

CONFERENCIAS STATA LATAM 2023

Aplicaciones estadísticas al servicio
de distintas áreas del conocimiento

GEE una metodología analítica para datos con medidas repetidas y correlacionadas



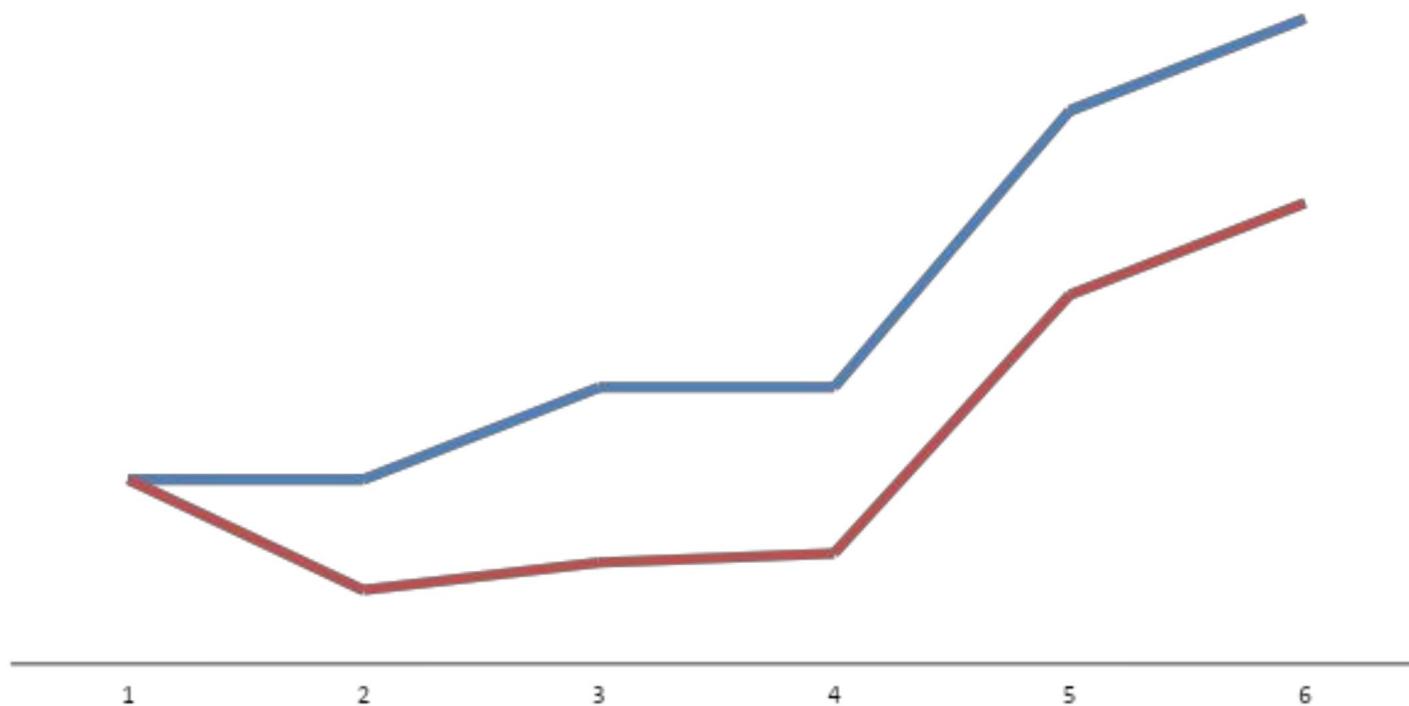
Algunos elementos relacionados

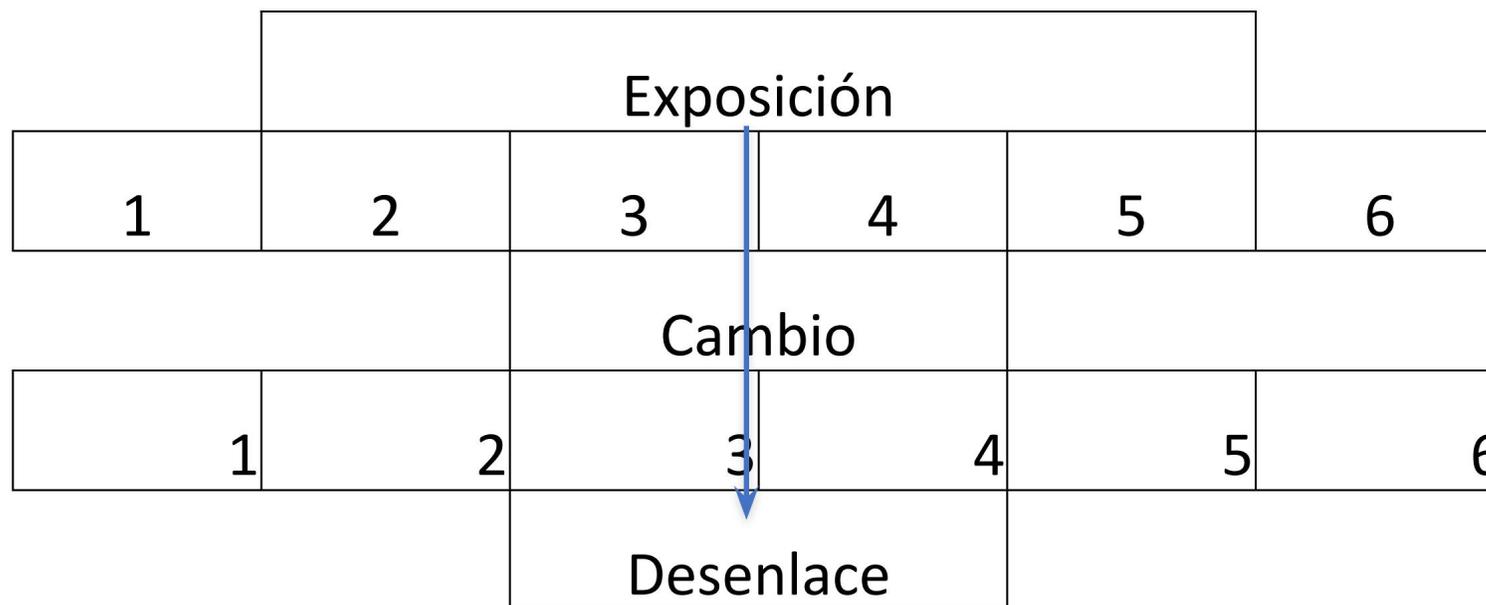
- Si $Y_{i:k:l}$ representa una observación desde el i^{th} miembro anidado dentro del k^{th} grupo que se encuentra anidado en la l^{th} condición, la anidación se representa por (:)

- Los datos en estas condiciones pueden por las jerarquías tener una correlación, lo que puede amenazar el supuesto de independencia de las medidas y además en estos grupos la medida repetida en el tiempo reduce variabilidad.

Aplicaciones estadísticas al servicio de distintas áreas del conocimiento

Un estudio longitudinal dentro y entre con seis medidas repetidas en dos grupos





Aplicaciones estadísticas al servicio de distintas áreas del conocimiento

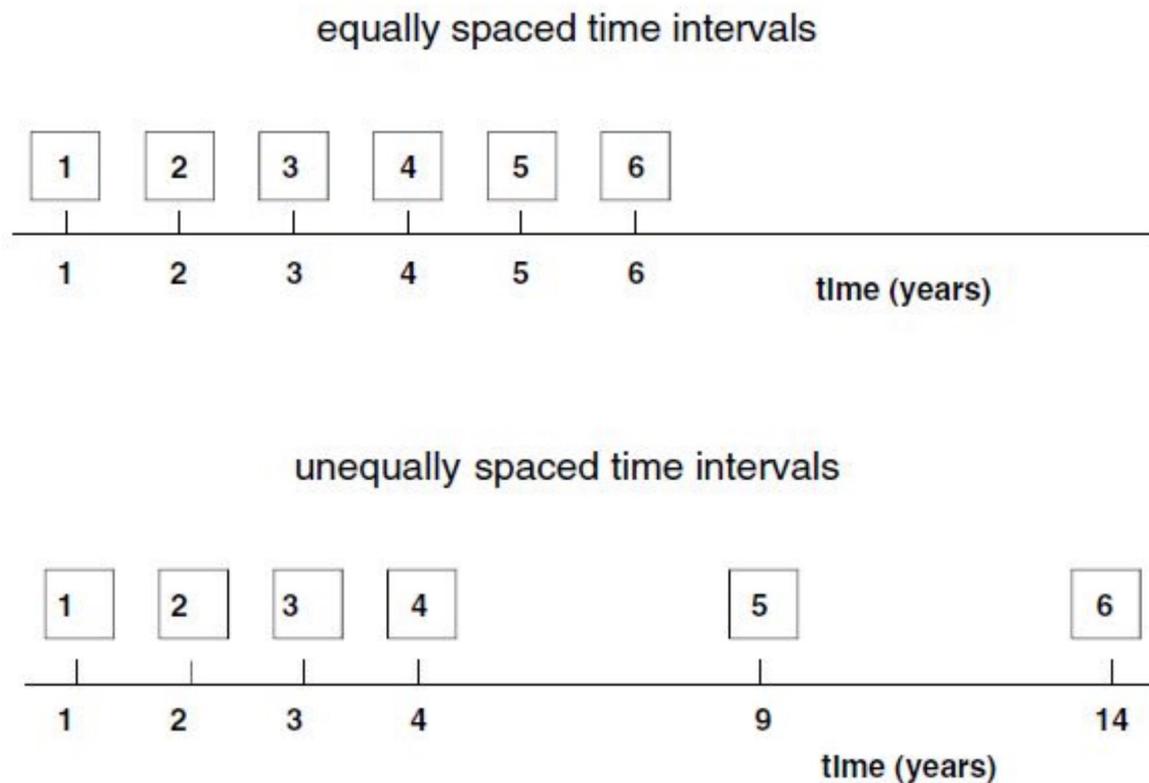


Figure 1.1. In the example dataset equally spaced time intervals and unequally spaced time intervals are used.

Datos longitudinales

- Los datos *longitudinales* presentan información acerca de qué pasa en un grupo de unidades de investigación (individuos, compañías, comunidades, ...) durante *distintos momentos a lo largo del tiempo*
- Usualmente son recogidos en estudios de diseño longitudinal, como los *estudios de cohortes*
- Los datos permiten la *comparación intra-unidad* a través del tiempo.

Relaciones Simples

- Regresión Y sobre X en un punto específico del tiempo
- Regresión de Y sobre X usando todos los datos en todos los puntos del tiempo

Relaciones Simples

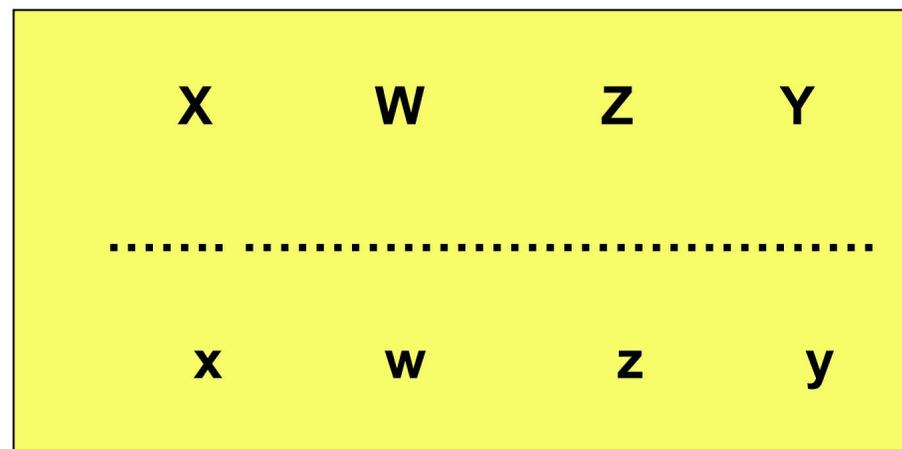
- Regresión de Y sobre X usando todos los datos en todos los puntos del tiempo de cada sujeto
- Regresión de Y sobre X usando todos los datos de todos los individuos dentro de cada conglomerado

Relaciones que nos puedan interesar

- Nivel macro

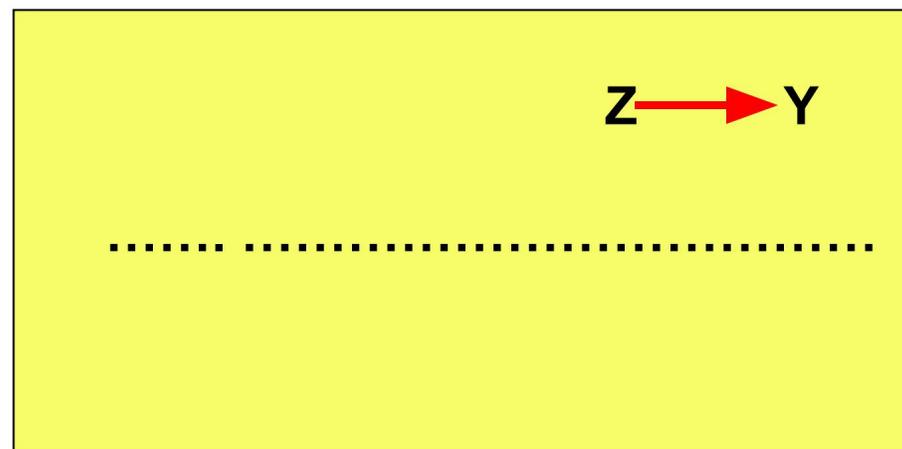
.....

- Nivel micro



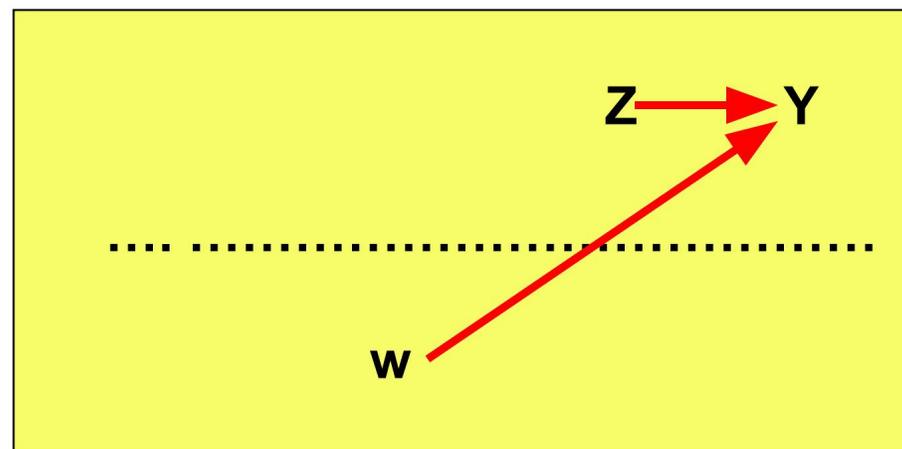
Relaciones que nos puedan interesar

- Nivel macro
.....
- Nivel micro



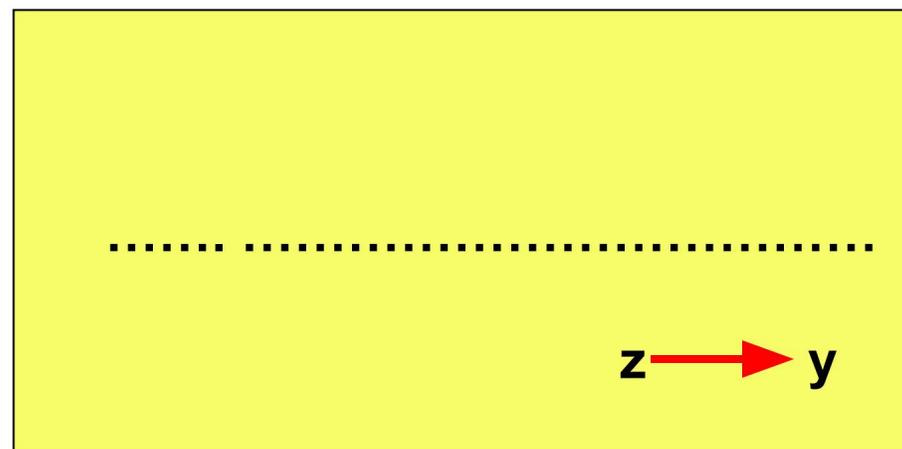
Relaciones que nos puedan interesar

- Nivel macro
.....
- Nivel micro



Relaciones que nos puedan interesar

- Nivel macro
.....
- Nivel micro

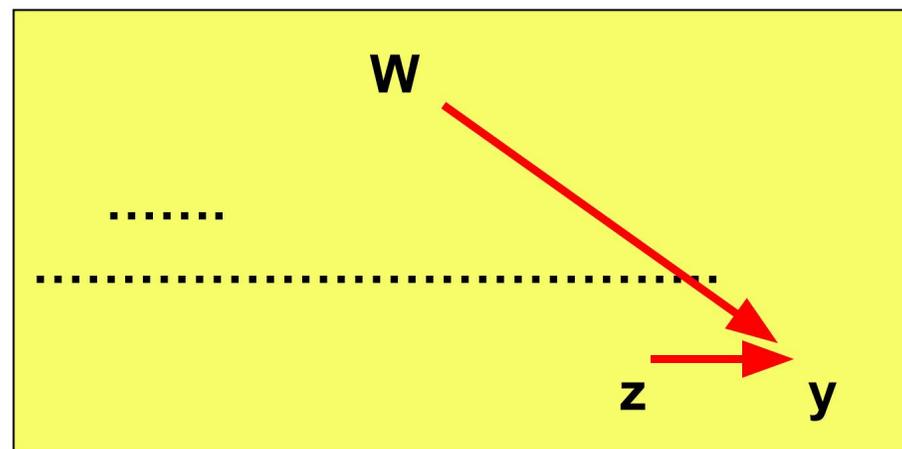


Relaciones que nos puedan interesar

- Nivel macro

.....

- Nivel micro

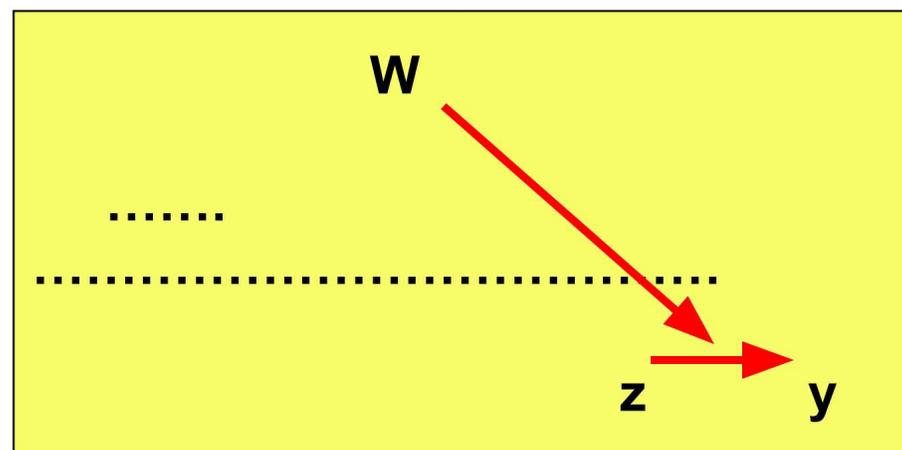


Relaciones que nos puedan interesar

- Nivel macro

.....

- Nivel micro



Enfoques

- Ignorar el problema
 - Usar todos los datos e ignorar la correlación entre las observaciones

$$Y = \alpha + \beta X$$

~ ~

Enfoques

- Simplificar el problema
 - Múltiples análisis transversales

$$Y_k = \alpha + \beta X \text{ a cada tiempo } k = 0, 1, \dots, 6$$

- Definir 'variables de cambio'

$$Y_6 - Y_0 = \alpha + \beta X$$

Enfoques

- Simplificar el problema
 - Usar solamente el resultado 'final' en el análisis

$$Y_6 = \alpha + \beta X$$

- Asumir una relación lineal con el tiempo:

$$Y = \alpha + \gamma t + \beta X$$

Enfoques

- Cambiar el problema
 - Re-definido como “análisis de sobrevivida”

$$\lambda(t) = \lambda_0(t) \exp^{\alpha + \beta X}$$

donde t = tiempo de Y alcanzar cierto umbral

Enfoque longitudinal

- Relación longitudinal verdadera

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \varepsilon_{it}$$

donde i es sujeto, t es tiempo, j es variable

El subíndice t implica que Y es medido repetidamente a lo largo del tiempo en i^{mo} sujeto, y que las variables X_j pueden variar también a lo largo del tiempo en cada sujeto

- Se puede incorporar un indicador del tiempo

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t + \varepsilon_{it}$$

- Se pueden agregar K covariables Z que dependen del tiempo y M covariables G que son tiempo-independientes

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t + \sum_{k=1}^K \beta_{3k} Z_{itk} + \sum_{m=1}^M \beta_{4m} G_{im} + \varepsilon_{it}$$

Enfoque GEE

- El modelo multinivel en general se puede describir como uno con efectos fijos y correlaciones entre los sujetos **dentro** del nivel macro
- Nos interesa conocer los valores de los efectos fijos β
- Nos interesa corregir por la correlación que pueda existir necesitamos imponer una estructura para las correlaciones

$$\underset{\sim}{Y} = \underset{\sim}{X} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

$$\underset{\sim}{\varepsilon} = \sigma_e^2 \begin{bmatrix} \underset{\sim}{\Sigma} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Sigma} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{\Sigma} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{\Sigma} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{\Sigma} \end{bmatrix}$$

Procedimientos de estimación

- Ecuaciones generalizadas para estimaciones (GEE)
 - Análisis simultáneo de las relaciones en el modelo
 - Basados en la 'cuasi-probabilidad' y en el proceso iterativo computacional
 - Toma en cuenta la **estructura de correlación** entre las múltiples observaciones en el individuo
 - Asume una 'correlación operativa' *a priori*

GEE

- Típicas *estructuras de correlación* para múltiples observaciones en el individuo
 - Independiente
 - Intercambiable
 - Estacionariamente m-dependiente
 - Autoregresiva
 - No estructurada

GEE

Estructura de correlación independiente

$$\Sigma \approx = \begin{array}{|c|c|c|c|c|c|} \hline 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

GEE

Estructura de correlación intercambiable

$$\Sigma \sim = \begin{array}{|c|c|c|c|c|c|} \hline 1 & \rho & \rho & \rho & \rho & \rho \\ \hline \rho & 1 & \rho & \rho & \rho & \rho \\ \hline \rho & \rho & 1 & \rho & \rho & \rho \\ \hline \rho & \rho & \rho & 1 & \rho & \rho \\ \hline \rho & \rho & \rho & \rho & 1 & \rho \\ \hline \rho & \rho & \rho & \rho & \rho & 1 \\ \hline \end{array}$$

También
llamado
*compound
symmetric*

GEE

Estructura de correlación estacionariamente m-dependiente

$$\Sigma_{\sim} = \begin{array}{|c|c|c|c|c|c|} \hline 1 & \rho_1 & \rho_2 & 0 & 0 & 0 \\ \hline \rho_1 & 1 & \rho_1 & \rho_2 & 0 & 0 \\ \hline \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & 0 \\ \hline 0 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \hline 0 & 0 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \hline 0 & 0 & 0 & \rho_2 & \rho_1 & 1 \\ \hline \end{array}$$

Asume 6 espacios de tiempo iguales de medición por individuo; m=2

GEE

Estructura de correlación autoregresiva

$$\Sigma \sim = \begin{array}{|c|c|c|c|c|c|} \hline 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \hline \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \hline \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \hline \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \hline \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \hline \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \\ \hline \end{array}$$

Asume 6 espacios de tiempo iguales de medición por individuo

GEE

Estructura de correlación no estructurada

$$\Sigma \sim = \begin{array}{|c|c|c|c|c|c|} \hline 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} \\ \hline \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} \\ \hline \rho_{13} & \rho_{23} & 1 & \rho_{34} & \rho_{35} & \rho_{36} \\ \hline \rho_{14} & \rho_{24} & \rho_{34} & 1 & \rho_{45} & \rho_{46} \\ \hline \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 & \rho_{56} \\ \hline \rho_{16} & \rho_{26} & \rho_{36} & \rho_{46} & \rho_{56} & 1 \\ \hline \end{array}$$

GEE

- Esencialmente, GEE trata la estructura de correlación intra-individuo como una variable 'de molestia', y corrige la dependencia entre las observaciones en el sujeto según la siguiente ecuación

$$Y_{it} = \beta_0 + \sum_{j=1}^J \beta_{1j} X_{itj} + \beta_2 t + \dots + \text{CÓRR}_{it}^* + \varepsilon_{it}$$

Enfoques

Usar GEE

- Tratar la estructura de correlación entre tiempos como una variable ‘de molestia’, y corregir la dependencia entre las mediciones en el sujeto
- Usar ‘intercambiable’ como la estructura de correlación operativa por defecto
- En oposición al caso de resultados continuos, los resultados **son** robustos a la mala-especificación de la estructura de correlación!
- Si usa ‘independencia’, el coeficiente de la regresión usando GEE es **exactamente** el mismo de la regresión logística, pero el error estándar es más grande en la GEE

Enfoques

- Para variables binarias, los modelos usando GEE darán resultados diferentes para las estimaciones de los parámetros que los modelos de efectos aleatorios
- ¿Cómo elegir?
- Basados en la pregunta de investigación:
 - Si está interesado en la relación entre la variable resultado y las variables predictoras GEE
 - Si está interesado en el desarrollo a lo largo del tiempo de la variable resultado modelo de efectos mixtos

CONFERENCIAS STATA LATAM 2023

Aplicaciones estadísticas al servicio
de distintas áreas del conocimiento

Conozca más sobre STATA escaneando el código QR.

