,678)

7.4. Modelo lineal mixto propuesto

Mediante la inspección del semivariograma muestral (figura 7.4.1) y pruebas de hipótesis para efectos aleatorios (ver Anexo 9.1) se decidió plantear un modelo de efectos mixtos con ordenada y pendiente aleatoria. De esta manera, se incorpora al modelo la correlación serial y el efecto entre pacientes. La CVT se incorpora en su forma original ya que se evaluó su exogeneidad. El modelo resulta:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ mes_j + \beta_5\ mes_j\ adherencia_{ij} + mes_j\ b_{1i} + \varepsilon_{ij} \tag{7.4.1}$$

Se supone que $\boldsymbol{\varepsilon}_i$ y $\boldsymbol{b}_i$ son independientes.

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in} \end{pmatrix} \sim N_n(0, \boldsymbol{R}_i) \quad \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N_k(0, \boldsymbol{D})$$

Donde $\boldsymbol{D}$ y $\boldsymbol{R}_i$ son las matrices de variancias y covariancias de los vectores $\boldsymbol{b}_i$ y $\boldsymbol{\varepsilon}_i$ respectivamente.

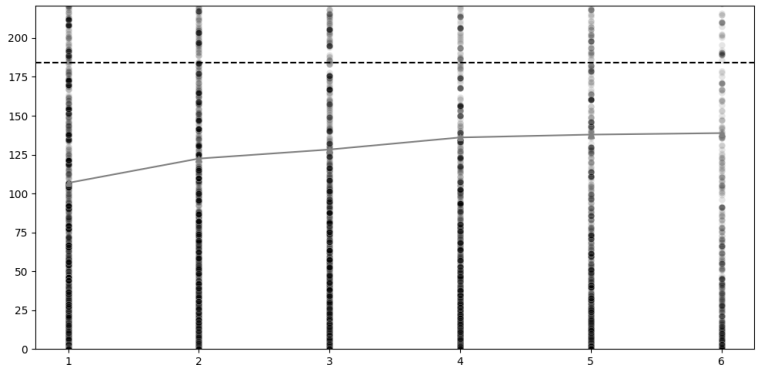

Figura 7.4.1: Semivariograma muestral

Los coeficientes del modelo estimado se presentan en la tabla 7.4.1.

Tabla 7.4.1: Modelo 1

Log-Likelihood			-15390.73	
AIC			30801.47	
BIC			30864.21	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
intercepto	122,270	2,502	48,874	< 0,001
$sexo_i$	3,760	0,747	5,036	< 0,001
$edad_i$	0,167	0,038	4,357	< 0,001
$adherencia_{ij}$	-1,524	1,143	-1,333	0,183
mes_j	0,371	0,240	1,547	0,122
$mes_j\ adherencia_{ij}$	-0,792	0,262	3,018	0,003

Como se mencionó anteriormente, la CVT adherencia contiene ambos efectos entre e intra paciente, por lo tanto su coeficiente no será interpretable y será dificil sacar conclusiones sobre como afecta la adherencia a la TAS. En la siguiente sección se verán distintas formas de incorporar esta CVT al modelo, las cuales algunas si serán interpretables.

7.5. Incorporación de la CVT

Hay más de una manera de incorporar una CVT exógena a un modelo lineal mixto, en esta sección compararemos algunas de ellas.

7.5.1. Incorporación de covariable fija

Hay diversas formas de convertir una CVT en CNVT, en este apartado se mostrarán 2 que resultan de interés para el estudio. Es importante destacar que éstas transformaciones se pueden utilizar para incorporar CVT tanto exógenas como endógenas.

Una de las transformaciones que puede aplicarse sobre la covariable adherencia es convertirla en una variable dicotómica fija, cuyo valor es 1 si el paciente adhirió en todo el tratamiento y 0 en otro caso. En base a los resultados presentados en la tabla 7.5.1, controlando por el resto de las variables, los pacientes que adhieren en la totalidad del tratamiento, inician el tratamiento con una TAS promedio menor en 2,763 unidades, y luego ésta continúa descendiendo en 0,272 unidades por mes.

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia\ perfecta_i + \beta_4\ mes_j + \beta_5\ mes_j\ adherencia\ perfecta_i + b_{1i}\ mes_j + \varepsilon_{ij} \quad (7.5.1)$$

Tabla 7.5.1: Modelo 2: Incorporación adherencia perfecta

Log-Likelihood			-15415.08	
AIC			30850.16	
BIC			30912.9	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercept$	121,926	2,376	51,320	$<$ 0,001
$sexo_i$	3,801	0,740	5,133	$<$ 0,001
$edad_i$	0,176	0,038	4,590	$<$ 0,001
$adherencia\ perfecta_i$	-2,491	1,176	-2,119	0,034
mes_j	-0,197	0,149	-1,322	0,186
$mes_j\ adherencia\ perfecta_i$	-0,272	0,212	-1,280	0,201

Otra manera de incorporar la covariable fija, conservando más información, es usar la proporción de adherencia total al final del estudio. Es decir, si de los 7 meses el paciente adhiere en solo 5, el valor que se le asignará es $\frac{5}{7}$ ($\approx$ 0,71). Los coeficientes presentados en la tabla 7.5.2 indican que, controlando por el resto de las variables, los pacientes presentan una disminución promedio en la TAS de 2,838 en proporción a la adherencia total al inicio del tratamiento y luego disminuirá en 1,010 proporcional a la adherencia total por mes. Por ejemplo, un paciente que adhiere sólo al 50 % del tratamiento, tendrá una disminución promedio en su TAS de 1.419 ($2,838\ *\ 0{,}5$) al inicio del tratamiento, y luego una disminución de 0,505 por mes ($1,010\ *\ 0{,}5$).

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ \overline{adherencia}_i$$
$$+ \beta_4\ mes_j + \beta_5\ mes_j\ \overline{adherencia}_i + b_{1i}\ mes_j + \varepsilon_{ij} \quad (7.5.2)$$

Tabla 7.5.2: Modelo 3: incorporación adherencia total

Log-Likelihood			15417.66	
AIC			30855.32	
BIC			30918.06	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercept$	122,419	3,036	40,318	$<$ 0,001
$sexo_i$	3,732	0,747	4,995	$<$ 0,001
$edad_i$	0,172	0,039	4,469	$<$ 0,001
$\overline{adherencia}_i$	-1,828	2,158	-0,736	0,462
mes_j	0,498	0,380	1,308	0,191
$mes_j\ \overline{adherencia}_i$	-1,010	0,445	-2,270	0,023

7.5.2. Incorporación como CVT

Se proponen otras dos maneras de incorporar la CVT al modelo, la adherencia en el mes anterior y la adherencia promedio hasta el mes actual, ya que podria pensarse que el tratamiento no es de efecto inmediato y la TAS podria estar influenciada por la adherencia en meses anteriores.

Los resultados de ajustar el modelo utilizando la adherencia en el mes anterior se presentan en la tabla 7.5.3.

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij-1}$$
$$+ \beta_4\ mes_j + \beta_5\ mes_j\ adherencia_{ij-1} + b_{1i}\ mes_j + \varepsilon_{ij} \quad (7.5.3)$$

Tabla 7.5.3: Modelo 4: incorporación adherencia en el mes anterior

Log-Likelihood			-15422.50	
AIC			30865.0	
BIC			30927.74	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercept$	121,125	2,388	50,714	$< 0{,}001$
$sexo_i$	3,869	0,751	5,151	$< 0{,}001$
$edad_i$	0,162	0,039	4,197	$< 0{,}001$
$adherencia_{ij-1}$	0,189	0,815	0,232	0,817
mes_j	0,064	0,179	0,359	0,720
$mes_j\ adherencia_{ij-1}$	-0,443	0,215	-2,061	0,039

En la tabla 7.5.4 se presentan los resultados de un modelo ajustando utilizando la adherencia promedio en cada mes. Es decir, si en el mes 4 un paciente adhirio solo en 2 meses, su adherencia promedio será de $\frac{2}{4}$ (0,5). Sin embargo, si adhiere en la ocasión 5, entonces la adherencia promedio en ese mes será $\frac{3}{5}$ (0,6).

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ overlineadherencia_{ij} + \beta_4\ mes_j + \beta_5\ mes_j\ \overline{adherencia}_{ij} + b_{1i}\ mes_j + \varepsilon_{ij} \quad (7.5.4)$$

Tabla 7.5.4: Modelo 5: incorporación adherencia acumulada

Log-Likelihood			-15409.38	
AIC			30838.77	
BIC			30901.51	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercept$	123,107	2,586	47,603	$< 0{,}001$
$sexo_i$	3,721	0,749	4,966	$< 0{,}001$
$edad_i$	0,177	0,039	4,574	$< 0{,}001$
$\overline{adherencia}_{ij}$	-3,247	1,421	-2,285	0,022
mes_j	0,382	0,306	1,248	0,212
$mes_j\ \overline{adherencia}_{ij}$	-0,835	0,355	-2,348	0,019

Al igual que con el modelo propuesto en una primera instancia, estas CVT siguen conteniendo ambos efectos entre e intra paciente, por lo que sus coeficientes no pueden ser fácilmente interpretables.

7.5.3. Incorporación dividiendo efecto entre e intra

Cuando la CVT es dicotómica, con 0 indicando la ausencia del atributo y 1 la presencia, entonces $\overline{X}_i$ es la proporción en la que una persona presentó el valor codificado con 1 de dicha covariable, por lo que el método presentado en la sección 6.5 resultará en valores extraños para β_W si usamos $X_i - \overline{X}_i$. En este caso, si paciente adhirio al tratamiento en el 50 % de los meses, $\overline{X}_i$ tendrá un valor de 0.5 y entonces el término que acompaña a β_W será de -0.5 en los meses que el paciente no adhiera y 0.5 en los meses que si adhiera. En términos de la estimación del modelo esto no genera ningún problema, pero será confuso en la interpretación de los parámetros, dado que el parámetro β_W estará siempre presente (nunca estará acompañado de un 0). Por lo tanto, para evitar esto, el efecto intra-unidad estará acompañado solo de X_{ij}.

Los resultados pueden observarse en la tabla 7.5.5. Los coeficientes -2.658 y 0,603 quieren decir que, manteniendo constante la proporción de adherencia total, se espera que los pacientes que adhieren al tratamiento presenten una disminución promedio en la TAS de 2.658 unidades al inicio del tratamiento y que esta disminuya en promedio 0,603 los meses en los que se adhiere al tratamiento. Por otro lado, los coeficientes -0,137 y -0,525 quieren decir que, después de controlar por la adherencia en ese mes, se espera que la TAS al inicio del tratamiento sea mayor en 0.137 unidades relativo a la proporción de adherencia total y luego disminuya, en promedio, 0,525 unidades en proporcion a la adherencia total al tratamiento.

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ \overline{adherencia}_i + \beta_5\ mes_j + \beta_6\ mes_j\ adherencia_{ij} + \beta_7\ mes_j\ \overline{adherencia}_i + b_{1i}\ mes_j + \varepsilon_{ij} \quad (7.5.5)$$

Tabla 7.5.5: Modelo 6: incorporación la adherencia dividiendo efecto entre e intra persona

Log-Likelihood			-15389.83	
AIC			30803.66	
BIC			30878.95	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercept$	121,987	3,032	40,235	$< 0{,}001$
$sexo_i$	3,723	0,747	4,983	$< 0{,}001$
$edad_i$	0,171	0,039	4,429	$< 0{,}001$
$adherencia_{ij}$	-2,055	1,325	-1,550	0,121
$\overline{adherencia}_i$	0,662	2,784	0,238	0,812
mes_j	0,644	0,379	1,701	0,089
$mes_j\ adherencia_{ij}$	-0,603	0,314	-1,919	0,055
$mes_j\ \overline{adherencia}_i$	-0,525	0,529	-0,992	0,321

7.5.4. Comparación de los métodos

Tabla 7.5.6: Coeficientes estimados con sus respectivos p-values y criterio de Akaike de cada modelo

Modelo	AIC	BIC
Modelo 1	30801.47	30864.21
Modelo 2	30850.16	30912.9
Modelo 3	30855.32	30918.06
Modelo 4	30865.0	30927.74
Modelo 5	30838.77	30901.51
Modelo 6	30803.66	30878.95

En la tabla 7.5.6 se puede observar que, basándose en el AIC y el BIC, los modelos que mejor ajustan los datos son el modelo propuesto y el modelo 5, es decir, los modelos que incorporan la CVT en su manera natural sin aplicarse ninguna transformación. El modelo 5 presenta valores un poco superiores, esto se debe a que se agregan coeficientes que no son significativos pero logran facilitar su interpretación.

8. Conclusiones

Tradicionalmente, los modelos longitudinales fueron pensados para ser ajustados sobre covariables fijas a través del tiempo, pero esto no es algo que suceda siempre en la vida real.

En este informe se ha introducido una manera de categorizar a las covariables variables en el tiempo, específicamente como *exógenas* o *endógenas*, como así también un método para verificar esta clasificación a través de ajustar disintos modelos individualmente para cada ocasión.

Cuando las variables son exógenas pueden añadirse al modelo de la manera tradicional. Sin embargo, se han propuesto diversas transformaciones que pueden ayudar tanto a ajustar de mejor manera los datos como a la interpretación de los coeficientes.

También se mostró un ejemplo con un caso de estudio de pacientes de un programa de atención y control de pacientes hipertensos. En base a estos datos se ajustaron diferentes modelos con las distintas formas mencionadas de introducir la CVT y se realizó una comparación de los mismos.

Como futuros pasos se propone estudiar más en profundidad la incorporación de variables endógenas. Dado que muchas de las técnicas existentes hasta el momento no están basadas en el ajuste de modelos lineales mixtos, quedan fuera del alcance de esta tesina.

9. Anexo

9.1. Elección de efectos aleatorios

Para probar la siginificación de los efectos aleatorios se ajustaron 3 modelos, uno con ordenada y pendiente aleatoria, otro solo con pendiente aleatoria y otro solo con ordenada aleatoria. Los parametros de los 3 modelos fueron estimados con el método de máxima verosimilitud restringida. El modelo completo es el siguiente y sus resultados pueden observarse en la tabla 9.1.1.

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ mes_j + \beta_5\ mes_j\ adherencia_{ij} + b_{1i}\ mes_j + \varepsilon_{ij}$$

Se supone que $\boldsymbol{\varepsilon}_i$ y $\boldsymbol{b}_i$ son independientes.

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in} \end{pmatrix} \sim N_n(0, \boldsymbol{R}_i) \quad \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N_k(0, \boldsymbol{D})$$

Donde $\boldsymbol{D}$ y $\boldsymbol{R}_i$ son las matrices de variancias y covariancias de los vectores $\boldsymbol{b}_i$ y $\boldsymbol{\varepsilon}_i$ respectivamente.

Siendo

$$Var(\boldsymbol{b_i}) = Var \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} = \boldsymbol{D} = \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix}$$

Tabla 9.1.1: Modelo con ordenada y pendiente aleatoria

Log-Likelihood			-15393,93	
Var($intercepto$)			112,78	
Var(mes_j)			2,16	
Cov($Intercepto$, mes_j)			-10,6	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	122,271	2,508	48,751	$< 0{,}001$
$sexo_i$	3,760	0,749	5,022	$< 0{,}001$
$edad_i$	0,167	0,039	4,344	$< 0{,}001$
$adherencia_{ij}$	-1,525	1,144	-1,333	0,183
mes_j	0,371	0,240	1,545	0,122
$mes_j\ adherencia_{ij}$	-0,792	0,263	-3,015	0,003

Se ajusta un modelo reducido solo con pendiente aleatoria, obteniendose los resultados presentados en 9.1.2:

$$Y_{ij} = \beta_0 + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ mes_j + \beta_5\ mes_j\ adherencia_{ij} + b_{1i}\ mes_j + \varepsilon_{ij}$$

Tabla 9.1.2: Modelo con pendiente aleatoria

Log-Likelihood			-15572,14	
Var(mes_j)			2,24	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	122,914	1,946	63,162	$< 0{,}001$
$sexo_i$	3,8	0,566	6,713	$< 0{,}001$
$edad_i$	0,156	0,029	5,361	$< 0{,}001$
$adherencia_{ij}$	-1,539	1,091	-1,41	0,158
mes_j	0,376	0,255	1,474	0,141
$mes_j\ adherencia_{ij}$	-0,798	0,276	-2,886	0,004

Y se plantea el siguiente test de hipótesis:

$H_0) D_{00} = D_{01} = 0 \qquad H_1$) *al menos uno distinto de cero*

Siendo $G^2 = (-2) * LogLikelihood$, MC el modelo completo y MR el modelo reducido obtenemos:

$G^2(MR) - G^2(MC) = 31144{,}28 - 30787{,}86 = 356{,}52$

Al comparar este valor con una $\chi_{50:50;1;0,05} = 5{,}14$ resulta mayor, por lo tanto se rechaza la hipótesis nula y se considera que se debe incluir la ordenada al origen aleatoria.

Al repetir los mismos pasos para un modelo solo con ordenada aleatoria se obtienen los resultados presentados en la tabla 9.1.3:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ mes_j + \beta_5\ mes_j\ adherencia_{ij} + \varepsilon_{ij}$$

Tabla 9.1.3: Modelo con ordenada aleatoria

Log-Likelihood			-15419,49	
Var(*intercepto*)			61,168	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
intercepto	122.102	2.482	49.204	$<$ 0,001
$sexo_i$	3.771	0.752	5.016	$<$ 0,001
$edad_i$	0.165	0.039	4.261	$<$ 0,001
$adherencia_{ij}$	-1.129	1.086	-1.039	0,299
mes_j	0.430	0.224	1.924	0,054
mes_j $adherencia_{ij}$	-0.866	0.250	-3.467	0,001

Se plantea el test de hipótesis:

$H_0) \, D_{11} \; = \; D_{01} \; = \; 0 \qquad H_1 \,) \, al \; menos \; uno \; distinto \; de \; cero$

Siendo $G^2 = (-2) * Log - Likelihood$, MC el modelo completo y MR el modelo reducido obtenemos:

$G^2(MR) - G^2(MC) = 30838{,}98 - 30787{,}86 = 51{,}12$

Al comparar este valor con una $\chi_{50:50;1;0,05} = 5{,}14$ vemos es que mayor, por lo tanto se rechaza la hipótesis nula y se considera que se debe incluir la pendiente aleatoria.

En conclusión, ambos efectos aleatorios son añadidos al modelo.

Bibliografía

[1] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied Longitudinal Analysis.* John Wiley & Sons, 2004.

[2] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data.* Oxford University Press, 2002.

[3] L. Hoffman, *Longitudinal Analysis, Modeling Within-Person Fluctuation and Change.* Routledge, 2015.

[4] R. E. Weiss, *Modeling Longitudinal Data.* Springer, 2005.

[5] T. L. Lalonde, "Modeling time-dependent covariates in longitudinal data analyses," in *Innovative Statistical Methods for Public Health Data* (D.-G. Chen and J. R. Wilson, eds.), ch. 4, pp. 57–79, Springer Cham, 2015.

[6] D. E. Morisky, L. W. Green, and D. M. Levine, "Concurrent and predictive-validity of a self-reported measure of medication adherence," *Medical Care*, vol. 24, pp. 67–74, 1986.

[7] T. L. Lai and D. Small, "Marginal regression analysis of longitudinal data with time-dependent covariates: A generalized method-of-moments approach," *Royal Statistical Society. Series B (Statistical Methodology)*, vol. 69, pp. 79–99, 2007.