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

1 de febrero de 2022

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:

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

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:

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

3.6.2. 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:

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

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

3.7. Pruebas de hipótesis

Generalmente, la inferencia en estos modelos se centra en los parámetros de la parte media. Es decir, en combinaciones lineales de los parámetros de β. La hipótesis lineal general para los test se plantea construyendo dichas combinaciones a través de $L'\beta$, siendo L' una matriz de dimensión $(r * p)$.

El estimador de $L'\beta$ resulta $L'\hat{\beta}$ y asintóticamente su distribución muestral es aproximadamente:

$$L'\beta \sim N_p(L'\beta, L'V_\beta L)$$

Para probar las hipótesis se proponen dos métodos basados en la función de verosimilitud:

3.7.1. Test de Wald

Se plantean las hipótesis:

$$H_0)L'\beta = 0 \quad H_1)L'\beta \neq 0$$

La estadística de prueba resulta:

$$W = (L'\hat{\beta})'(L'\hat{V}_\beta L)^{-1}(L'\hat{\beta})$$

Donde W se distribuye aproximadamente como χ^2_r.

Este test provee inferencias válidas cuando el N es grande, ya que utiliza la aproximación asintótica a la distribución Normal. Si el N es chico se propone reemplazar la distribución de la chi-cuadrado por una F de Snedecor. El problema con el uso de esta estadística es que no se conocen los grados de libertad del denominador y deben ser calculados con los datos.

3.7.2. Test de cociente de verosimilitud

Se basa en la teoría asintótica y va a suplir las dificultades del test de Wald. El test se obtiene comparando las verosimilitudes de dos modelos, uno de los cuales incorpora la restricción $L'\beta = 0$ (modelo reducido) y el otro no tiene restricciones (mdoelo completo). Estos dos modelos están anidados, ya que el modelo reducido es un caso particular del modelo completo.

Al igual que anteriormente, se plantean las hipótesis:

$$H_0)L'\beta = 0 \quad H_1)L'\beta \neq 0$$

La función de log-verosimilitud maximizada del modelo completo es:

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

Y la del modelo reducido es:

$$-\frac{1}{2}\sum_{i=1}^{N} n_i ln(2\pi) - \frac{1}{2}ln|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{N}[(Y_i - X_i\beta_0)'\Sigma_i^{-1}(Y_i - X_i\beta_0)]$$

Donde β_0 es el vector de parámetros resultante de haber impuesto $L'\beta = 0$.

Si el N es grande, la distribución muestral aproximada de la estadística es:

$$G^2 = -2(\hat{l}_{red} - \hat{l}_{comp}) \sim \chi^2_v$$

Donde v es el número de parámetros del modelo completo menos el número de parámetros del modelo reducido

Si lo que se quiere comparar mediante este test es la estructura media de dos modelos, se recomienda estimar los parámetros mediante el método ML. En cambio, si lo que se desea comparar son patrones de covariancia anidados entre dos modelos se recomienda estimar los parámetros con el método REML.

Cuando se desea hacer inferencia sobre las componentes de variancia (por ejemplo, postular que una variancia es nula), la diferencia entre los dos modelos va a ser de $k+1$ parámetros (1 de variancia y k de covariancia). Si sucede esto, la distribución de la estadística G^2 ya no será una chi-cuadrado común, si no que será una mezcla entre dos chi-cuadrado, una de k grados de libertad y otra de $k+1$ grados de libertad.

3.8. Elección entre modelos de covariancia

Cuando los modelos no están anidados, como sucede generalmente cuando se plantean modelos con distintas estructuras de covariancia, no se puede usar el método del cociente de verosimilitud para compararlos.

Como en la matriz de covariancia siempre intervienen los residuos, y en ellos aparecen los parámetros de la parte media del modelo, para asegurarse que la parte media esté bien especificada se elige un modelo maximal (modelo con todos los parámetros que queremos incorporar). Dado dicho modelo, se plantean distintos modelos que se van a diferenciar únicamente en la estructura de la matriz de covariancia.

Los enfoques que se proponen a continuación se basan en la comparación de versiones penalizadas de las log-verosimilitudes de los modelos. Como es conocido, a medida que se incorporan parámetros a los modelos, mayor va a ser la verosimilitud. Para comparar modelos con distinto número de parámetros, se penalizan a los mismos, surgiendo varios criterios A continuación se destacan dos de estos:

- *Criterio de Akaike (AIC)*: $-2\hat{l} + 2p$
- *Criterio bayesiano se Schwarz (BIC)*: $-2\hat{l} + pln(N)$

Donde:

$\hat{l}$: log-verosimilitud maximizada del modelo

p: número de parámetros del modelo

El BIC penaliza más la verosimilitud dando modelos más sencillos. El criterio para seleccionar un modelo es aquel que minimice los valores de AIC o BIC

3.9. Predicción de los efectos aleatorios

En muchas aplicaciones la inferencia está centrada solamente en los efectos fijos. Estos parámetros se interpretan como los cambios de la respuesta media en el tiempo, pero muchas veces se desea conocer además los perfiles individuales, por lo que necesitamos conocer los valores de b_i. Como b_i es una cantidad aleatoria se habla de predicción de b_i.

Llamando $c(Y_i)$ al predictor de los efectos aleatorios, se debe minimizar $E[b_i - c(Y_i)]^2$. La función $c(Y_i)$ que hace que esa esperanza sea lo más chica posible se llama esperanza condicional de b_i dado Y_i y se simboliza:

$$E(b_i/Y_i) = D_i Z_i' \Sigma_i^{-1}(Y_i - X_i\beta)$$

Esto se conoce como mejor predicción lineal insesgada (BLUP) y el BLUP empírico (EBLUP) o estimador empírico de Bayes resulta:

$$\hat{b_i} = \hat{D}_i Z_i' \hat{\Sigma_i}^{-1}(Y_i - X_i\hat{\beta})$$

A través de esto se pueden conocer los perfiles individuales, siendo $\hat{Y_i} = X_i\hat{\beta} + Z_i\hat{b_i}$, operando algebraicamente se llega a la expresión:

$$\hat{Y_i} = (\hat{R}_i\hat{\Sigma_i}^{-1})X_i\beta + (I_n - \hat{R}_i\hat{\Sigma_i}^{-1})Y_i$$

Resulta que el perfil individual estimado es un promedio ponderado entre el perfil de la respuesta media poblacional ($X_i\hat{\beta}$) y el perfil de la respuesta observada en la unidad i (Y_i).

3.10. Examen de residuos

Como en todo análisis de datos en donde se utiliza un modelo estadístico, es necesario realizar un estudio del cumplimiento de los supuestos del modelo utilizando los gráficos de residuos.

El modelo lineal mixto para la unidad i es:

$$Y_i = X_i\beta + Z_i\beta + \varepsilon_i$$

Con b_i y ε_i independientes y $\varepsilon \sim N_{n_i}(0, R_i)$, $b_i \sim N_k(0, D_i)$.

En los modelos lineales mixtos se va a distinguir entre tres tipos de residuos:

3.10.1. Residuos marginales

Se definen como la diferencia entre el vector de respuestas y su media marginal estimada.

$$r_{iM} = Y_i - \widehat{E(Y_i)} = Y_i - X_i\hat{\beta}$$

3.10.2. Residuos condicionales

Se definen como la diferencia entre el vector de respuestas y su media condicional estimada.

$$r_{iC} = Y_i - \widehat{E(Y_i/b_i)} = Y_i - X_i\hat{\beta} - Z_i\hat{b_i}$$

3.10.3. Residuos de los efectos aleatorios

$$r_{iEA} = \widehat{E(Y_i/b_i)} - \widehat{E(Y_i)} = Z_i\hat{b_i}$$

Estos residuos así calculados reciben el nombre de residuos ordinarios, van a estar correlacionados y sus variancias no van a ser necesariamente contrastes. Para evitar este inconveniente es necesario realizar una estandarización de los mismos:

- Si se divide el residuo por $\sqrt{\widehat{Var(r_i)}}$ se obtienen los residuos estudentizados.
- Si se divide el residuo por $\sqrt{\widehat{Var(Y_i)}}$ se obtienen los residuos de Pearson.

Se define como residuo puro al residuo que depende de solo los componentes fijos y del error que predice, y residuo confundido si además de depender de las componentes fijas y del error que predicen están en función de otros residuos.

4. Covariables que varían en el tiempo

En los estudios longitudinales, las variables independientes pueden ser clasificadas en dos categorías: covariables fijas en el tiempo, 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.

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. Por ejemplo, el sexo biológico de una persona o el grupo de tratamiento.

Las CVT son variables independientes que incluyen tanto la variación intra-sujeto y 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).

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.

4.1. 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 segunas son: valor del colesterol, ingesta de alcohol (si/no), ingesta de grasas, etc.

4.2. 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:

$$f(X_{ij+1}|X_{i1}, ..., X_{ij}, Y_{i1}, ..., Y_{ij}) = f(X_{ij+1}|X_{i1}, ..., X_{ij}) \tag{3}$$

Y en consecuencia:

$$E(Y_i|X_i) = E(Y_i|X_{i1}, ..., X_{in_i}) = E(Y_i|X_{i1}, ..., X_{ij}) \quad (4)$$

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. Por ejemplo, cuando se evalúa si la cantidad de actividad física está asociada al nivel de glicemia. El nivel de actividad física en un determinado momento puede estar (o no) asociado a niveles previos y también puede estar asociado a valores previos de la glicemia (un paciente con valor de glicemia alto en una visita puede decidir aumentar su nivel de actividad física para reducir este valor)

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.

4.3. Otros tipos de CVT

Trabajos más recientes han definido una nueva categorización de las CVT para facilitar las interpretaciones y los métodos de estimación adecuados para el modelo. Se pueden definir cuatro tipos de CVT relacionados con el grado de no exogeneidad con respecto a la respuesta.

4.3.1. CVT tipo I

Se clasifica una CVT como de tipo I si satisface:

$$E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij}) \tag{5}$$

En otras palabras, una CVT se considera de tipo I si la variable respuesta en la j-ésima ocasión es independiente de todos los valores de la CVT en diferentes momentos, aún de los previos a la ocasión. Variables que involucran cambios predecibles en el tiempo son tratadas como CVT tipo I, por ejemplo la edad o el momento de observación.

4.3.2. CVT tipo II

Una CVT se clasficia de tipo II si:

$$E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{i1}, ..., X_{ij}) \tag{6}$$

Cabe destacar que la clase de covariables de tipo I es un subconjunto de la clase de covariables de tipo II. Esta condición dice que el proceso de la CVT $X_{ij+1}, ..., X_{in_i}$ no se ve afectado por la respuesta Y_{ij}.

En otras palabras, la variable respuesta en la j-ésima ocasión puede estar asociada a valores previos de la CVT. Esta definición es similar pero no equivalente a la definición de exogeneidad. Se puede demostrar que la exogeneidad es suficiente es condición suficiente para que una CVT sea de tipo II. Este tipo de CVT incluyen covariables que pueden tener una asociación rezagada con la respuesta (los valores anteriores de la CVT pueden afectar a la respuesta actual) pero los valores de la covariable en un momento determinado no se verán afectados por los valores previos de la variable respuesta. Un ejemplo de este tipo de CVT es el "tratamiento farmacológico para la hipertensión arterial" con la variable respuesta "presión arterial".

4.3.3. CVT tipo III

Se clasifica a una CVT como de tipo III si no es de tipo II. No se asume independencia entre la respuesta y la covariable, por lo tanto, uede existir un *feedback* entre ambas en donde los valores de la CVT pueden estar afectados por valores previos de la variable respuesta. Un ejemplo de este tipo de CVT es el "tratamiento farmacológico para la

hipertensión arterial" con la variable respuesta "infarto de miocardio". Mientras que es esperable que la medicación impacte en la probabilidad de tener un infarto de miocardio, también tener un infarto de miocardio puede impactar en el cambio del tratamiento farmacológico.

4.3.4. CVT tipo IV

Una CVT se define como de tipo IV si:

$$E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij+1}, ..., X_{in_i}) \tag{7}$$

La CVT puede estar asociada con valores previos de la variable respuesta, pero la variable respuesta no está asociada con valores previos de la covariable, está sólo asociada con el valor observado de la covariable en la misma ocasión. Un ejemplo de este tipo de CVT es "presión arterial" con la variable respuesta "peso". En una determinada ocasión hay relación entre ambas variables, es esperable que valores previos del peso impacten en la presión arterial pero no al revés

4.4. Covariables rezagadas

En la mayoria de los casos, se suele utilizar solo la exposición que ocurre antes del tiempo t para predecir Y_{it}. 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 utiliar y sobre la estructura de sus coeficientes.

4.4.1. Una sola covariable rezagada

En algunas aplicaciones hay justificacion previa para considerar la covariable en un solo 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.

4.4.2. Múltiples covaribales rezagadas

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

$$Y_{it} = \beta_0 + \beta_1 X_{it-1} + \beta_2 X_{it-2} + ... + \beta_L X_{it-L},$$

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

A pesar de que los modelos rezagados distribuidos permiten modelar parsimoniosamente las covariables rezagadas múltiples, la especificación de del número de rezagos, L, 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.

4.5. Confusores variables en el tiempo

El análisis de regresión epidemiológica tradicional considera una clasificación de variables que están relacionadas tanto con una exposición de interés como con el resultado, ya sea como confusoras o intermedias. Una confusora se define vagamente como una variable asociada tanto con la exposición de interés como con la respuesta, y que, si se ignora en el análisis, conducirá a estimaciones sesgadas del efecto de la exposición. Una variable intermedia es aquella que se encuentra en la vía causal desde la exposición hasta el resultado y un análisis no debe controlar dicha variable, ya que se pierde el efecto de la exposición mediado por la variable intermedia.

Por ejemplo, en un estudio observacional de pacientes con hipertensión elevada, podemos esperar determinar la magnitud del beneficio (o daño) atribuible a un tratamiento

en la supervivencia del paciente o medidas longitudinales como la hipertensión en cierto momento. Sin embargo, podemos encontrar que la hipertensión en el tiempo t prediga tanto la hipertensión en momentos posteriores como las opciones de tratamiento posteriores. En este caso, la hipertensión en el momento $s < t$ es la variable de respuesta para las opciones de tratamiento recibidas antes de s, pero también es un predictor y, por lo tanto, un factor de confusión potencial para el tratamiento administrado en momentos futuros, $t > s$.

4.5.1. Feedback

Para aclarar los problemas que surgen con las covariables dependientes del tiempo, se consiidera un solo par de tiempos de estudio, $t = 0, 1$, con medidas de exposición y respuesta (X_{t-1}, Y_t). Sea Y_t un indicador de gravedad de la enfermedad o de los síntomas (1 = enfermedad/síntomas presentes, 0 = enfermedad/síntomas ausentes) y sea $X_t = 1$ si se administra tratamiento y 0 en caso contrario. Se supone que la exposición X_{t-1} precede a Y_t para $t = 1, 2$ y que Y_t precede o se mide simultáneamente con X_t. La figura 1 presenta un grafo que representa los modelos secuenciales condicionales:

$$logitE(Y_1/X_0 = x_0) = -0{,}5 - 0{,}5x_0,$$

$$logitE(X_1/Y_1 = y_1, X_0 = x_0) = -0{,}5 + 1y_1,$$

$$logitE(Y_2/H_1^X = h_1^X, Y_1 = y_1) = -1 + 1{,}5y_1 - 0{,}5x_1,$$

$$dondeH_1^x = X_0, X_1 y h_1^X = x_0, x_1$$

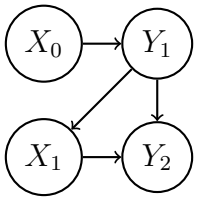

Figura 1: Covariable variable en el tiempo, X_{t-1}, y respuesta Y_t.

Estos modelos especifican un efecto beneficioso del tratamiento X_0 sobre el resultado en el momento 1 con una razón de odds logarítmica de -0,5 en el modelo $[Y_1|X_0]$. Sin embargo, el segundo modelo especifica que el tratamiento recibido en el momento 2 depende en gran medida del resultado en el momento uno. Para $X_0 = 0$ o $X_0 = 1$, si los pacientes

Cuadro 1: Recuentos esperados cuando se tratan inicialmente 500 sujetos, $X_0 = 1$, y 500 sujetos no reciben tratamiento, $X_0 = 0$, cuando el tratamiento en el momento 2, X_1, se predice por el resultado en el momento 1, Y_1, según el modelo dado por

X_0	**0**								**1**							
n	500								500							
Y_1	**0**				**1**				**0**				**1**			
n	311				189				365				135			
X_1	**0**		**1**		**0**		**1**		**0**		**1**		**0**		**1**	
n	194		117		71		118		227		138		51		84	
Y_2	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
n	142	52	96	21	27	44	59	59	166	61	113	25	19	32	42	42

tienen una respuesta inicial deficiente ($Y_1 = 1$), es más probable que reciban tratamiento en el momento 2 que si respondieran bien ($Y_1 = 0$). Finalmente, la respuesta en el segundo tiempo está fuertemente correlacionada con la respuesta inicial y está influenciada por el tratamiento en el tiempo 2.