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

20 de octubre de 2022

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 clasficar en 2 categorías: covariables fijas y variables en el tiempo. La naturaleza diferente de estas covariables conduce a considerar distintos enfoques para cada una de ellas en el análisis.

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

Las covariables que varían en el tiempo (CVT) son variables independientes que contienen ambas variaciones, intra y entre sujeto, es decir que el valor de la covariable cambia para un individuo determinado a lo largo del tiempo y además puede cambiar para diferentes sujetos. Este tipo de covariables tienen los mismos usos que las covariables no variables en el tiempo (CNVT) pero además describen 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 con el tiempo en modelos mixtos para datos longitudinales, presentando diferentes definiciones de las mismas y enfoques metodológicos.

Se aplican estos conceptos al Programa de Atención y Control de pacientes hipertensos de Fundación ECLA. Este estudio observacional se se realizó entre 2014 y XXXX en Rosario y realiza un seguimiento de pacientes hipertensos, registrando en cada visita el tratamiento farmacológico dado al paciente, los valores del tensión arterial sistólica (TAS) y la adherencia a dicho tratamiento entre otras características. Uno de los objetivos que persigue este estudio es 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 evaluarán 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 respecto a 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 en un estudio sobre la tendencia de la presión arterial en el tiempo para pacientes que siguen cierto tratamiento.

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. Los individuos tienen diferentes reacciones frente a los mismos estímulos.
- *Variación biológica intra-unidad:* Se piensa que existe algún proceso que genera las respuestas de una unidad, el cual cambia en forma continua y suave a través del tiempo produciendo que las medidas más cercanas sean más parecidas.
- *Error de medición:* Surge debido a los errores de medida.

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)

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.

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 $N^* = Nn$. Se denomina Y_{ij} a la medición sobre la unidad i en la ocasión j, con $i = 1, ..., N; j = 1, ..., n$

A cada unidad se le observan las covariables X_{ij} medidas sobre la unidad i en la ocasión j. 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: $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), ..., (X_{in_i}, Y_{in_i})$

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.

4. Modelos lineales mixtos

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. Cada unidad tiene una trayectoria individual caracterizada por parámetros y un subconjunto de esos parámetros se consideran aleatorios.

La respuesta media es modelada 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:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i; \quad i = 1, ..., N; \quad \boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ...Y_{in_i})'$$

Donde:

- $\boldsymbol{Y}_i$: Vector de la variable respuesta de la i-ésima unidad, de dimensión $(n_i \times 1)$
- $\boldsymbol{X}_i$: Matriz de diseño de la i-ésima unidad, que caracteriza la parte sistemática de la respuesta, de dimensión $(n_i \times p)$
- $\boldsymbol{\beta}$: 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 $(n_i \times k)$
- $\boldsymbol{b}_i$: 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 $(n_i \times 1)$

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

$$\boldsymbol{\varepsilon}_i \sim N_{n_i}(0, \boldsymbol{R}_i)$$

$$\boldsymbol{b}_i \sim N_k(0, \boldsymbol{D})$$

$\boldsymbol{D}$ 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 cualquier otra estructura.

4.1. 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

4.1.1. Método de máxima verosimilitud (ML)

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

$$l = -\frac{1}{2}\sum_{i=1}^{N} n_i ln(2\pi) - \frac{1}{2}ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{N}[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})] \tag{4.1.1}$$

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):

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

4.1.2. 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 $\boldsymbol{\theta}$. 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 MV 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:

$$\boldsymbol{l}^* = -\frac{1}{2}\sum_{i=1}^{N} n_i ln(2\pi) - \frac{1}{2}ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{N}[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})] - -\frac{1}{2}ln|\sum_{i=1}^{N} \boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}_i^{-1}}\boldsymbol{X}_i| \tag{4.1.2}$$

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

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

Donde $\hat{\boldsymbol{\Sigma}}_i$ es el estimador REML de $\boldsymbol{\Sigma}_i$

4.1.3. Problemas con la estimación

Pepe y Anderson (1994) mostraron que las ecuaciones anteriores llegan a cero solo si los datos cumplen con el supuesto:

$$E[\boldsymbol{Y}_{ij}|\boldsymbol{X}_{ij}] = E[\boldsymbol{Y}_{ij}|\boldsymbol{X}_{ij}, j = 1, ..., n] \tag{4.1.3}$$

Esto quiere decir que la media condicional de la variable respuesta en una determinada ocasión, dados todos los valores de la covariable, depende solo del valor de la covariable en esa ocasión. Este supuesto se cumple tanto con covariables no variables del tiempo como con covariables variables en el tiempo no estocáticas. Sin embargo, para las CVT estocásticas puede no necesariamente cumplirse: valores anteriores o posteriores de la CVT pueden confundir la relacion entre la variable respuesta y la CVT en una determinada ocasión. En consecuencia, esto puede conducir a estimaciones sesgadas de los efectos fijos del modelo.

Frente a este escenario, varios autores recomendaron plantear el modelo longitudinal marginal y realizar las estimaciones mediante GEE (ecuaciones de estimación generalizadas) con estructura de correlación independiente o bien utilizar el GMM (método generalizado de los momentos), donde es posible incorporar información sobre la naturaleza de la CVT que se está analizando.

5. Covariables en datos longitudinales

En los estudios longitudinales, las variables independientes pueden ser clasificadas en dos categorías: covariables fijas, es decir que no varían en el tiempo (CNVT) o covariables que varían en el tiempo (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.

5.1. Covariables fijas en el tiempo

Las CNVT son variables independientes que no presentan variación intra-sujeto, es decir, los valores de estas covariables no cambian a lo largo del estudio para un individuo 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.

5.2. Covariables variables en el tiempo

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

A continuación se describen diferentes tipos de CVT.

5.2.1. 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, edad, grupo de tratamiento. Ejemplos de las segundas son: valor del colesterol, ingesta de alcohol (si/no), ingesta de grasas, etc.

5.2.2. Covariables exógenas y endógenas

Otra clasificación de las CVT es en exógenas y endógenas.

Covariables exógenas

Se dice que una CVT es exógena cuando los valores actuales y anteriores de la respuesta en la ocasión $j(\boldsymbol{Y}_{i1}, ..., \boldsymbol{Y}_{ij})$, dados los valores actuales y precedentes de la CVT $(\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{ij})$, no predicen el valor posterior de $\boldsymbol{X}_{ij+1}$. Formalmente, una CVT es exógena cuando:

$$f(\boldsymbol{X}_{ij+1}|\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{ij}, \boldsymbol{Y}_{i1}, ..., \boldsymbol{Y}_{ij}) = f(\boldsymbol{X}_{ij+1}|\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{ij}) \tag{5.2.1}$$

Y en consecuencia:

$$E(\boldsymbol{Y}_i|\boldsymbol{X}_i) = E(\boldsymbol{Y}_i|\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{in_i}) = E(\boldsymbol{Y}_i|\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{ij}) \tag{5.2.2}$$

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 nivel de polución en el aire está asociado a la función pulmonar, es de esperar que el nivel de polución del aire en una determinada ocasión dependa de los niveles observados previamente, pero no se espera que dependa de los niveles de la función pulmonar observados previamente en el sujeto.

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 de respuesta, se dice que hay *feedback* entre la respuesta y la CVT endógena. Este tipo de relación debe tenerse en cuenta en cualquier modelo longitudinal.

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.

6. Transformaciones de interés de CVT

En la mayoria de los casos, se suele utilizar solo la exposición que ocurre antes de la ocasión j para predecir $\boldsymbol{Y}_{ij}$. Sin embargo, en algunas aplicaciones, el historial completo de la covariable $X_{i1}, ..., X_{in_i}$ está disponible y es considerado como potencial predictor de la respuesta. En otras, solo un pequeño subconjunto de las mediciones más recientes son usados, ya que se supone que el efecto en la respuesta

está concentrado en ellas. En cualquier caso, el uso de más de una covariable rezagada puede llevar a predictores altamente correlacionados, lo que lleva a preguntarse sobre la elección de cuantos predictores rezagados utilizar y sobre la estructura de sus coeficientes.

6.1. Una sola covariable rezagada

En algunas aplicaciones hay justificacion previa para considerar la covariable en un el rezago k momentos antes de la medición de la respuesta. Por ejemplo, muchos agentes farmacológicos son rápidamente limpiados del cuerpo, por lo que sólo mantienen efectos por una corta duración. En este caso, si la covariable es exógena, puede ajustarse el modelo mixto sin más consideraciones. Lo más común es que se desconozca el valor k apropiado y se consideren varias opciones diferentes.

6.2. Múltiples covariables rezagadas

La literatura de series de tiempo ha considerado modelos para infinitos o finitos rezagos de la covariable. Dado que los datos longitudinales son series de tiempo cortas, se puede proponer un modelo de menor dimensión para los coeficientes de las covariables rezagadas. En este modelo, se asume que los coeficientes rezagados siguen una función paramétrica suave de orden inferior. Por ejemplo, para un rezago finito k, se puede usar un modelo polinomial de orden p, con $p < k$ para obtener coeficientes de regresión suaves.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij-1} + \beta_2 X_{ij-2} + ... + \beta_k X_{ij-k},$$

$$\beta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2 + ... + \gamma_p j^p$$

A pesar de que estos modelos permiten modelar parsimoniosamente las covariables rezagadas múltiples, la especificación del número de rezagos, kx, y el órden del modelo para el coeficiente, p, deben ser consideradas. Ésto puede realizarse a través de tests para modelos anidados, como el test del cociente de verosimilitud o el test de Wald.

6.3. Funcion de las covariables rezagadas

Una alternativa cuando se quiere utilizar la información de las covariables rezagadas pero quiere mantenerse un modelo parsimonioso, es decir, con el menor número posible de variables independientes, 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.

7. Formas de introducir una CVT al modelo

7.1. Convertirla en CNVT

Una solución rápida al problema de las CVT 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 mucha información, dado que se usa una covariable más simple que no refleja la relación dinámica entre la covariable y la respuesta en el tiempo.

7.2. CVT exógena

Si la CVT es exógena se puede introducir en el modelo en su forma original o con cualquiera de sus transformaciones mencionadas en la sección anterior, dado que no habrá problemas con la estimación de los parámetros, debido a que se cumple el supuesto de independencia condicional y pueden recibir una interpretación causal.

Otra forma de incorporar la CVT exógena es dividiéndola en dos componentes que reflejen la variación intra-sujeto y la variación entre-sujeto respecto a la CVT. Entonces, el término del modelo que representa a la covariable se puede descomponer en dos términos:

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

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

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 dicha covariable, por lo que el método anterior resultará en valores extraños para β_W. Por ejemplo, si la CVT se presentó en el 50 % de las ocasiones, $\overline{X}_i$ tendrá un valor de 0.5 y entonces el término que acompaña a β_W será de -0.5 en las ocasiones que la CVT dicotómica no se presente y 0.5 en las que se presente. En términos de la estimación del modelo esto no genera ningún problema, pero será raro en la interpretación de los parámetros, dado que el parámetro β_W estará siempre presente (nunca estará acompañado de un 0). Una forma de evitar esto es dividiendo el efecto entre-persona e intra-persona de la siguiente manera:

$$\beta X_{ij} = \beta_W X_{ij} + \beta_B \overline{X}_i$$

7.3. CVT endógena

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, por lo que es una opción "segura". 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 es sólo un 60 % eficiente en relación con la estructura de correlación verdadera.

Por otro lado, Lai y Small (2007) propusieron utilizar el "Método generalizado de los momento" (GMM) (Hansen, 1982). En este método de estimación se puede incorporar información sobre la naturaleza de la CVT que se está analizando. Lai y Small (2007) definieron 3 tipos de CVT y luego Lalonde et. al (2014) definieron un cuarto tipo de CVT.

8. Aplicación

Se cuenta con un programa de atención y control de pacientes hipertensos iniciado en el año 2014 en Rosario que realiza un seguimiento exhaustivo de 560 pacientes de entre 30 y 86 años (M=58.84, DS=9.87) y, de los cuales un 49.28 % son hombres. Las visitas se realizaron una vez por mes durante 7 meses desde asignado el tratamiento y en cada una de ellas se registró si la persona estaba adhiriendo correctamente al tratamiento y el valor de la tensión arterial sistólica (TAS) (M=134.10, DS=16.07). Es decir, el seguimiento comenzó a medirse luego de un mes de que los pacientes empezaron el tratamiento.

A fines de centrarse en la CVT se dejaron de lado algunas covariables fijas derivadas de estudios de laboratorio, manteniendo solo algunas covariables sociodemográficas de interés para lograr una estabilidad entre modelos interpretables pero no demasiado complejos, sin perder el objetivo principal de este informe.

8.1. Análisis descriptivo

En esta sección se presentaran diversos gráficos para ayudar a entender un poco mejor la población en estudio.

En la figura 8.1.1 se puede observar que en general la TAS se mantiene constante (o con una muy leve pendiente decreciente) a través del tiempo. Esto a simple vista no resultaría muy alentador, dado que el propósito del tratamiento es disminuir la TAS a niveles más saludables. Sin embargo, como se mencionó anteriormente, los pacientes no adhirieron al 100 % el tratamiento, este efecto es el que estudiaremos más adelante. Cabe destacar que las mediciones de los pacientes son equiespaciadas en el tiempo, las desviaciones del eje en el tiempo se deben a una técnica llamada "jitter" que nos permite mover levemente los puntos en el eje x para poder observar mejor la densidad de los mismos.

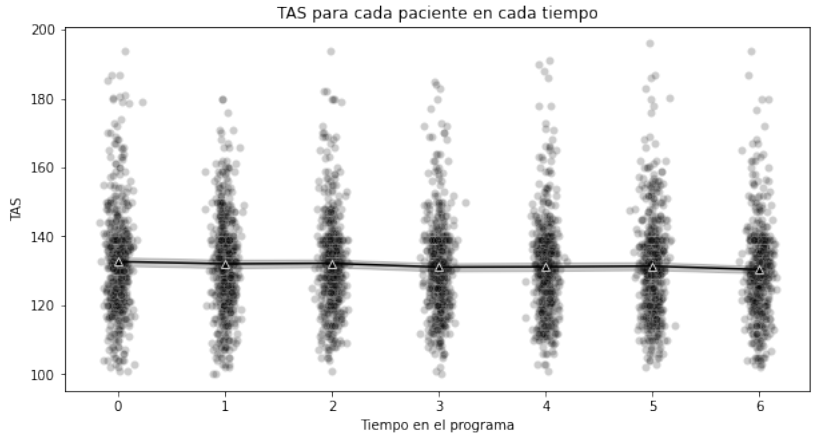

Figura 8.1.1: TAS de cada paciente en cada tiempo

En la figura 8.1.2 se pueden observar las trayectorias individuales de 15 pacientes seleccionados al azar, las pendientes son muy similares entre sí, sin embargo hay variación en la ordenada al origen.

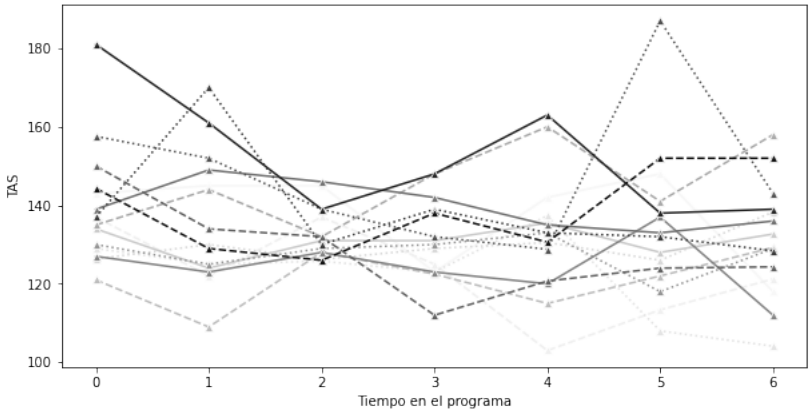

Figura 8.1.2: TAS a través del tiempo de 15 pacientes al azar

Otro gráfico que resulta de interés es observar la evolución de la TAS a través del tiempo pero sobre cada grupo de las covariables fijas, el resultado se expresa en la figura 8.1.3. Para las variables continuas, se utilizó como punto de corte para segmentar en grupos la mediana de sus valores. Podemos observar que en general la TAS disminuye, ya sea en mayor o menor medida, a los largo del tiempo. Analizando las covariables de a una, el IMC parece no tener un efecto significativo, dado que, aunque los pacientes con menor IMC presentan menor TAS, los promedios en cada tiempo caen dentro de los intervalos de confianza de ambos grupos. En cuanto a la edad y el sexo, no parece haber una diferencia significativa en las pendientes pero si parece haber una leve diferencia en las ordenadas al origen, presentando menor TAS los pacientes más jóvenes y de sexo femenino. Por último, los pacientes con antecedentes de diabetes parecen tener una TAS mayor en un comienzo con una pendiente negativa a través del tiempo, mientras que los pacientes que no tienen antecedentes de diabetes comienzan con una TAS mas inferior manteniendola más constante.

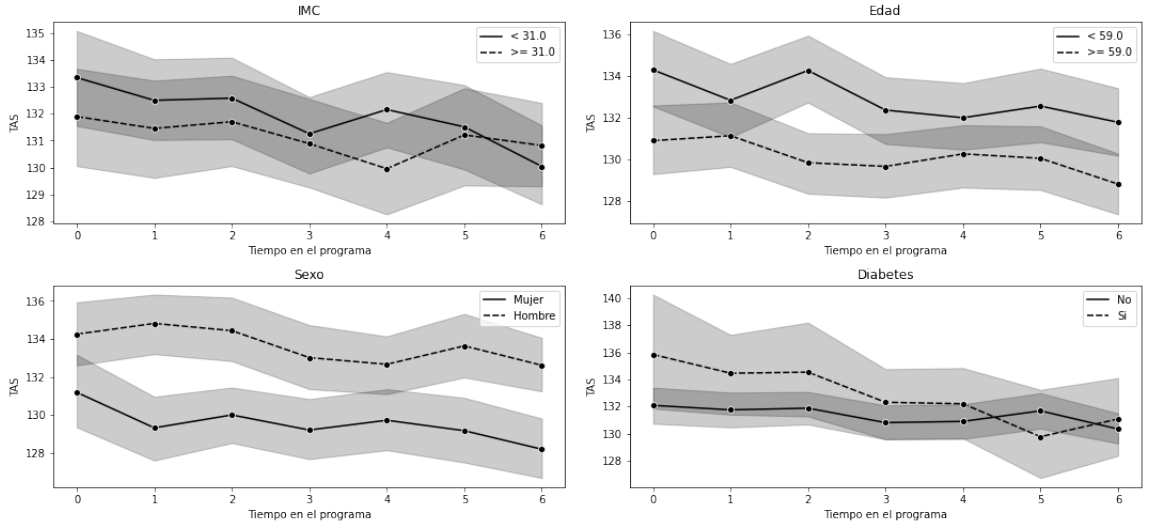

Figura 8.1.3: TAS a través del tiempo según grupos de covariables

Por último, la figura 8.1.4 refleja el efecto de la adherencia al tratamiento sobre la TAS. Para mantener

perfiles no variables en el tiempo, se usa la adherencia total al tratamiento y al igual que antes se divide en base a su mediana, teniendo por un lado los pacientes que adhieren correctamente a más del 86 % del tratamiento (al menos 6 de las 7 ocasiones) y por el otro a los pacientes que adhieren correctamente menos del 86 % del tratamiento (hasta 5 de las 7 ocasiones). Aqui se puede observar que los pacientes que adhieren correctamente al tratamiento parecen tener una pendiente negativa en la TAS a través del tiempo, mientras que los pacientes que no adhieren correctamente mantienen la TAS más constante.

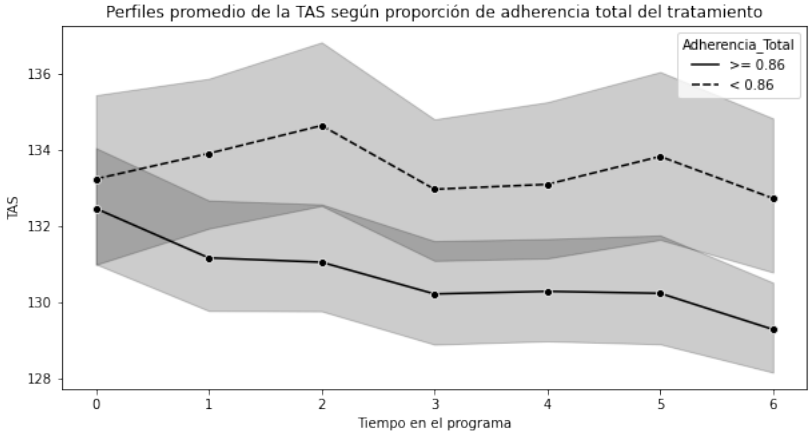

Figura 8.1.4: TAS a través del tiempo según adherencia al tratamiento

8.2. Modelo propuesto

8.2.1. Modelo propuesto para la media

El modelo 8.2.1 se obtuvo mediante el método stepwise, el cual consiste en ajustar un modelo con todas las covariables y en cada paso quitar la de mayor p-value (siempre y cuando no sea significativa) mientras que también se añaden las covariables que no están presentes en el modelo con el fin de volver a agregarlas si alguna tiene un efecto significativo. En el anexo pueden encontrarse todos los pasos de dicho proceso, el modelo obtenido es el siguiente:

$$\begin{aligned}\hat{y}_{ij} = 121{,}011 \; + \; b_{0i} \; + \; 3{,}917 \; sexo_i \; + \; 0{,}156 \; edad_i \; - \; 3{,}337 \; diabetes_i \\ - \; 0{,}238 \; mes_i \; - \; 0{,}689 \; mes_i * diabetes_i \; + \; b_{1i} * mes_i \; + \; \epsilon_{ij}\end{aligned} \quad (8.2.1)$$

8.2.2. Modelo propuesto para la covariancia

En la figura 8.2.1 puede notarse que la mayor parte de la variabilidad total está compuesta por el error de medición dado que la curva no inicia en el cero. Al no haber una pendiente muy pronunciada, la correlación serial también puede considerarse pequeña. Por último, como la curva no llega a la variancia total, esto nos indica que debe explicarse la variancia entre individuos agregando una ordenada aleatoria.

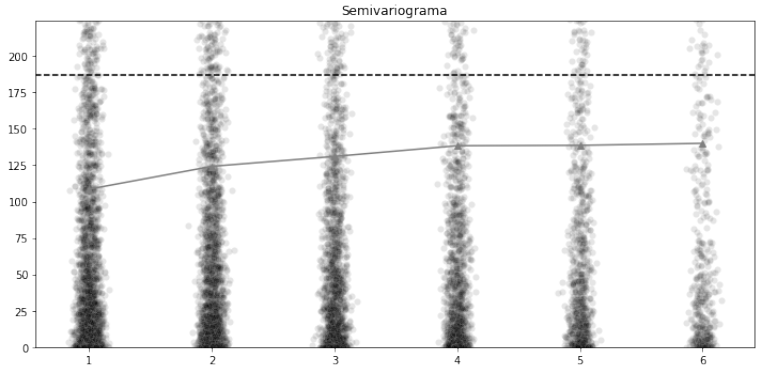

Figura 8.2.1: Semivariograma

Además, se realizaron tests de hipótesis para probar la significación de los efectos aleatorios para la ordenada y la pendiente, resultando ambos significativos. Por lo que la estructura de covariancia elegida para el modelos es una estructura independiente con ordenada y pendiente aleatoria.

Los resultados obtenidos con el modelo propuesto pueden observarse en la tabla 8.2.2.

Tabla 8.2.1: Modelo 1: Modelo propuesto sin CVT

Log-Likelihood			-15424.7921	
AIC			30869.584	
BIC			30932.323	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	121.011	2.357	51.352	< 0.001
Sexo	3.917	0.755	5.187	< 0.001
Edad	0.156	0.039	3.999	< 0.001
DBT	3.337	1.490	2.239	0.025
tpo programa	-0.238	0.114	-2.094	0.036
tpo programa*DBT	-0.689	0.309	-2.230	0.026

8.3. Evaluación de la exogeneidad

Para evaluar la exogeneidad de la adherencia se ajustaron modelos de regresion logística en cada ocasión, usando como variable respuesta la adherencia en dicha ocasión y como covariables la TAS y la adherencia en la ocasión anterior y el promedio y proporción hasta la ocasión anterior. En la tabla 8.3 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 8.3.1: Resultados de la prueba de exogeneidad

Ocasión (t)	x_{t-1}	$\overline{x}_{t-2}$	y_{t-1}	$\overline{y}_{t-2}$
1	1,9302 ($<$ 0,001)	–	0,0057 (0,45)	–
2	2,3047 ($<$ 0,001)	0,5683 (0,044)	−0,0088 (0,343)	0,0075 (0,419)
3	1,9689 ($<$ 0,001)	1,0734 (0,002)	0,0138 (0,17)	−0,017 (0,138)
4	2,2945 ($<$ 0,001)	1,0617 (0,007)	0,0092 (0,441)	−0,0141 (0,307)
5	2,2741 ($<$ 0,001)	1,0698 (0,015)	−0,0008 (0,938)	$<$ 0,0001 (0,996)
6	2,5812 ($<$ 0,001)	1,4609 (0,003)	−0,0005 (0,966)	−0,0072 (0,678)

8.4. Incorporación de la CVT

Como se mencionó anteriormente, hay más de una manera de incorporar una CVT a un modelo mixto, en esta sección compararemos algunas de éstas.

8.4.1. Incorporación de covariable fija

Una de las transformaciones que puede aplicarse sobre la covariable .ªdherencia al tratamiento.ᵉˢ convertirla en una variable dicotómica, cuyo valor es 1 si el paciente adhirio correctamente en todo el estudio y 0 si adhirio de manera incorrecta en algún mes. Puede observarse en la tabla 8.4.1 que la covariable solo es significativa en la ordenada al origen. Este modelo indica que, controlando por el resto de las variables, los pacientes con adherencia perfecta tienen en promedio 2.792 menor TAS en el primer mes del tratamiento y continúa disminuyendo en 0.265 por cada mes.

Tabla 8.4.1: Modelo 2: Incorporación adherencia perfecta

Log-Likelihood			-15412.1077	
AIC			30848.215	
BIC			30923.501	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	121.442	2.341	51.874	$<$ 0.001
Sexo	3.807	0.740	5.147	$<$ 0.001
Edad	0.173	0.038	4.513	$<$ 0.001
DBT	3.364	1.481	2.272	0.023
Adherencia Perfecta	-2.792	1.015	-2.750	0.006
tpo programa	-0.108	0.154	-0.702	0.483
tpo programa*DBT	-0.683	0.308	-2.215	0.027
tpo programa*Adherencia Perfecta	-0.265	0.211	-1.253	0.210

Otra manera de incorporar la covariable fija sin perder tanta información es usar la proporción de adherencia correcta al final del estudio. Es decir, si de las 7 ocasiones el paciente adhirio correctamente al tratamiento en solo 5, el valor que se le asignará es 5/7 ($\approx$ 0,71). Los coeficientes de este modelo indican

que, controlando por el resto de las variables, los pacientes tienen en promedio un porcentaje de 3.063 menor TAS al primer mes de tratamiento equivalente al porcentaje de adherencia total, y continuará disminuyendo en un porcentaje de 0.965 por mes también equivalente al porcentaje de adherencia total. Por ejemplo, un paciente que adhirió correctamente sólo al 50 % del tratamiento, tendra una TAS en promedio un 1.5315 (3,063x0,5) menor que un paciente que no adhirio correctamente en ninguna ocasión.

Tabla 8.4.2: Modelo 3: incorporación adherencia total

Log-Likelihood			-15414.7464	
AIC			30853.492	
BIC			30928.779	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	122.805	2.843	43.193	< 0.001
Sexo	3.738	0.746	5.008	< 0.001
Edad	0.169	0.039	4.390	< 0.001
DBT	3.412	1.489	2.291	0.022
Adherencia Total	-3.063	2.150	-1.425	0.154
tpo programa	0.549	0.380	1.448	0.148
tpo programa*DBT	-0.657	0.308	-2.133	0.033
tpo programa*Adherencia Total	-0.965	0.444	-2.174	0.030

Puede notarse que ambas maneras de introducir la covariable ayudan para explicar el comportamiento de la TAS en el estudio (siendo la primera opción levemente mejor, ya que el AIC y el BIC son menores).

8.4.2. Incorporación como CVT

Como vimos anteriormente, la CVT puede considerarse exógena, por lo tanto puede introducirse al modelo en su forma natural. Los coeficientes de Adherencia de este modelo y de los siguientes no son fácilmente interpretables, dado que contienen los efectos tanto entre-persona como intra-persona.

Tabla 8.4.3: Modelo 4: incorporación la adherencia en cada ocasión

Log-Likelihood			-15388.2302	
AIC			30800.460	
BIC			30875.746	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	122.388	2.426	50.444	< 0.001
Sexo	3.768	0.746	5.054	< 0.001
Edad	0.164	0.038	4.271	< 0.001
DBT	3.212	1.485	2.163	0.031
Adherencia	-2.321	0.922	-2.517	0.012
tpo programa	0.442	0.242	1.828	0.067
tpo programa*DBT	-0.584	0.306	-1.906	0.057
tpo programa*Adherencia	-0.783	0.262	-2.984	0.003

Al comparar los criterios de AIC y BIC de este modelo con los anteriores se puede notar que han disminuido en gran medida, lo cual quiere decir que este modelo explica la variación de la TAS de mejor manera. Además, ambos efectos de adherencia (ordenada y pendiente) son significativos en este modelo, lo cual no ocurria en los modelos anteriores. Este comportamiento es el esperado, dado que se está usando toda la información disponible de la CVT en cada momento en vez de resumirla.

Además de incorporar la adherencia en cada ocasión, también puede acompañarse de la adherencia en momentos anteriores o algunas medidas resumenes, algunos ejemplos pueden observarse en las siguientes tablas.

Tabla 8.4.4: Modelo 5: incorporación la adherencia en cada ocasión y en ocasión anterior

Log-Likelihood			-15387.1933	
AIC			30800.386	
BIC			30881.946	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	122.467	2.426	50.485	< 0.001
Sexo	3.755	0.745	5.038	< 0.001
Edad	0.166	0.038	4.309	< 0.001
DBT	3.205	1.486	2.157	0.031
Adherencia	-2.527	0.933	-2.709	0.007
tpo programa	0.543	0.252	2.157	0.031
tpo programa*DBT	-0.579	0.307	-1.888	0.059
tpo programa*Adherencia	-0.671	0.274	-2.451	0.014
tpo programa*Adherencia lag1	-0.232	0.161	-1.440	0.150

Tabla 8.4.5: Modelo 6: incorporación la adherencia en cada ocasión y adherencia acumulada

Log-Likelihood			-15387.1802	
AIC			30800.360	
BIC			30881.920	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	122.454	2.425	50.503	< 0.001
Sexo	3.735	0.745	5.011	< 0.001
Edad	0.168	0.039	4.370	< 0.001
DBT	3.199	1.486	2.153	0.031
Adherencia	-2.694	0.957	-2.816	0.005
tpo programa	0.678	0.291	2.328	0.020
tpo programa*DBT	-0.581	0.306	-1.896	0.058
tpo programa*Adherencia	-0.547	0.309	-1.772	0.076
tpo programa*Adherencia Acumulada	-0.527	0.364	-1.450	0.147

En las tablas 8.4.2 y 8.4.2 puede notarse que las covariables adicionales no son significativas. Otra forma de comparar estos modelos con el modelo 4 es a través del criterio BIC, que es mayor en ambos casos. Como conclusión, la mejor decisión hasta el momento es el modelo 4 que se mantiene más parsimonioso.

8.4.3. Incorporación dividiendo efecto entre e intra

El efecto de la CVT puede también dividirse en distintos coeficientes para el efecto entre-persona y el efecto intra-persona, los valores de éstos coeficientes pueden observarse en la tabla 8.4.3. El coeficiente -0.607 quiere decir que, manteniendo constante la proporción de adherencia total al tratamiento, se espera que la TAS disminuya en promedio 0.607 los meses en los que se adhiere correctamente al tratamiento. Por otro lado, el coeficiente -0.481 se interpreta como, después de controlar por la adherencia al tratamiento en ese mes, se espera que la TAS disminuya en promedio en 0.481 relativo a la proporción de adherencia total.

Tabla 8.4.6: Modelo 7: incorporación la adherencia dividiendo efecto entre e intra persona

Log-Likelihood			-15387.3388	
AIC			30802.677	
BIC			30890.511	
Covariable	Coef.	Std. Err.	z	$P < \|z\|$
Intercept	122.542	2.841	43.141	< 0.001
Sexo	3.729	0.746	4.997	< 0.001
Edad	0.168	0.039	4.348	< 0.001
DBT	3.220	1.487	2.166	0.030
Adherencia	-2.611	1.054	-2.478	0.013
Adherencia Total	-0.121	2.380	-0.051	0.959
tpo programa	0.689	0.378	1.822	0.068
tpo programa*DBT	-0.578	0.306	-1.887	0.059
tpo programa*Adherencia	-0.607	0.314	-1.930	0.054
tpo programa*Adherencia Total	-0.481	0.528	-0.911	0.362

9. 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.

Como futuros pasos se propone estudiar de manera más profunda 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.

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] D.-G. Chen and J. R. Wilson, *Innovative Statistical Methods for Public Health Data*. Springer International Publishing, 2015.

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