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

16 de mayo de 2022

Índice

1. Introducción

Los datos longitudinales tienen la particularidad de estar conformados por mediciones repetidas sobre una unidad, las cuales pueden surgir por ser medidas en diferentes momentos o conidiciones. Su principal interés es estudiar los cambios en el tiempo y los factores que influencian el cambio.

Los modelos mixtos permiten ajustar datos con estas particularidades, donde la respuesta es modelada por una parte sistemática que está formada 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 covariables en los estudios longitudinales se pueden clasificar en dos categorias: fijas y variables en el tiempo. Las diferencias entre estos tipos de covariables pueden llevar a diferentes intereses de investigación, diferentes tipos de análisis y diferentes conclusiones.

Las covariables fijas son variables independientes que no tienen variación intra-sujeto, lo que significa que el valor de la covariable no cambia para un individuo determinado en el estudio longitudinal. Este tipo de covariable se puede usar para realizar comparaciones entre poblaciones y describir diferentes tendencias en el tiempo, pero no permite una relación dinámica entre la covariable y la respuesta.

Las covariables variables en el tiempo (CVT) son variables independientes que contienen ambas variaciones, intra y entre sujeto, lo que significa que el valor de la covariable cambia para un individuo determinado a lo largo del tiempo y además puede cambiar para diferentes sujetos. Una CVT se puede usar para hacer comparaciones entre poblaciones, describir tendencias en el tiempo y también la relación dinámica entre la CVT y la respuesta

Se puede ver que las CVT permiten diferentes tipos de relaciones y conclusiones que las covariables fijas. Por ejemplo, covariables como la edad pueden cambiar a través del tiempo, pero cambian de manera predecible. Por otro lado, covariables como la precipitación diaria pueden cambiar a través del tiempo pero no pueden ser predecidas. En esos casos es importante considerar las relaciones entre la CVT y la respuesta a través del tiempo.

En el presente informe se cuenta con un programa de atención y control de pacientes hipertensos iniciado en el año 2014 en Rosario. En cada visita se registra el seguimiento del tratamiento y los valores de la tensión arterial sistólica. En particular, se desea evaluar si la adherencia al tratamiento farmacológico influye en los valores de la tensión arterial sistólica (TAS) a lo largo del seguimiento. Como la variable "adherencia al tratamiento farmacológico" es una CVT estocástica se evaluaran diferentes enfoques para incluirla en un modelo longitudinal que pueda explicar el cambio en la tensión arterial sistólica media a lo largo del tiempo.

Un aspecto a tener en cuenta en este trabajo es que, si bien contamos con mucha otra información para obtener modelos que describan de mejor manera el comportamiento de la TAS, nos centraremos en modelos más simples con respecto a las covariables fijas con el fin de no perder de vista la relación entre la variable respuesta y la CVT.

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.
- Evaluar 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. Los Datos Longitudinales

Los datos longitudinales están conformados por mediciones repetidas de una misma variable realizadas a la misma unidad. Estas mediciones surgen de observar unidades en diferentes ocasiones, es decir 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 acerca de la independencia entre las respuestas de cada unidad y la homogeneidad de variancias frecuentemente no son válidos

Las ocasiones en las que se registran las mediciones repetidas no necesariamente serán iguales para todos los individuos, por lo tanto se pueden obtener tanto estudios balanceados (todos los individuos tienen el mismo número de mediciones durante un conjunto de ocasiones comunes) como desbalanceados (la secuencia de tiempos de observaciones no es igual para todos los individuos). Otra característica de estos datos es que en ocasiones se pueden obtener valores perdidos, obteniendo datos incompletos aunque se cuente con un estudio balanceado.

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 le llama Y_{ij} a la medición sobre la unidad i en la ocasión j, con $i = 1, ..., N; j = 1, ..., n$

Asociadas a cada unidad se observan las covariables X_{ij} medidas también 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: $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), ..., (X_{in_i}, Y_{in_i})$

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 la secuencia de medidas repetidas de una unidad tiene un comportamiento determinado, que produce que las mediciones más cercanas sean más parecidas.
- *Error de medición:* Surge debido a los errores de medida, se puede disminuir usando instrumentos de medición más precisos

Estas tres fuentes de variación pueden clasificarse en *"variabilidad entre"*, es decir, la variación entre las unidades (heterogeneidad entre unidades) y *"variabilidad intra"*, es decir, la variación entre las mediciones de las misma unidad (variación biológica intra-unidad y error de medición)

Dado que, como se mencionó anteriormente, las mediciones están correlacionadas entre sí, si se utilizaran las técnicas habituales basadas en la independencia entre mediciones, los errores estándares nominales

van a ser incorrectos, lo cual nos llevaría a inferencias incorrectas sobre los parámetros del modelo. En base a esto, surgen técnicas que consideran esa correlación modelando los datos considerando la modelación de dos estructuras: la parte media y la estructura de covariancia.

4. Modelos lineales mixtos

En estos modelos, cada unidad tiene una trayectoria individual caracterizada por parámetros y un subconjunto de esos parámetros ahora 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).

Se consideran las dos fuentes de variación (intra y entre) presentes en los datos longitudinales. Entonces, este modelo va a ser similar al modelo lineal general con respecto a la parte media del mismo, pero se va a diferenciar en cuanto a la estructura de covariancia.

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

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i; \quad i = 1, ..., N; \quad Y_i = (Y_{i1}, Y_{i2}, ...Y_{in_i})'$$

Donde:

- Y_i: Vector de la variable respuesta de la i-ésima unidad, de dimensión $(n_i * 1)$
- X_i: Matriz de diseño de la i-ésima unidad, que caracteriza la parte sistemática de la respuesta, de dimensión $(n_i * p)$
- β: Vector de parámetros de dimensión $(p * 1)$
- Z_i: Matriz de diseño de la i-ésima unidad, que caracteriza la parte aleatoria de la respuesta, de dimensión $(n_i * k)$
- b_i: Vector de efectos aleatorios de la i-ésima unidad, de dimensión $(k * 1)$
- ε_i: Vector de errores aleatorios de la i-ésima unidad, de dimensión $(n_i * 1)$

ε_i y b_i son independientes.

$$\varepsilon_i \sim N_{n_i}(0, R_i)$$

$$b_i \sim N_k(0, D_i)$$

Las matrices D_i y R_i contienen las variancias y covariancias de los elementos de los vectores b_i y ε_i respectivamente. A partir de este modelo se obtiene:

- $E(y_i/b_i) = X_i\beta + Z_i b_i$ (media condicional o específica de la i-ésima unidad)
- $E(Y_i) = X_i\beta$ (media marginal)

- $Cov(Y_i/b_i) = R_i$ (variancia condicional)
- $Cov(Y_i) = Z_i D_i Z_i' + R_i = \Sigma_i$ (variancia marginal)

Generalmente, la matriz D_i adopta una estructura de covariancia arbitraria, mientras que la matriz R_i adopta cualquiera de las vistas anteriormente

4.1. Estimación de los parámetros del modelo

Bajo el supuesto de que ε_i y 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 $Y_i \sim N_{n_i}(X_i\beta, \Sigma_i)$ y las Y_i son independientes entre sí, se obtiene la siguiente función de log-verosimilitud:

$$l = -\frac{1}{2}\sum_{i=1}^{N} n_i ln(2\pi) - \frac{1}{2} ln|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{N}[(Y_i - X_i\beta)'\Sigma_i^{-1}(Y_i - X_i\beta)] \tag{4.1.1}$$

Siendo Σ_i función del vector θ que contiene los parámetros de covariancia.

La ecuación anterior se debe derivar con respecto a β y θ y luego debe igualarse a cero, de esta manera se obtienen sus estimadores. Cuando θ es desconocido (lo que generalmente sucede) se obtiene una ecuación no lineal, por lo que no se puede obtener una expresión explícita de $\hat{\theta}$, para encontrar su solución se recurren a algoritmos numéricos. El estimador del vector β resulta:

$$\hat{\beta} = (\sum_{i=1}^{N} X_i' \hat{\Sigma_i}^{-1} X_i)^{-1} \sum_{i=1}^{N} X_i' \hat{\Sigma_i}^{-1} Y_i$$

El estimador $\hat{\beta}$ resulta insesgado de β. Cuando θ es conocido se conoce la distribución exacta del estimador. Sin embargo, cuando es desconocido, no se puede calcular de manera exacta la matriz de covariancias de $\hat{\beta}$. Si el número de unidades es grande se puede demostrar que asintóticamente:

$$\hat{\beta} \sim N_p(\beta, V_\beta) \quad donde \quad V_\beta = (\sum_{i=1}^{N} X_i' \hat{\Sigma_i}^{-1} X_i)^{-1}$$

4.1.2. Método de máxima verosimilitud restringida (REML)

El inconveniente que posee el método de MV es que los parámetros de covariancia resultan sesgados. Es decir, a pesar de que la estimación de β resulta insesgada, no pasa lo mismo con θ. 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 β es estimado a partir de los datos.

El método REML separa la parte de los datos usada para estimar β de aquella usada para estimar los parámetros de Σ_i, la función de log-verosimilitud restringida que se propone es:

$$l^* = -\frac{1}{2}\sum_{i=1}^{N} n_i ln(2\pi) - \frac{1}{2}ln|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{N}[(Y_i - X_i\beta)'\Sigma_i^{-1}(Y_i - X_i\beta)] - -\frac{1}{2}ln|\sum_{i=1}^{N} X_i'\hat{\Sigma_i^{-1}}X_i| \quad (4.1.2)$$

Maximizando esta funcion con respecto a β y θ se obtiene:

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

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

4.1.3. Problemas con la estimación

Pepe y Anderson (1994) mostraron que estas ecuaciones llegan a cero solo si la data cumple con el supuesto

$$E[Y_{ij}|X_{ij}] = E[Y_{ij}|X_{ij}, j = 1, ..., T] \quad (4.1.3)$$

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 y las CVT ueden ser utilizadas 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

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, valor de la presión arterial o condición de fumar (si/no).

5.2.1. Covariables 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 en los estudios cross-over. Ejemplos de las segundas son: valor del colesterol, ingesta de alcohol (si/no), ingesta de grasas, etc.

5.2.2. Covariables exógenas y endógenas

Se dice que una CVT es exógena cuando los valores actuales y anteriores de la respuesta en la ocasión $j(Y_{i1}, ..., Y_{ij})$, dados los valores actuales y precedentes de la CVT $(X_{i1}, ..., X_{ij})$, no predicen el valor posterior de X_{ij+1}. Más formalmente, una CVT es exógena cuando:

$$f(X_{ij+1}|X_{i1}, ..., X_{ij}, Y_{i1}, ..., Y_{ij}) = f(X_{ij+1}|X_{i1}, ..., X_{ij}) \tag{5.2.1}$$

Y en consecuencia:

$$E(Y_i|X_i) = E(Y_i|X_{i1}, ..., X_{in_i}) = E(Y_i|X_{i1}, ..., X_{ij}) \tag{5.2.2}$$

Esta definición implica que la respuesta en cualquier momento puede depender de 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.

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 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.

A los parámetros de regresión se les puede dar una interpretación causal sólo cuando se puede asumir que las CVT son exógenas con respecto a la variable respuesta.

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.