Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística

Anteproyecto de Tesina

Modelos longitudinales con covariables que varían en el tiempo

Autor: **Esteban Cometto**

Directora: Noelia Castellana

Codirectora: Cecilia Rapelli

19 de octubre de 2021

Índice

1. Introducción

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

El objetivo principal de estos estudios es estudiar los cambios en el tiempo y los factores que influencian el cambio.

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.

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. Estos modelos permiten, además, hacer predicciones del perfil de una unidad específica. La selección de la estructura de covariancia apropiada produce estimadores más eficientes y consecuentemente, pruebas estadísticas más robustas para los efectos de interés

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, una CVT puede estar involucrada en efectos acumulados para diferentes valores a través del tiempo (Fitzmaurice y Lard 1995). Además, ciertas CVT transmiten diferente información que otras. 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 que realiza un seguimiento exhaustivo de 560 pacientes. Este programa contempla: efectores no médicos supervisados, tratamiento farmacológico genérico para la hipertensión y utilización de un algoritmo terapéutico sistematizado. En cada visita se registran tantocaracterísticas de los pacientes, del tratamiento y de los valores de la tensión arterial. En particular, se desea evaluar si la adherencia al tratamiento farmacológico influye en los valores de la tensión arterial sistólica 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

Profundizar en el estudio de propuestas metodológicas para utilizar la información obtenida de la covariable que varía en el tiempo dentro de un modelo mixto.

2.2. Objetivos Específicos

- Específicar distintos tipos de CVT
- Transformaciones a realizar sobre la CVT antes de incluirla al modelo, incluyendo conversión a covariable fija
- Consideraciones sobre interpretación de los parámetros sobre las CVT
- Indagar sobre feedback entre la CVT y la variable respuesta

3. Metodología

3.1. 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.

El objetivo principal de estos estudios es estudiar los cambios en el tiempo y los factores que influencian el cambio.

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}, de las cuales existen dos tipos: variables en el tiempo (estocásticas) e invariables en el tiempo (estacionarias)

Existen estudios empíricos que llevan a pensar que 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 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.

3.2. Análisis exploratorio

Antes de ajustar algún modelo, lo primero siempre es realizar un análisis exploratorio para estudiar cómo se comportan los datos. A continuación se presentan técnicas gráficas para cada estructura.

Evaluación de la parte media

- *Perfil individual:* Consiste en un gráfico de dispersión en el cual se representan las respuestas vs las ocasiones. Cada respuesta tiene un punto y se une con un segmento los puntos de la misma unidad. Sirven para detectar si hay mucha variabilidad entre y dentro de las unidades y si hay valores atípicos.

- *Perfiles promedio por grupo:* En general son más informativos. Para cada tiempo calculamos un promedio para cada grupo y luego se unen los puntos. Permiten ver la tendencia de las variables a través de las ocasiones. Se superponen en un mismo gráfico los perfiles promedio de cada grupo

Evaluación de la estructura de covariancia

- *Matriz de diagrama de dispersión:* Para cada par de ocasiones se grafican los valores esperados de la respuesta y todos estos gráficos se acomodan dentro de una matriz. En general se utiliza cuando las ocasiones son las mismas para todas las unidades.

- *Gráfico de Draftman:* Es similar al gráfico anterior pero utilizando variables estandarizadas. La utilización de la variable respuesta estandarizada ayuda a eliminar la variabilidad de los datos asociada con diferencias en las medias y variancias en el tiempo, permitiendo visualizar más claramente el patrón de correlación.

- *Gráfico PRISM (Partial Regression on Intervenors Scatterplot Matrix):* Utilizando la variable estandarizada se crea una matriz de gráficos de dispersión. En la primera diagonal se encuentran gráficos de dispersión entre la variable respuesta en los tiempos t_j y t_{j+1}. Luego, en la k-ésima diagonal, se obtienen gráficos de regresión parcial de las respuestas en los tiempos t_j y t_{j+k}, ajustadas por las respuestas en los tiempos intermedios. Estos gráficos permiten ver con mayor claridad ciertos tipos de estructuras seriales que se dan entre las medidas repetidas.

- *Correlograma:* Representa las características que existen entre las respuestas de los individuos de cada grupo en tiempos que están separados una cantidad de periodos. Permite analizar cómo evoluciona la correlación a medida que aumenta el número de rezagos.

- *Semivariograma:* Cuando los datos están desbalanceados, el semivariograma permite distinguir las 3 fuentes de variabilidad. Después de haber estimado un modelo, el mismo permite confirmar si la estructura de correlación es adecuada. El semivariograma se define como una función:

$$\gamma(u) = \frac{1}{2}E[(\varepsilon_{ij} - \varepsilon_{ij'})^2]$$

$$u_{ijj'} = |t_{ij} - t_{ij'}|$$

$$\widehat{\gamma(u)} = v_{ijj'} = \frac{1}{2}(r_{ij} - r_{ij'})$$

donde r_{ij} y $r_{ij'}$ son los residuos estandarizados obtenidos después de ajustar un modelo de regresión considerando las observaciones independientes.

Se va a obtener un gráfico donde la variabilidad total va a estar dividida en 3 partes.

Si la curva no empieza en cero significa que hay error de medición, si tiene pendiente quiere decir que hay un error debido a una causa biológica (correlación serial) y si la misma no llega a la variancia total significa que se debe explicar la variabilidad entre.

3.3. Modelo lineal general para datos longitudinales

Si se piensa que existe una tendencia en el tiempo de las respuestas, y esta tendencia se puede expresar como una función, se puede escribir o representar a las medidas repetidas de una unidad en un vector Y_i. Entonces, un modelo lineal para representar la evolución en el tiempo va a ser:

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

Donde:

- Y_{ij}: respuesta obtenida de la i-ésima unidad en la ocasión t_{ij}.
- X_i: matriz de diseño de la i-ésima unidad, de dimensión $(n_i * p)$
- β: vector de parámetros de dimensión $(p * 1)$
- ε_i: vector de errores aleatorios de la i-ésima unidad, de dimensión $(n_i * 1)$, este mismo representa todas las fuentes de variabilidad de los datos longitudinales

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

- θ: vector de parámetros desconocidos de covariancia, de dimensión $(q * 1)$

3.4. Modelación de la estructura de covariancia

Al tenerse tantos parámetros de variancia (n) y covariancia $n(n-1)/2$ para estimar, se proponen modelos específicos para la estructura de correlación. Se trata de elegir una estructura que no tenga tantos parámetros. Sin embargo, se debe tener cuidado de no seleccionar estructuras demasiado parcas con las que se pierda información.

La matriz de covariancia de cada unidad va a ser función de θ. El número de parámetros de este vector depende de la estructura de la matriz.

A continuación se mencionan algunas estructuras que se pueden utilizar, se llamará R a la matriz de correlaciones

3.4.1. Arbitraria o no estructurada (datos balanceados)

Considera variancias y covariancias distintas entre las mediciones repetidas. Siendo $\sigma_{jj'} = Cov(Y_{ij}Y_{ij'})$ se expresa como:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}$$

La ausencia de restricciones hace que haya que estimar una gran cantidad de parámetros

3.4.2. Simetria compuesta (datos balanceados o no balanceados)

La correlación entre pares de observaciones es la misma, sin importar la cantidad de rezagos entre ellas, $Corr(Y_{ij}, Y_{ik}) = \rho$ para todo $j \neq k$

$$R_i = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

3.4.3. Toeplitz (datos equiespaciados)

Se plantea para que cualquier par de respuestas que estén igualmente separadas en el tiempo la correlación es la misma, $Corr(Y_{ij}, Y_{ij+k}) = \rho_k$ para todo j y k.

$$R_i = \begin{bmatrix} 1 & \rho_1 & \dots & \rho_n \\ \rho_1 & 1 & \dots & \rho_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_n & \rho_{n-1} & \dots & 1 \end{bmatrix}$$

3.4.4. Autorregresiva de primer orden (datos equiespaciados)

Es un caso especial de la estructura anterior, en la que $Corr(Y_{ij}, Y_{ij+k}) = \rho^k$. Esta estructura asume que la correlación entre medidas repetidas disminuye a medida que aumenta el número de rezagos entre ellas.

$$R_i = \begin{bmatrix} 1 & \rho & \dots & \rho^n \\ \rho & 1 & \dots & \rho^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^n & \rho^{n-1} & \dots & 1 \end{bmatrix}$$

3.4.5. Markov (datos no equiespaciados)

Es una generalización de la estructura autorregresiva para mediciones no equiespaciadas. $Corr(Y_{ij}, Y_{ij'}) = \rho^{d_{jj'}}$, donde $d_{jj'} = |t_{ij} - t_{ij'}|$ para todo $j \neq j'$.

$$R_i = \begin{bmatrix} 1 & \rho^{d_{12}} & \dots & \rho^{d_{1n}} \\ \rho^{d_{21}} & 1 & \dots & \rho^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^n & \rho^{n-1} & \dots & 1 \end{bmatrix}$$

3.5. 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:

- $E(y_i/b_i) = X_i\beta + Z_ib_i$ (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)
- $Cov(Y_i) = Z_iD_iZ_i' + R_i = \Sigma_i$ (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

3.6. 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

3.6.1. 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:

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

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:

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

3.6.2. Método de máxima verosimilitud restringida (REML)