Análisis de Datos Longitudinales

Maestría en Demografía y Estudios de Población

Alejandra Marroig

Instituto de Estadística, Universidad de la República, Uruguay

August 2025

Contenidos

En esta clase veremos

1. Contexto de investigaciones del cambio en el tiempo
2. Análisis exploratorio del cambio con datos longitudinales
3. **Análisis de datos con el modelo multinivel**
4. Aspectos avanzados de análisis de datos longitudinales (no linealidades, flexibilización de la variable tiempo, otros)

Modelo multinivel para el cambio en el tiempo

El modelo estadístico para representar el proceso de cambio con datos longitudinales debe poder dar respuesta a dos tipos de preguntas:

Preguntas del nivel 1: cambios en el tiempo intra individuos (*within-individual change*)

Preguntas del nivel 2: diferencias entre individuos en el cambio (*interindividual difference in change*)

Modelo lineal para el cambio individual

Modelo nivel 1:

$$Y_{ij} = \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij}$$

i representa cada individuo $i = 1, \ldots, N$

j representa cada ocasión $j = 1, \ldots, J$.

Partes del modelo

- parte estructural $[\pi_{0i} + \pi_{1i} time_{ij}]$, verdadera trayectoria individual
- parte estocástica ϵ_{ij}, no observable, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

Nota: Puede relajarse el supuesto de la parte estocástica y considerar autocorrelación y heteroscedasticidad en los residuos.

Comentarios: Modelo lineal para el cambio individual

Importante:

- Este modelo puede utilizarse para datos donde el tiempo y espaciamiento entre olas difieren entre individuos.
- Se asume implícitamente que todas las trayectorias de cambio individuales verdaderas tienen una forma algebraica común, pero **no** todos tienen exactamente la misma trayectoria.

Diferencias interindividuales en el cambio

Caso sencillo: característica invariante que separa los individuos en 2 grupos $X_i = 0$ o $X_i = 1$.

Diferencias interindividuales en el cambio

Caso sencillo: característica invariante que separa los individuos en 2 grupos $X_i = 0$ o $X_i = 1$.

Modelo nivel 2:

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \gamma_{01} X_i + b_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} X_i + b_{1i} \end{aligned}$$

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right)$$

Modelo lineal multinivel para el cambio en el tiempo

Modelo nivel 1 y nivel 2:

$$
\begin{aligned}
Y_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\
\pi_{0i} &= \gamma_{00} + \gamma_{01} X_i + b_{0i} \\
\pi_{1i} &= \gamma_{10} + \gamma_{11} X_i + b_{1i}
\end{aligned}
$$

con $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ y $b \sim N(o, D)$

Coeficientes del modelo lineal multinivel

Coeficientes del modelo lineal multinivel

Nivel 1:

Coeficientes del crecimiento lineal individual

- π_{0i}: intercepto de la verdadera trayectoria para i
- π_{1i}: pendiente de la verdadera trayectoria para i

Componente de varianza

σ_ϵ^2 varianza residual a través de todas las ocasiones de medida de i

Coeficientes del modelo lineal multinivel

Nivel 1:

Coeficientes del crecimiento lineal individual

- π_{0i}: intercepto de la verdadera trayectoria para i
- π_{1i}: pendiente de la verdadera trayectoria para i

Componente de varianza

σ_ϵ^2 varianza residual a través de todas las ocasiones de medida de i

Nivel 2:

Efectos fijos

- γ_{00}: promedio poblacional de interceptos individuales (π_{0i}) para $X_i = 0$
- γ_{01}: diferencia en promedio poblacional de π_{0i} por cambio unitario en X_i
- γ_{10}: promedio poblacional de pendientes individuales (π_{1i}) para $X_i = 0$
- γ_{11}: diferencia en promedio poblacional de π_{1i} por cambio unitario en X_i

Componente de varianza

- σ_0^2: varianza residual de π_{0i} a través de los individuos
- σ_1^2: varianza residual de π_{1i} a través de los individuos
- σ_{01}: covarianza π_{0i} y π_{1i} a través de los individuos

Parámetros a estimar

4 efectos fijos, 3 varianzas de efectos aleatorios y la varianza residual del modelo $\rightarrow$ 8 parámetros.

Especificación compuesta: modelo lineal multinivel condicional

Modelo lineal multinivel condicional:

$$
\begin{aligned}
Y_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\
\pi_{0i} &= \gamma_{00} + \gamma_{01} X_i + b_{0i} \\
\pi_{1i} &= \gamma_{10} + \gamma_{11} X_i + b_{1i}
\end{aligned}
$$

con $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ y $b \sim N(o, D)$

Especificación compuesta: modelo lineal multinivel condicional

Modelo lineal multinivel condicional:

$$
\begin{aligned}
Y_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\
\pi_{0i} &= \gamma_{00} + \gamma_{01} X_i + b_{0i} \\
\pi_{1i} &= \gamma_{10} + \gamma_{11} X_i + b_{1i}
\end{aligned}
$$

con $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ y $b \sim N(o, D)$

$$Y_{ij} = \gamma_{00} + \gamma_{01} X_i + b_{0i} + (\gamma_{10} + \gamma_{11} X_i + b_{1i})\, time_{ij} + \epsilon_{ij}$$

Especificación compuesta: modelo lineal multinivel incondicional

Modelo lineal multinivel incondicional:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + b_{0i} \\ \pi_{1i} &= \gamma_{10} + b_{1i} \end{aligned}$$

con $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ y $b \sim N(o, D)$

Especificación compuesta: modelo lineal multinivel incondicional

Modelo lineal multinivel incondicional:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + b_{0i} \\ \pi_{1i} &= \gamma_{10} + b_{1i} \end{aligned}$$

con $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ y $b \sim N(o, D)$

$$Y_{ij} = \gamma_{00} + b_{0i} + (\gamma_{10} + b_{1i})\, time_{ij} + \epsilon_{ij}$$

Modelo multinivel

Modelo multinivel:

$$Y_i = X_i\gamma + Z_i b_i + \epsilon_i$$

con $\epsilon_i \sim N(o, \Sigma_{\epsilon_i})$ y $b_i \sim N(o, D)$ independientes.

Modelo multinivel

Modelo multinivel:

$$Y_i = X_i\gamma + Z_i b_i + \epsilon_i$$

con $\epsilon_i \sim N(o, \Sigma_{\epsilon_i})$ y $b_i \sim N(o, D)$ independientes.

Especificación del modelo con trayectoria lineal condicional:

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_i + b_{0i} + (\gamma_{10} + \gamma_{11}X_i + b_{1i})\, time_{ij} + \epsilon_{ij}$$

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_i + \gamma_{10} time_{ij} + \gamma_{11}X_i time_{ij} + b_{0i} + b_{1i} time_{ij} + \epsilon_{ij}$$

Método de estimación

- Máxima verosimilitud
- Máxima verosimilitud restricta

Ejemplo de Singer & Willett

Modelo nivel 1 y nivel 2:

$$
\begin{aligned}
cog_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\
\pi_{0i} &= \gamma_{00} + \gamma_{01} prog_i + b_{0i} \\
\pi_{1i} &= \gamma_{10} + \gamma_{11} prog_i + b_{1i}
\end{aligned}
$$

Recordar que $time_{ij} = age_{ij} - 1$ donde la edad está medida en años.

Resultados de la estimación

Efectos fijos			Est	aSE	z
π_{0i}	Intercept	γ_{00}	107.84***	2.04	52.97
	Prog	γ_{01}	6.85*	2.71	2.53
π_{1i}	Intercept	γ_{10}	-21.13***	1.89	-11.18
	Prog	γ_{11}	5.27*	2.52	2.09

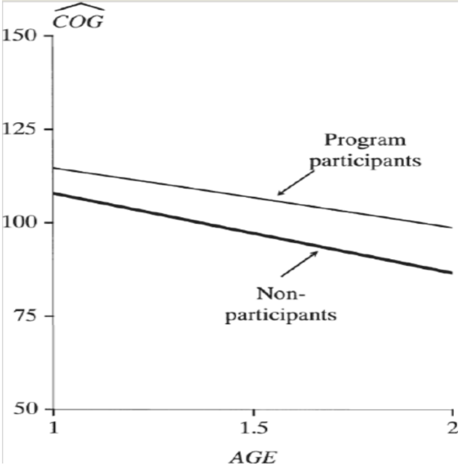
Varianza			Est	aSE	z
Nivel 1:	Intraindiv	σ_ϵ^2	74.24***	10.34	7.17
Nivel 2:	Intercept	σ_0^2	124.64***	27.38	4.55
	Pendiente	σ_1^2	12.29	30.50	0.40
	Covar	σ_{01}	-36.41	22.74	-1.60

Interpretación

- Se estima que el estado cognitivo promedio al inicio (con 1 año de edad) para quienes no participaron del programa es 107.84
- Se estima que el estado cognitivo promedio al inicio para los participantes es 6.85 puntos mayor (114.69) que para los que no participaron.
- Se estima que la tasa de cambio anual verdadera para quienes no participaron fue -21.13 puntos menor (cae 21.13 puntos en promedio por año)
- Se estima que la tasa de cambio anual verdadera para los participantes es en promedio 5.27 puntos mayor (-15.86) respecto a los que no participaron

$\Rightarrow$ La función cognitiva de ambos grupos decrece con el tiempo y el programa enlentece la tasa de cambio (*rate of decline*).

Resultados del modelo multinivel ajustado

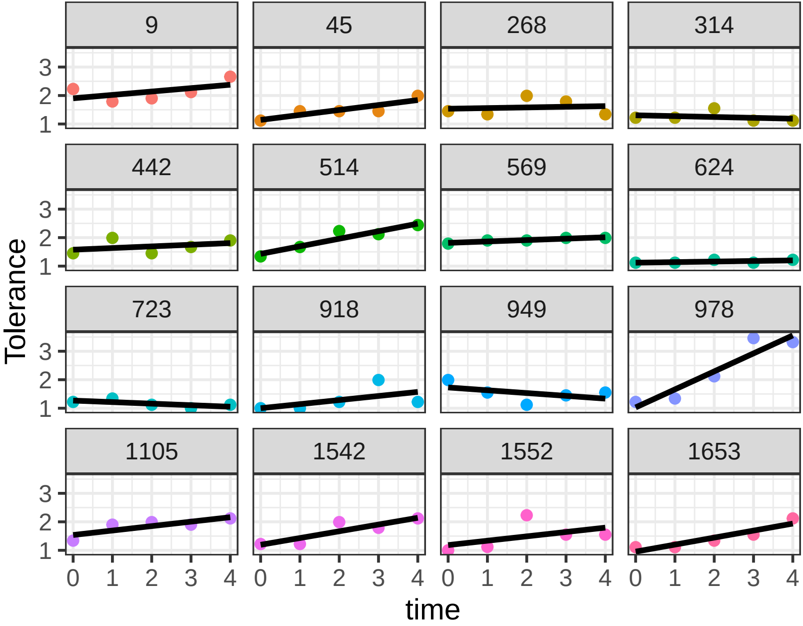


¿Y los datos?

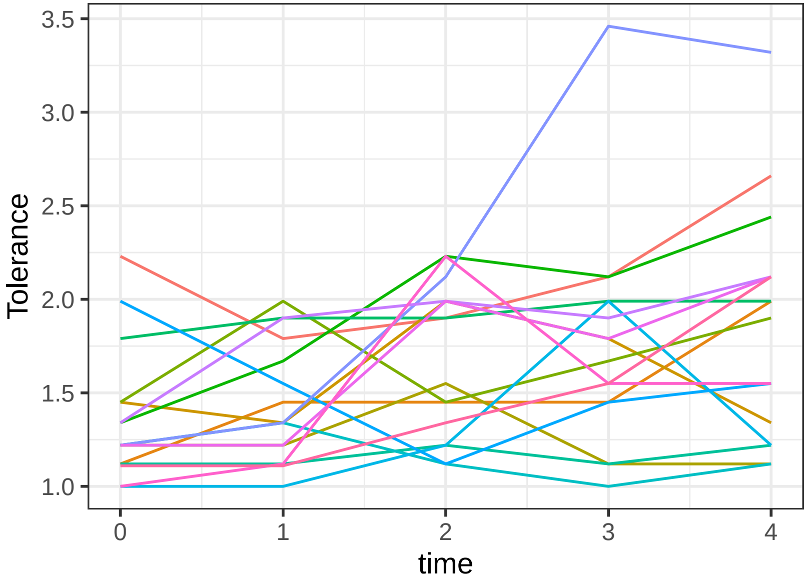
¿Y los datos?

- 16 niñas/os de la Encuesta Nacional sobre la Juventud (NYS).
- Edades: 11, 12, 13, 14 y 15 años
- *Outcome*: `tolerance`

```
##    id age tolerance male exposure time idf
## 1   9  11      2.23    0     1.54    0   9
## 2   9  12      1.79    0     1.54    1   9
## 3   9  13      1.90    0     1.54    2   9
## 4   9  14      2.12    0     1.54    3   9
## 5   9  15      2.66    0     1.54    4   9
## 6  45  11      1.12    1     1.16    0  45
## 7  45  12      1.45    1     1.16    1  45
## 8  45  13      1.45    1     1.16    2  45
## 9  45  14      1.45    1     1.16    3  45
## 10 45  15      1.99    1     1.16    4  45
```


9
45
268
314
442
514
569
624
723
918
949
978
1105
1542
1552
1653
Tolerance
1
2
3
0
1
2
3
4
time

Spaghetti plot



Preguntas específicas en modelo multinivel

Modelo nivel 1 y nivel 2:

$$\begin{aligned} tol_{ij} &= \pi_{0i} + \pi_{1i} time_{ij} + \epsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + \gamma_{01} male_i + b_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} male_i + b_{1i} \end{aligned}$$

con $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ y $b_i \sim N(o, D)$ independientes.

1. ¿Cómo cambia la tolerancia de jóvenes en el tiempo?
2. ¿Hay diferencias en las trayectorias de la tolerancia según sexo?

Modelo incondicional

```
model0      <- lm(tolerance ~ time,
                  data=tolerance.pp)
library(nlme)
model1      <- lme(tolerance ~ time,
                   random= ~time | id,
                   data=tolerance.pp,
                   method="ML")
anova(model1,model0)

library(lme4)
model1_lme4 <- lmer(tolerance ~ time + (1 + time | id),
                    data=tolerance.pp,
                    REML=FALSE)
```

Modelo con covariable

```
model0cov   <- lm(tolerance ~ time + male,
                  data=tolerance.pp)
library(nlme)
model1cov   <- lme(tolerance ~ time + male,
                   random= ~time | id,
                   data=tolerance.pp,
                   method="ML")
anova(model1cov,model0cov)

library(lme4)
model1_lme4c<- lmer(tolerance ~ time + male + (1 + time | id),
                    data=tolerance.pp,
                    REML=FALSE)
```

Interpretación de los parámetros y bondad de ajuste

```
library(nlme)
model1      <- lme(tolerance ~ time + male,
                   random= ~time | id,
                   data=tolerance.pp,
                   method="ML")
summary(model1)
ranef(model1)
fixef(model1)
coef(model1)
```

Pruebas de hipótesis de un único efecto fijo y comparación de modelos

$$\begin{aligned} H_0 &: \ \gamma = 0 \\ H_1 &: \ \gamma \neq 0 \end{aligned}$$

Se puede hacer un contraste de hipótesis asintótico:

$$z = \frac{\hat{\gamma}}{ase(\hat{\gamma})}$$

- Se pueden comparar modelos con los criterios de información ya que la estimación se hace por ML (AIC, BIC, LR). También está disponible un análisis de varianzas que se puede usar con cautela `anova()`

¡¡Gacias!!

alejandra.marroig@fcea.edu.uy