Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística

Tesina

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

Autor: **Esteban Cometto**

Directora: Noelia Castellana

Codirectora: Cecilia Rapelli

30 de agosto 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 sobre 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, para el análisis de datos longitudinales no se pueden utilizar las técnicas estadísticas clásicas, sino que se deben utilizar métodos estadísticos especiales que reconozcan las diferentes fuentes de variabilidad presentes en los datos. Los modelos lineales mixtos constituyen la herramienta más utilizada para representar datos correlacionados.

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.

Para introducir los conceptos referentes a los tipos de covariables en los estudios longitudinales se supone que se cuenta con la variable respuesta Y y una sola covariable X. Se obtiene una muestra aleatoria de N unidades, cada una con n mediciones repetidas de la variable respuesta y de la covariable observadas en los tiempos $t_1, ..., t_n$ (se asume que los tiempos de medición son los mismos para todas las unidades). El número total de observaciones es $N* = Nn$.

Sean Y_{ij} y X_{ij} los valores de la variable respuesta y de la variable independiente respectivamente, medidos para la unidad i en la ocasión t_j con $i = 1, ..., N$ y $j = 1, ..., n$. Si se asume que Y_{ij} y X_{ij} son simultáneamente medidas, en un análisis de corte transversal, Y_{ij} y X_{ij} se correlacionarían directamente. Sin embargo, para un análisis longitudinal se debe asumir que existe un orden preestablecido: $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), ..., (X_{in}, Y_{in})$

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. En consecuencia, $X_{ij} = X_i$ para todo $j = 1, ..., n$

Estas 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 varía 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

Las CVT también se pueden clasificar en exógenas y endógenas.

Covariables exógenas

Una CVT estocástica se define como exógena, respecto a la variable respuesta, si el valor de la covariable en un determinado momento, dado los valores previos de la covariable y de la respuesta, 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:

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

Y, en consecuencia (Fitzmaurice et al., 2004):

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

Esta definición implica que la media condicional de la variable respuesta en un determinado momento, dado todo los valores de la covariable (previos y posteriores), sólo depende de los valores previos de la covariable. Por ejemplo, en un estudio longitudinal que evalúa si la cantidad de actividad física (variable explicativa) está asociada al nivel de glucosa en sangre (variable respuesta), según (4.2.2), la media condicional de la glucosa en un determinado momento, dado todos los registros de actividad física (previos y posteriores), sólo depende de los registros previos de actividad física. También, (4.2.1) sugiere que es de esperar que la cantidad de actividad física en una determinada ocasión dependa de la cantidad de actividad física observada en momentos previos, pero no se espera que dependa de los niveles de glucosa observados previamente.

Es posible examinar empíricamente la suposición de que una CVT es exógena, ajustando un modelo de regresión en donde se considera, como variable respuesta a la covariable en un momento determinado (X_{ij}), y como variables explicativas tanto a los valores previos de la covariable ($X_{i1},...,X_{ij-1}$) como a los valores previos de la variable respuesta ($Y_{i1},...,Y_{ij-1}$). Si, después de controlar por los valores previos de la covariable, el valor actual de la covariable no muestra una asociación con los valores previos de la variable respuesta, puede considerarse que la covariable 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. Ésta 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 (Lalonde et al., 2015). Por ejemplo, cuando se evalúa si la cantidad de actividad física está asociada al nivel de glucosa. 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 glucosa (un paciente con valor de glucosa alto en una visita puede decidir aumentar su nivel de actividad física 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:

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

Maximizando esta función 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 sólo si se cumple con el supuesto de independencia condicional:

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

Con las CNVT, esta suposición se mantiene necesariamente ya que $X_{ij} = X_{ik}$ para todo $j, k = 1, ..., n$. Con las CVT no 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 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 (2004) 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.

Lai y Small (2007) y Lalonde et al. (2015) definieron cuatro tipos de CVT y propusieron utilizar el "Método generalizado de los momentos" (Hansen, 1982) en donde es posible incorporar información sobre la naturaleza de la CVT que se está analizando.

En conclusión, si la CVT es exógena, puede introducirse el modelo lineal mixto de manera tradicional y mediante ciertas transformaciones. Sin embargo, si la CVT es endógena, no puede introducirse al modelo

lineal mixto en su formato original y deben explorarse las propuestas planteadas anteriormente.

6. Formas de introducir una CVT exógena al modelo lineal mixto

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 a un modelo lineal mixto con ordenada aleatoria. 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 con $i = 1, ..., N; j = 1, ..., n$.

Para todos los modelos se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$.

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:

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

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:

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

6.3. Covariable rezagada

En algunas aplicaciones hay justificación 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 interesar su valor en la ocasión anterior ($k = 1$). Lo más común es que se desconozca el valor k apropiado y se considere mas de una opción. El modelo lineal mixto se definiría de la siguiente manera:

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

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

6.4. Función de las covariables rezagadas

Una alternativa, cuando se quiere utilizar toda la información con la que se cuenta de la covariable hasta la ocasión actual, es resumir la misma a través de una función. Un ejemplo puede ser el valor promedio o acumulado hasta la ocasión actual. Sin embargo, la elección de está función 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:

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

Donde $\overline{X}_{ij}$ es el IMC promedio calculado hasta la ocasión j para el i-ésimo paciente.

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-unidades respecto de la CVT. Por lo tanto, el término del modelo que representa a la covariable se puede descomponer en dos términos:

$$\beta X_{ij} \rightarrow \beta_W(X_{ij} - \overline{X}_i) + \beta_B\overline{X}_i$$

El modelo lineal mixto queda planteado del siguiente modo:

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

Donde, $\overline{X}_i$ representa el promedio de todos los valores observados en el tiempo de la CVT para la unidad i, es decir, el promedio del IMC para cada paciente. β_W representa el efecto intra-unidad y β_B el efecto entre-unidades.

Cabe destacar que cuando la covariable es dicotómica (que toma valores 0 y 1) la componente $(X_{ij} - \overline{X}_i)$ tomará solamente dos valores y en consecuencia se sugiere dejar esta componente solamente con el valor de X_{ij} sin centrar respecto al valor promedio (Hoffman, 2015).

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 hipertensos de entre 30 y 86 años, con una edad media de 58.84 (desvío estándar 9,9 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), cuyo resultado categoriza a los pacientes como adherentes o no adherentes al tratamiento. Al ser evaluada en todas las visitas mensuales, esta variable dicotómica captura como fue la adherencia durante el período 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 seguimiento.

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 al tratamiento, sexo y edad basal como variables explicativas. Cabe destacar que para este estudio la adherencia al tratamiento es una CVT estocástica, mientras que la edad basal y el sexo son CNVT.

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 variables derivadas que se han construido a partir de las variables originales mediante diferentes transformaciones.

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

- TAS_{ij}: tensión arterial sistólica (mmHg) del paciente i en el mes j.
- $\overline{TAS}_i$: tensión arterial sistólica (mmHg) promedio del paciente i a lo largo del seguimiento ($\sum_{k=1}^{7} \frac{TAS_{ik}}{n}$).
- $\overline{TAS}_{ij}$: tensión arterial sistólica (mmHg) promedio del paciente i hasta el mes j ($\sum_{k=1}^{j} \frac{TAS_{ik}}{j}$).
- $sexo_i$: sexo del paciente i medido como una variable dicotómica (0=mujer, 1=hombre) 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 seguimiento 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 ($\sum_{k=1}^{7} \frac{adherencia_{ik}}{n}$).
- $\overline{adherencia_{ij}}$: proporción de visitas en las que el paciente i adhirió al tratamiento hasta el mes j ($\sum_{k=1}^{j} \frac{adherencia_{ik}}{j}$).
- $adherencia\ perfecta_i$: 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 que luego de un mes de seguimiento (mes 1) la TAS promedio es de aproximadamente 133 mmHg, la cual fue disminuyendo levemente de manera lineal hasta un promedio de aproximadamente 130 al final del seguimiento.

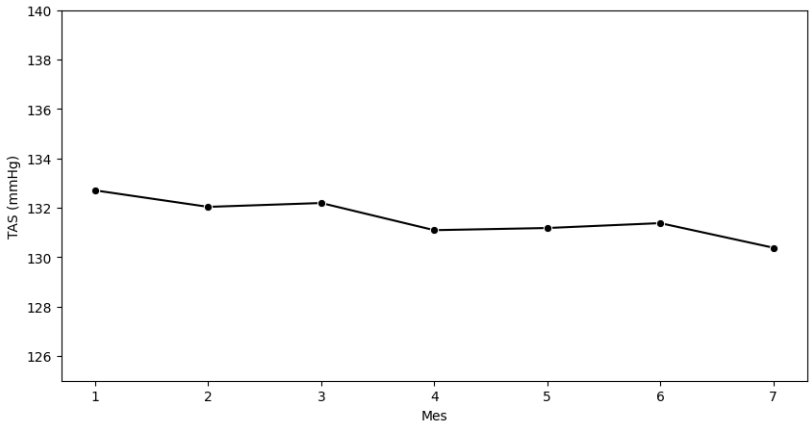

Figura 7.2.1: Evolución de la TAS promedio a lo largo del seguimiento

También resulta de interés observar la evolución de la TAS en el tiempo según la edad basal y sexo de los pacientes. Como la variable edad es continua, para tal fin, se la categorizó en 2 grupos (< 59 años y $>= 59$ años, donde 59 es la mediana de la edad de los pacientes estudiados). En la figura (7.2.2) se observa que los perfiles promedios presentan (en general) una pendiente decreciente, es decir, la TAS media disminuye con el transcurso del seguimiento. Este comportamiento es similar para los grupos, observando valores superiores de TAS para los pacientes de sexo masculino y para los pacientes mayores de 59 años.

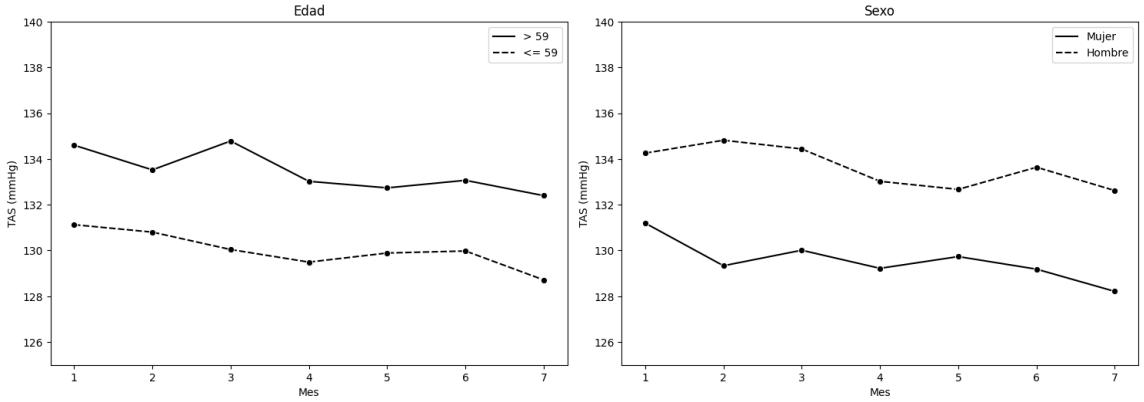

Figura 7.2.2: Evolución de la TAS promedio a lo largo del seguimiento según sexo y edad

Para visualizar la relación entre la adherencia al tratamiento, el tiempo de seguimiento y la TAS media, los gráficos de perfiles promedio no resultan adecuados. La covariable "adherencia al tratamiento" es una CVT y, en consecuencia, para cada individuo, puede presentar distintos valores en cada ocasión, es decir, los pacientes no mantienen un perfil constante a lo largo del tiempo. En una primera instancia, es posible realizar un diagrama de dispersión entre la TAS y el tiempo (mes) según adherencia. Para poder observar con mayor claridad la relación entre estas variables se utiliza la técnina "jitter", la cual agrega un pequeño desplazamiento en los puntos, evitando que queden perfectamente solapados. A partir de la figura (7.2.3), se puede notar que esta manera de visualizar el efecto de la adherencia sobre la TAS resulta confusa.

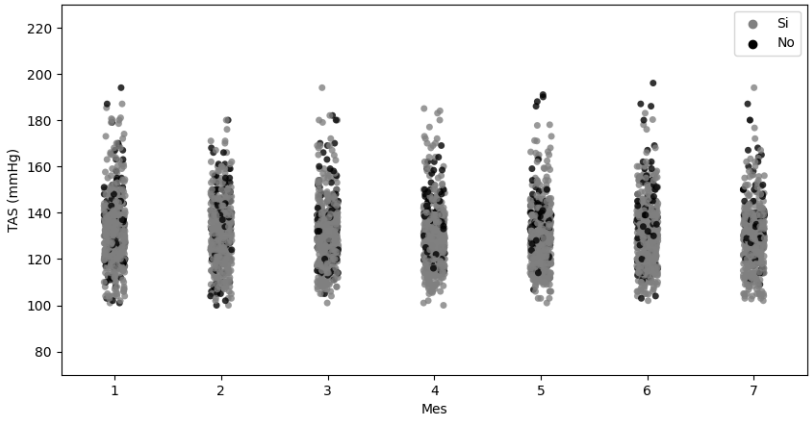

Figura 7.2.3: Valores de la TAS a través del tiempo según adherencia

Como una alternativa a la presentada previamente, se propone realizar un gráfico de perfiles promedio convirtiendo a la CVT (adherencia) en una CVNT, para que de esta manera, cada paciente pertenezca únicamente a un sólo grupo durante todo el período del estudio. Para esto, se dividieron a los pacientes según la cantidad de meses que adhirieron al tratamiento, formando los grupos: 3 meses o menos (adherencia baja), entre 4 y 6 meses (adherencia media/alta) o todos los meses (adherencia perfecta). En la figura (7.2.4) se puede observar que para el grupo de pacientes que adhirieron al tratamiento 3 meses o menos, la TAS presenta una leve pendiente creciente a lo largo del estudio. Para los otros 2 grupos, la TAS disminuye a

lo largo del seguimiento, observándose una disminución mayor para el grupo de pacientes que adhirieron la totalidad de los meses.

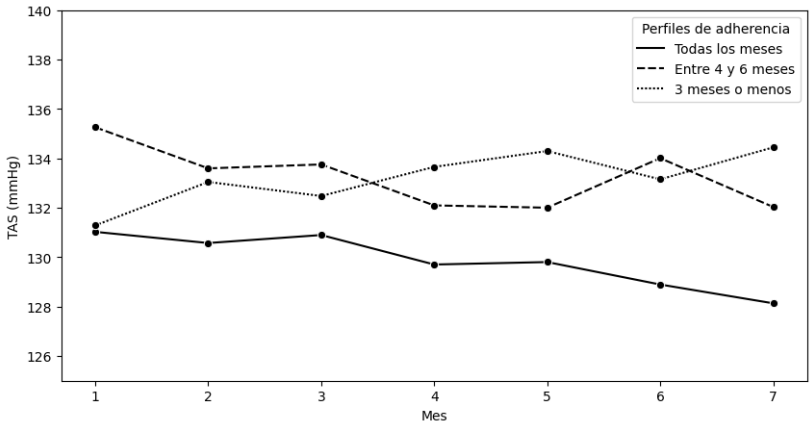

Figura 7.2.4: TAS a través del tiempo según perfiles de adherencia al tratamiento

7.3. Evaluación de la exogeneidad

Para evaluar la exogeneidad de la variable adherencia al tratamiento es necesario verificar el supuesto de independencia condicional descripto en (4.2.2). Esto puede realizarse ajustando 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. Si, en estos modelos, las variables referentes a la TAS registrada en ocasiones previas no resultan ser significativas, puede decirse que la adherencia es una CVT exógena. Cómo la CVT adherencia es una variable dicotómica, se ajustan modelos de regresión logística.

Para ajustar estos modelos se utilizarán como variables explicativas 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 para la ocasión j $(j = 2, ..., 7)$ resulta:

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

En la tabla (7.3.1) se presentan los valores de los coeficientes estimados en cada modelo y entre paréntesis la probabilidad asociada a cada coefiente (referente al test parcial $H_0)\, Bk = 0, k = 1, 2, 3, 4$). Como se puede notar, en ninguna ocasión la adherencia depende de valores anteriores de la TAS (cuando se controla por los valores previos de la adherencia), por lo tanto puede considerarse como una covariable exógena.

Tabla 7.3.1: Estimación de coeficientes de los modelos logit y sus respectivas probabilidades asociadas

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. Modelos propuestos

Se decidió plantear un modelo lineal de efectos mixtos (MLM) con ordenada aleatoria y estructura de covariancia autorregresiva de orden 1. De esta manera, se incorpora al modelo la correlación serial y el efecto entre pacientes presente en los datos (ver anexo). Además, se incorpora el tiempo (mes) y como covariables: el sexo, la edad y la adherencia al tratamiento. Esta última variable, como se determinó en el punto anterior, es una CVT exógena y puede incorporase al modelo de diferentes maneras. A continuación se presentan 6 modelos que surgen de plantear las distintas formas de incorporar la CVT adherencia al tratamiento al MLM.

7.4.1. Incorporación de la CVT como CNVT

Hay diversas formas de convertir una CVT en CNVT. En este apartado se presentarán dos que resultan de interés para el estudio:

- adherencia perfecta (si/no).
- proporción de adherencia durante todo el seguimiento.

Adherencia perfecta

Una de las transformaciones que puede aplicarse sobre la covariable adherencia al tratamiento es convertirla en una variable dicotómica fija, cuyo valor es 1 si el paciente adhirió en todo el seguimiento y 0 en otro caso. El modelo resulta:

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

Se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$, se supone AR(1).

Tabla 7.4.1: Parámetros estimados y medidas de bondad de ajuste del Modelo 1 que incorpora la CVT como Adherencia perfecta (CNVT)

Log-Likelihood			-15401.83	
AIC			30819.66	
BIC			30869.85	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	122,615	2,332	52,574	$<$ 0,001
$sexo_i$	3,776	0,744	5,078	$<$ 0,001
$edad_i$	0,174	0,038	4,533	$<$ 0,001
$adherencia\ perfecta_i$	-3,608	0,745	-4,84	$<$ 0,001
mes_j	-0,341	0,102	-3,352	0,001

En base a los resultados presentados en la tabla (7.4.1), controlando por el resto de las variables, los pacientes que adhieren siempre a lo largo del seguimiento presentan una TAS promedio menor que los pacientes que adhieren a veces o no adhieren en ningún momento. Esta diferencia es constante en el tiempo.

Proporción de adherencia

Otra manera de transformar a la CVT, conservando más información, es calcular la proporción de adherencia al final del seguimiento. Por ejemplo, si un paciente manifestó adherir al tratamiento en 5 de las 7 visitas, entonces la proporción de adherencia es: $5/7 = 0{,}71$. El modelo resulta:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ \overline{adherencia}_i + \beta_4\ mes_j + \varepsilon_{ij} \tag{7.4.2}$$

Se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$, se supone con AR(1).

Tabla 7.4.2: Parámetros estimados y medidas de bondad de ajuste del Modelo 2 que incorpora la CVT como Proporción de adherencia (CNVT)

Log-Likelihood			-15406.67	
AIC			30829.33	
BIC			30879.52	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	125,859	2,599	48,422	$< 0{,}001$
$sexo_i$	3,718	0,751	4,949	$< 0{,}001$
$edad_i$	0,17	0,039	4,392	$< 0{,}001$
$\overline{adherencia_i}$	-5,806	1,583	-3,667	$< 0{,}001$
mes_j	-0,341	0,102	-3,354	0,001

En base a los resultados presentados en la tabla (7.4.2), controlando por el resto de las variables, a mayor proporción de adherencia durante el seguimiento menores valores de TAS promedio.

7.4.2. Incorporación de la CVT sin modificación

Como se observó en el punto (7.3), la CVT adherencia al tratamiento es exógena y, en consecuencia, puede incorporarse en su forma original al modelo. El MLM resulta en:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ mes_j + \varepsilon_{ij} \tag{7.4.3}$$

Se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$, se supone con AR(1).

Tabla 7.4.3: Parámetros estimados y medidas de bondad de ajuste del Modelo 3 que incorpora la Adherencia al tratamiento sin transformar

Log-Likelihood			-15382.71	
AIC			30781.42	
BIC			30831.61	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	124,682	2,376	52,485	$< 0{,}001$
$sexo_i$	3,754	0,751	5,002	$< 0{,}001$
$edad_i$	0,167	0,039	4,321	$< 0{,}001$
$adherencia_{ij}$	-4,423	0,563	-7,851	$< 0{,}001$
mes_j	-0,289	0,101	-2,85	0,004

En base a los resultados presentados en la tabla (7.4.3), controlando por el resto de las variables y para un momento de tiempo determinado, los pacientes que adhieren al tratamiento presentan una TAS promedio menor que los pacientes que no adhieren al tratamiento.

7.4.3. Incorporación de la CVT considerando la covariable rezagada

Puede pensarse que el efecto de la adherencia al tratamiento no es inmediato y entonces se desea considerar la adherencia al tratamiento observada en el mes anterior a la visita (rezago $k = 1$). A continuación, se presenta el MLM resultante:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij-1} + \beta_4\ mes_j + \varepsilon_{ij} \tag{7.4.4}$$

Se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$, se supone con AR(1).

Tabla 7.4.4: Parámetros estimados y medidas de bondad de ajuste del Modelo 4 que incorpora la Adherencia al tratamiento en la visita anterior

Log-Likelihood			-15411.76	
AIC			30839.53	
BIC			30889.72	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
intercepto	121,889	2,367	51,5	$<$ 0,001
$sexo_i$	3,867	0,757	5,111	$<$ 0,001
$edad_i$	0,159	0,039	4,089	$<$ 0,001
$adherencia_{ij-1}$	-0,854	0,481	-1,773	0,076
mes_j	-0,254	0,113	-2,252	0,024

En base a los resultados presentados en la tabla (7.4.4), controlando por el resto de las variables y para un momento de tiempo determinado, no se observan diferencias significativas en la TAS promedio entre los pacientes que adhieren al tratamiento en la visita anterior y los que no adhirieron.

7.4.4. Incorporación de la CVT como función de covariables rezagadas

Puede pensarse que el efecto de la adherencia al tratamiento en la TAS no sólo depende del mes anterior sino también del comportamiento del paciente en las visitas previas. Por lo tanto, se decidió calcular para cada paciente la proporción de adherencia al tratamiento hasta el mes actual. Por ejemplo, si hasta la visita del mes 5 el paciente manifestó adherir al tratamiento en 3 de ellas y en 2 no, la proporción de adherencia al tratamiento al mes 5 es de 3/5. Esta nueva covariable resulta también ser una CVT.

A continuación, se presenta el MLM resultante:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ \overline{adherencia}_{ij} + \beta_4\ mes_j + \varepsilon_{ij} \tag{7.4.5}$$

Se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$, se supone con AR(1).

Tabla 7.4.5: Parámetros estimados y medidas de bondad de ajuste del Modelo 5 que incorpora la Proporción de adherencia al tratamiento hasta la visita actual

Log-Likelihood			-15399.91	
AIC			30815.82	
BIC			30866.01	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	124,98	2,434	51,346	$< 0{,}001$
$sexo_i$	3,739	0,753	4,965	$< 0{,}001$
$edad_i$	0,171	0,039	4,411	$< 0{,}001$
$\overline{adherencia}_{ij}$	-5,12	0,987	-5,187	$< 0{,}001$
mes_j	-0,305	0,102	-2,999	0,003

En base a los resultados presentados en la tabla (7.4.5), controlando por el resto de las variables y para un momento de tiempo determinado, a mayor proporción de adherencia al tratamiento hasta ese momento, se esperan menores valores de TAS promedio.

7.4.5. Incorporación de la CVT, dividiendo su efecto en dos componentes

Como se presentó en el apartado (6.5), el efecto de la CVT puede descomponerse en dos componentes: efecto intra paciente y efecto entre pacientes. Puesto que la variable adherencia es dicotómica, se considera el efecto intra paciente solamente con X_{ij}. El MLM resultante es:

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1\ sexo_i + \beta_2\ edad_i + \beta_3\ adherencia_{ij} + \beta_4\ \overline{adherencia}_i + \beta_5\ mes_j + \varepsilon_{ij} \qquad (7.4.6)$$

Se supone que $\boldsymbol{\varepsilon}_i$ y b_{0i} son independientes.

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

Donde $\boldsymbol{R}_i$, matriz de variancias y covariancias del vector $\boldsymbol{\varepsilon}_i$, se supone con AR(1).

Tabla 7.4.6: Parámetros estimados y medidas de bondad de ajuste del Modelo 6 que incorpora dos componentes para la adherencia al tratamiento

Log-Likelihood			-15382.25	
AIC			30782.5	
BIC			30838.96	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
$intercepto$	125,696	2,599	48,36	$< 0{,}001$
$sexo_i$	3,713	0,751	4,943	$< 0{,}001$
$edad_i$	0,17	0,039	4,388	$< 0{,}001$
$adherencia_{ij}$	-4,22	0,602	-7,008	$< 0{,}001$
$\overline{adherencia_i}$	-1,622	1,692	-0,958	0,338
mes_j	-0,291	0,101	-2,873	0,004

En base a los resultados presentados en la tabla (7.4.6), controlando por el resto de las variables y para un momento de tiempo determinado, la proporción de aherencia en todo el seguimiento no resulta ser signficatva. En cambio, la adherencia en cada visita si lo es, evidenciando que la TAS promedio es menor para los pacientes que adhieren al tratamiento en la visita que para los que no.

7.4.6. Comparación de los modelos propuestos

A continuación, en la tabla 7.4.7, se presentan para cada uno de los 6 modelos ajustados la forma en la que se incorporó la covariable adherencia al tratamiento, el valor estimado del coeficiente que acompaña a la covariable conjuntamente con la probabilidad asociada y los valores de los criterios de información de Akaike (AIC) y Bayesiano de Schwarz (BIC).

Tabla 7.4.7: Resumen de los modelos ajustados

Modelo	Forma de incorporar la adherencia al tratamiento	Parámetro estimado asociado a la adherencia(prob asoc)	AIC	BIC
Modelo 1	Adherencia perfecta (CNVT)	$-3{,}608$ ($< 0{,}001$)	30819,66	30869,85
Modelo 2	Proporción de adherencia (CNVT)	$-5{,}806$ ($< 0{,}001$)	30829,33	30879,52
Modelo 3	Adherencia al tratamiento sin transformar	$-4{,}423$ ($< 0{,}001$)	30781,42	30831,61
Modelo 4	Adherencia al tratamiento en la visita anterior	$-0{,}854$ (0,076)	30839,53	30889,72
Modelo 5	Proporción de adherencia al tratamiento hasta la visita actual	$-5{,}12$ ($< 0{,}001$)	30815,82	30866,01
Modelo 6	Adherencia al tratamiento sin transformar	$-4{,}22$ ($< 0{,}001$)	30782.5	30838.96
	Proporción de adherencia (CNVT)	$-0,291$ (0,338)		

En la tabla (7.4.7) se puede observar que, basándose en el AIC y el BIC, el modelo que mejor ajusta a los datos es el modelo 3 (7.4.3), este modelo incorpora la CVT en su forma natural sin aplicarse ninguna transformación. El segundo modelo que mejor ajusta los datos es el modelo 6 (7.4.6), el cual también incorpora la covariable en su forma natural y además incorpora una transformación CNVT de ésta para dividir su efecto y facilitar la interpretación de sus parámetros.

8. Consideraciones finales

Los datos que surgen de estudios longitudinales pueden ser analizados mediante modelos lineales mixtos. En este tipo de estudios es frecuente contar con variables independientes que pueden ser fijas a lo largo de todo el período o bien puedan variar a lo largo del seguimiento. Estas covariables que varían con el tiempo son variables independientes que incluyen tanto la variación intra-sujeto y la variación entre-sujetos.

La relación entre la covariable que varía en el tiempo y la variable respuesta 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.

En esta tesina se presentaron diferentes definiciones y tipos de covariables que varían en el tiempo, específicamente se introdujo el concepto de covariable endógena y exógena. También se describió una forma empírica de evaluar esta clasificación.

Se presentaron diferentes formas de introducir una covariable que varía en el tiempo exógena al modelo lineal mixto, ya sea en su formato original o bien mediante transformaciones. Estos conceptos fueron aplicados a un conjunto de datos que surge del programa de atención y control de pacientes hipertensos de Fundación ECLA.

Uno de los objetivos que persiguió este estudio observacional fue evaluar si la adherencia al tratamiento (covariable que varía en el tiempo) influía en los valores de la TAS a lo largo del seguimiento. En primer lugar se evaluó la exogeneidad de ésta covariable, luego se presentaron diferentes formas de introducirla al modelo lineal mixto dando lugar a 6 propuestas diferentes. Posteriormente, se compararon estas propuestas y se arribó a un modelo final en el que se incorpora a la covariable adherencia en su formato original sin realizar transformaciones. En este modelo se concluye que, controlando por el resto de las variables y para un momento de tiempo determinado, los pacientes que adhieren al tratamiento presentan una TAS promedio menor que los pacientes que no adhieren al tratamiento.

Se propone como sugerencia para futuros estudios evaluar la incorporación de covariables continuas que varían en el tiempo, ya que en este trabajo la covariable de interés es dicotómica. También se propone evaluar el tratamiento de covariables que varian en el tiempo que resultan ser endógenas al evaluar su condición.

9. Anexo

9.1. Gráficos de interés

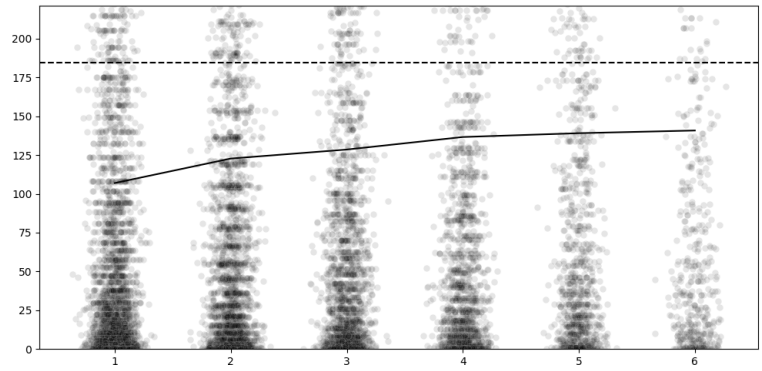

Figura 9.1.1: Semivariograma muestral

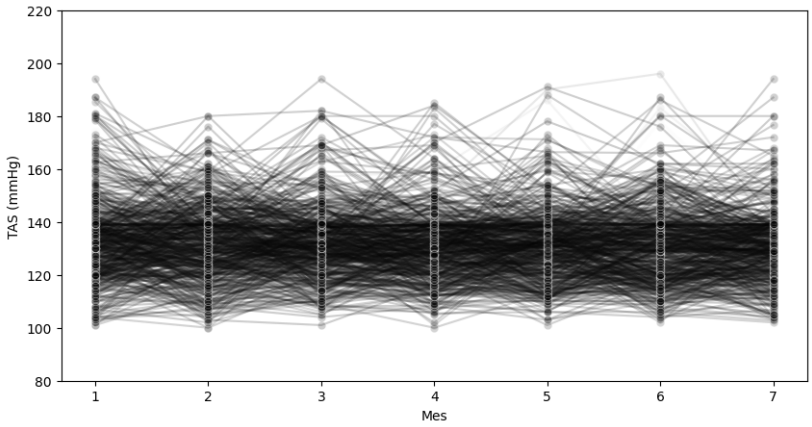

Figura 9.1.2: Evolución de la TAS a lo largo del tiempo para cada paciente

9.2. Código

A continuación se adjunta el código paso a paso en lenguaje Python con el que se realizó esta tesina.

9.2.1. Preprocesamiento

Importar paquetes necesarios.

```
import pandas as pd
```

Leer los datos.

```
df = pd.read_csv("../Datos/tesis_final.csv")
df.head()
```

Crear las transformaciones de las variables.

```
df_grouped_by_paciente = df.groupby("idPaciente")

# Crear Adherencia_Acum: variable de indice de performance de adherencia
al tratamiento hasta el momento t
df["Adherencia_Acumulada"] = (
    df_grouped_by_paciente["Adherencia"]
    .expanding()
    .mean()
    .to_list()
)

# Crear covariable Adherencia_Total no dependiente del tiempo con el
performance final de cada paciente
adherencia_promedio_por_paciente = (
    df_grouped_by_paciente["Adherencia"]
    .mean()
    .to_dict()
)
df["Adherencia_Total"] = df["idPaciente"].map(
    adherencia_promedio_por_paciente
)

# Crear Adherencia Perfecta (1 si adhiere todos los meses, 0 en otro caso)
df["Adherencia_Perfecta"] = (df["Adherencia_Total"] == 1).astype(int)

# Crear TAS_Media_Acum: variable de TAS media hasta el momento t
df["TAS_Media_Acumulada"] = (
    df_grouped_by_paciente["TAS"]
    .expanding()
    .mean()
    .to_list()
)
```

Crear variables rezagadas.

```
df["Adherencia_lag1"] = (
    df_grouped_by_paciente["Adherencia"]
```

```
        .shift(1)
        .fillna(0)
    )

    df["TAS_lag1"] = (
        df_grouped_by_paciente["TAS"]
        .shift(1)
        .fillna(df["tas_basal"])
    )

    df["Adherencia_Acumulada_lag2"] = (
        df_grouped_by_paciente["Adherencia_Acumulada"]
        .shift(2)
        .fillna(0)
    )

    df["TAS_Media_Acumulada_lag2"] = (
        df_grouped_by_paciente["TAS_Media_Acumulada"]
        .shift(2)
        .fillna(df["tas_basal"])
    )
```

Guardar datos preprocesados.

```
df.to_csv("../Datos/tesis_final_preprocesado.csv", index=False)
```

9.2.2. Análisis Descriptivo

Importar paquetes necesarios.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import rpy2.robjects as robjects
```

Setear variables globales de utilidad.

```
plt.rcParams["figure.figsize"] = (10, 5)
```

```
PATH_TO_IMG = "../Tesina/img"
YLIM = (125, 140)
```

Leer los datos.

```
df = pd.read_csv("../Datos/tesis_final_preprocesado.csv")
df_basal = df[df["Mes"] == 1]
```

Valores de interés.

```
prop_hombres = df_basal["Sexo"].mean()
edad_media = df_basal["Edad"].mean()
desvio_edad = df_basal["Edad"].std()
edad_min = df_basal["Edad"].min()
edad_max = df_basal["Edad"].max()
tas_media = df_basal["TAS"].mean()
tas_desvio = df_basal["TAS"].std()
```

Gráfico de perfiles individuales

```
lineplot = sns.lineplot(
    x=df["Mes"],
    y=df["TAS"],
    hue=df["idPaciente"],
    marker="o",
    palette="Greys",
    alpha=0.2
)
plt.ylabel("TAS (mmHg)")
lineplot.get_legend().remove()
plt.ylim((80, 220))
plt.savefig(
    f"{PATH_TO_IMG}/TAS_vs_tpo_perfiles_individuales.png",
    bbox_inches='tight'
)
```

Evolución de la TAS promedio.

```
sns.lineplot(
    x=df["Mes"],
    y=df["TAS"],
    marker="o",
```

```
        color="black",
        ci=None
    )
    plt.ylim(YLIM)
    plt.ylabel("TAS (mmHg)")
    plt.savefig(f"{PATH_TO_IMG}/TAS_vs_tpo.png", bbox_inches='tight')
```

Evolución de la TAS según covariables.

```
    plt.figure(figsize=(15,10))

    # Separar grupos por covariable
    edad_mediana = int(
        df_basal
        .drop_duplicates(["idPaciente", "Edad"])["Edad"]
        .median()
    )
    hue_edad = df["Edad"] > edad_mediana
    hue_edad = hue_edad.map(
        {
            True: f"> {edad_mediana}",
            False: f"<= {edad_mediana}"
        }
    )
    hue_sexo = df["Sexo"].map({0: "Mujer", 1: "Hombre"})

    hues = {
        "Edad": hue_edad,
        "Sexo": hue_sexo
    }

    covs = ["Edad", "Sexo"]
    for i, cov_name in enumerate(covs):
        axs = plt.subplot(int(len(covs)/2)+1,2,i+1)
        line = sns.lineplot(
            x=df["Mes"],
            y=df["TAS"],
```

```
            marker="o",
            hue=hues[cov_name],
            style=hues[cov_name],
            palette=["black", "black"],
            ci=None
        )
        plt.legend()
        plt.ylim(YLIM)
        plt.ylabel("TAS (mmHg)")
        plt.title(cov_name)

    plt.tight_layout()
    plt.savefig(
        f"{PATH_TO_IMG}/TAS_vs_tpo_with_covs.png",
        bbox_inches='tight'
    )
    plt.show()
```

Valores de la TAS a través del tiempo según adherencia.

```
    sns.stripplot(
        x=df["Mes"],
        y=df["TAS"],
        marker="o",
        hue=df["Adherencia"].map({0: "No", 1: "Si"}),
        palette=["gray", "black"],
        alpha=0.8,
    )
    plt.ylabel("TAS (mmHg)")
    plt.ylim((70, 230))
    plt.legend(loc="upper right")
    plt.savefig(
        f"{PATH_TO_IMG}/TAS_vs_tpo_with_adherencia_scatter.png",
        bbox_inches='tight'
    )
```

TAS a través del tiempo según perfiles de adherencia al tratamiento.

```
    # Crear perfiles
```

```
    aux = []
    for value in df["Adherencia_Total"]:
        if value <= 3/7:
            aux.append("3 meses o menos")
        elif value == 1:
            aux.append("Todas los meses")
        else:
            aux.append("Entre 4 y 6 meses")

    df["Perfiles de adherencia"] = aux

    # Plot
    plt.figure(figsize=(10,5))
    sns.lineplot(
        x=df["Mes"],
        y=df["TAS"],
        marker="o",
        hue=df["Perfiles de adherencia"],
        style=df["Perfiles de adherencia"],
        palette=["black", "black", "black"],
        ci=None
    )
    plt.ylim(YLIM)
    plt.ylabel("TAS (mmHg)")
    plt.savefig(
        f"{PATH_TO_IMG}/TAS_vs_tpo_with_adherencia.png",
        bbox_inches='tight'
    )

    df = df.drop("Perfiles de adherencia", axis=1)
```

Semivariograma

```
    # Funciones auxiliares
    def jitter(data, factor=0.1):
        jittered_data = data + np.random.normal(0, factor, len(data))
        return jittered_data
```

```
5
6   def variograma(id, x, y):
7       """
8       id = (nobs x 1) vector con los id
9       y = (nobs x 1) vector respuesta (residuos)
10      x = (nobs x 1) vector de covariables (tiempo)
11
12      RETURN: delta.y = vec( 0.5*(y_ij - y_ik)^2 )
13              delta.x = vec( abs( x_ij - x_ik ) )
14      """
15      uid = id.unique()
16      m = len(uid)
17      delta_y = []
18      delta_x = []
19      did = []
20      for i in range(m):
21          yi = y[id == uid[i]]
22          xi = x[id == uid[i]]
23          n = len(yi)
24          expand_j = [i for i in range(n)]*n
25          expand_k = np.repeat([i for i in range(n)], n)
26          keep = expand_j > expand_k
27          if keep.sum() <= 0:
28              continue
29          expand_j = [j for j, b in zip(expand_j, keep) if b]
30          expand_k = [k for k, b in zip(expand_k, keep) if b]
31          delta_yi = (
32              0.5*(np.array(yi.iloc[expand_j])
33              - np.array(yi.iloc[expand_k]))**2
34          )
35          delta_xi = (
36              abs(np.array(xi.iloc[expand_j])
37              - np.array(xi.iloc[expand_k]))
38          )
39          didi = np.repeat(uid[i], len(delta_yi))
40          delta_y.extend(list(delta_yi))
```

```
            delta_x.extend(list(delta_xi))
            did.extend(list(didi))
        return did, delta_x, delta_y

    # Agregar intercepto porque OLS lo pide
    df["Intercept"] = 1

    # Efectos a usar
    effects = [
        "Intercept",
        "Sexo",
        "Edad",
        "Adherencia",
        "Mes",
    ]

    # Modelo OLS
    model = sm.OLS(df["TAS"], df[effects]).fit()

    # Valores semivariograma
    did, delta_x, delta_y = variograma(
        df["idPaciente"], df["Mes"], model.resid
    )

    # Spline en R
    r_x = robjects.FloatVector(delta_x)
    r_y = robjects.FloatVector(delta_y)
    r_smooth_spline = robjects.r["smooth.spline"]
    spline = r_smooth_spline( r_x, r_y, df=5 )

    # Plot
    plt.figure(figsize=(10,5))
    var = np.var(model.resid)
    fig, ax = plt.subplots(1,1)
    scat = sns.scatterplot(
        x=jitter(delta_x),
```

```
        y=delta_y,
        alpha=0.1,
        color="black"
    )
    plt.axhline(var, color="black", linestyle="--")
    scat.set(ylim=(0, 1.2*var))
    plt.plot(spline[0], spline[1], color="black")
    plt.tight_layout()
    plt.savefig(f"{PATH_TO_IMG}/semivariogram.png", bbox_inches='tight')
```

9.2.3. Prueba de exogeneidad

Importar paquetes necesarios.

```
import pandas as pd
from statsmodels.discrete.discrete_model import Logit
```

Leer los datos.

```
df = pd.read_csv("../Datos/tesis_final_preprocesado.csv")
```

Mostrar el resumen de cada modelo logit.

```
max_range = df["Mes"].max()+1
df["Intercept"] = 1

for i in range(2, max_range):
    df_aux = df[df["Mes"] == i]

    print("\n", "*"*50, "Target time:", i, "*"*50, "\n")

    if i == 2:
        fixed_effects = [
            "Intercept",
            "Adherencia_lag1",
            "TAS_lag1"
        ]
    else:
        fixed_effects = [
            "Intercept",
```

```
18                  "Adherencia_lag1",
19                  "Adherencia_Acumulada_lag2",
20                  "TAS_lag1",
21                  "TAS_Media_Acumulada_lag2"
22              ]
23
24          lr = Logit(df_aux["Adherencia"], df_aux[fixed_effects]).fit()
25          print(lr.summary())
```

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.

[8] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica*, vol. 50, pp. 1029–1054, 1982.

[9] M. S. Pepe and G. L. Anderson, "A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data," *Communications in Statistics, Part B Simulation and Computation*, vol. 23, pp. 939–951, 1994.