

Datos en panel II

Gabriel Montes-Rojas

Modelo de componentes del error en una dirección

En un panel longitudinal el mismo individuo es observado a lo largo del tiempo.

$$y_{it} = \alpha + X'_{it}\beta + u_{it}$$

$i = 1, 2, \dots, N$ es el índice de individuos (firmas, familias, hogares, países),

$t = 1, 2, \dots, T$ es el índice de tiempo,

α es un escalar, β es $K \times 1$, X_{it} es la observación it de las K variables explicativas.

El error tiene esta estructura:

$$u_{it} = \mu_i + v_{it}$$

es el **error compuesto**. En inglés: **one-way error components model**.

- μ_i : efecto individual no observado que captura todos los factores constantes a lo largo del tiempo en y_{it} ,
- v_{it} : errores idiosincráticos o shocks.

Modelo de componentes del error en dos direcciones

Consideremos el modelo

$$y_{it} = \alpha + X'_{it}\beta + u_{it}$$

$i = 1, 2, \dots, N$ es el índice de individuos (firmas, familias, hogares, países),

$t = 1, 2, \dots, T$ es el índice de tiempo,

α es un escalar, β es $K \times 1$, X_{it} es la observación it de las K variables explicativas.

$$u_{it} = \mu_i + \lambda_t + v_{it}$$

es el **error compuesto**. En inglés: **two-way error components model**.

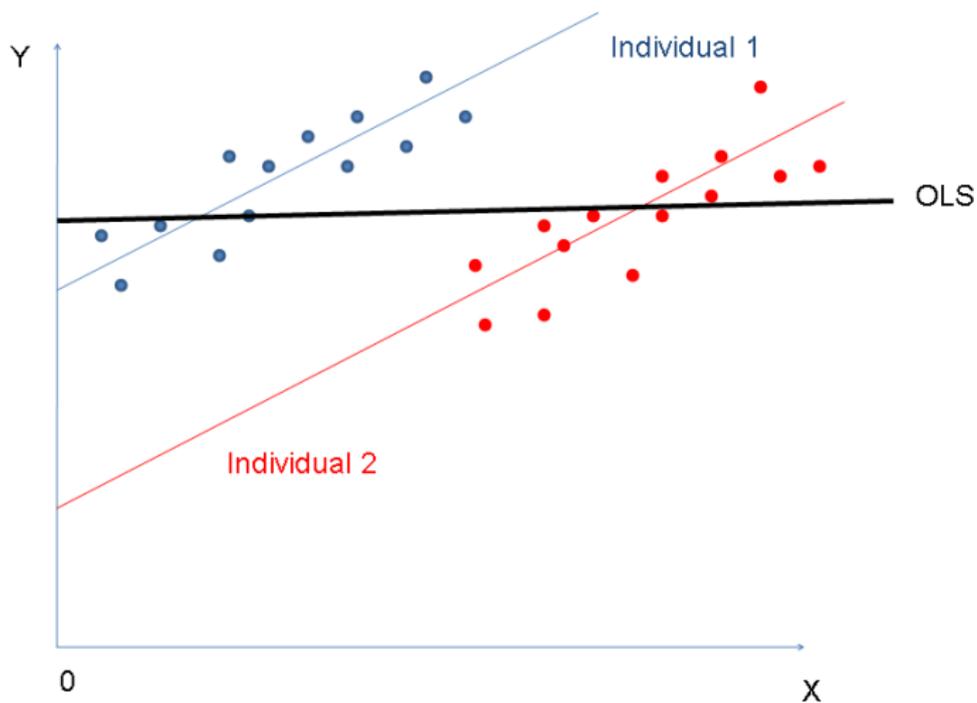
- μ_i : efecto individual no observado que captura todos los factores constantes a lo largo del tiempo en y_{it} ,
- λ_t : efecto temporal no observado que captura todos los factores constantes a lo de los individuos en y_{it} ,
- v_{it} : errores idiosincráticos o shocks.

El problema es que si $cov(x_{it}, u_{it}) \neq 0 \Rightarrow$ MCO es sesgado. ¿Por qué?

Los siguientes estimadores se proponen como soluciones a este problema:

- **Estimador en primeras diferencias**
- **Efectos fijos**

Panel data



Estimador en diferencias (first differences, FD)

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \mu_i + v_{i,t}$$

$$y_{i,t-1} = \beta_0 + \beta_1 x_{i,t-1} + \mu_i + v_{i,t-1}$$

$$\Rightarrow \Delta y_{it} = \beta_1 \Delta x_{it} + \Delta v_{it}$$

donde Δ es el operador de diferencias, $\Delta y_{it} = y_{it} - y_{it-1}$

¡Lo que pasa es que μ_i desaparece, entonces se acaba el problema!

Efectos fijos (fixed-effects, FE)

$$y_{it} = \beta_0 + \beta_1 x_{it} + \mu_i + v_{it}$$

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \bar{\mu}_i + \bar{v}_i$$

$$\Rightarrow \tilde{y}_{it} = \beta_1 \tilde{x}_{it} + \tilde{v}_{i,t}$$

donde $\tilde{\cdot}$ es una transformación que se aplica a cada individuo (se usa en inglés **within** transformation), $\tilde{y}_{it} = y_{it} - \bar{y}_i$

¡Otra vez desaparece μ_i ! (porque $\bar{\mu}_i = T^{-1} \sum_{t=1}^T \mu_i = T^{-1} T \mu_i = \mu_i$)

Efectos fijos (fixed-effects, FE)

En los estimadores de efectos fijos μ_i se asumen como **parámetros fijos a ser estimados** y $v_{it} \sim IID(0, \sigma_v^2)$. $x_{it} \perp v_{it}, \forall i, t$.

Otro modo de ver los modelos FE es $E(v_{it}|x_{it}) \neq 0$ pero que $E(v_{it}|x_{it}, \mu_i) = 0$.
Entonces necesitamos estimar μ_i .

Efectos fijos (fixed-effects, FE)

Notar que $\hat{\beta}_{FE}$ es equivalente a un modelo MCO con una dummy para cada individuo i , $\hat{\beta}_{LSDV}$.

La prueba es una aplicación del Teorema de **Frisch-Waugh-Lovell**

$$y = X_1\beta_1 + X_2\beta_2 + u$$

con estimadores OLS $\hat{\beta} = [\hat{\beta}_1 \hat{\beta}_2]$.

Usemos $M_1y = M_1X_2\beta_2 + M_1u$ donde M_1 es la proyección residual de X_1 , el teorema muestra que

$$\hat{\beta}_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 y$$

Prueba: Consideremos $y = P_X y + M_X y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + M_X y$. Multiplicar ambos lados por $X_2' M_1 y$ obtenemos $X_2' M_1 y = X_2' M_1 X_2 \hat{\beta}_2$. (usando $M_1 X_1 = 0$ y $M_X M_1 X_2 = 0$). Resolver.

Efectos fijos (fixed-effects, FE)

- Podemos plantear $\hat{\mu}_i = \hat{y}_i - \hat{\beta}_{FE} \hat{x}_i$ como el estimador de los "efectos fijos".
- Si $T \rightarrow \infty$, $(\hat{\mu}_{FE}, \hat{\beta}_{FE})$ son estimadores consistentes e insesgados.
- Pero si T está fijo y $N \rightarrow \infty$, sólo $\hat{\beta}_{FE}$ es consistente, aunque ambos son insesgados. El problema es conocido como el problema de parámetros incidentales de Neyman y Scott (1948).
- Contraste de efectos fijos: correr el modelo LSDV, y contrastar conjuntamente por la significatividad de las dummies.

Efectos fijos (fixed-effects, FE)

Usar efectos fijos tiene algunas desventajas:

- Hay un costo en comparación con MCO. La transformación within es equivalente a estimar una dummy para cada individuo. O sea estimar N parámetros adicionales. Más parámetros significa menos precisión.
- Entonces eso afecta los grados de libertad y por ende la precisión de lo que estimamos. Los grados de libertad son $NT - N - K$
- También la transformación within (y first-differences) elimina TODO aquello que esta fijo para cada individuo. Entonces no se puede medir por ejemplo el efecto de SEXO, para individuos, o CONTINENTE para países.

Efectos aleatorios (random-effects, RE)

Un modelo alternativo es el de efectos aleatorios. Tiene un supuesto MUY importante y restrictivo: $cov(X, \mu) = 0$.
Consideremos el modelo

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_{it} = \beta_0 + \beta_1 x_{it} + \mu_i + v_{it}$$

En este caso hay **correlación serial**:

$Cov(u_{it}, u_{ij}) = Cov(\mu_i + v_{i,t}, \mu_i + v_{i,j}) = Var(\mu_i) \neq 0$. El estimador de RE tiene en cuenta esta particularidad y produce un estimador eficiente:

$$\hat{\beta}_{RE} = (X' \Omega^{-1} X)^{-1} (X' \Omega^{-1} Y)$$

donde $\Omega = E(uu')$.

Una de las ventajas de RE es que se pueden reincorporar variables fijas por individuos:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \mu_i + v_{it}$$

donde z captura todas las variables que están fijas para cada individuo y que se pueden observar.

Efectos aleatorios (random-effects, RE)

En el modelo RE $\mu_i \sim iid(0, \sigma_\mu^2)$, $v_{it} \sim iid(0, \sigma_v^2)$, $\mu_i \perp v_{it}$. Además tenemos $X_{it} \perp \mu_i$ y $X_{it} \perp v_{it}$ para todo i y t .

$$\Omega = E(uu') = Z_\mu E(\mu\mu')Z_\mu' + E(vv') = \sigma_\mu^2(I_N \otimes J_T) + \sigma_v^2(I_N \otimes I_T).$$

Ω tiene esta estructura:

$$\begin{aligned} \text{cov}(u_{it}, u_{js}) &= \sigma_\mu^2 + \sigma_v^2 && \text{para } i = j, t = s \\ &= \sigma_\mu^2 && \text{para } i = j, t \neq s \\ &= 0 && \text{para } i \neq j \end{aligned}$$

Otra forma de verlo es como un modelo de **equicorrelación** intra cluster:

$$\begin{aligned} \text{correl}(u_{it}, u_{js}) &= 1 && \text{para } i = j, t = s \\ &= \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_v^2) := \rho && \text{para } i = j, t \neq s \\ &= 0 && \text{para } i \neq j \end{aligned}$$

Contraste de Hausman

El estimador de efectos fijos es siempre **consistente**. Sin embargo, el de efectos aleatorios es válido si las Xs no están correlacionadas con los efectos individuales (μ_i). Entonces, un contraste de la validez de efectos aleatorios es

$$H_0 : \hat{\beta}_{FE} - \hat{\beta}_{RE} = 0$$

$$H_A : \hat{\beta}_{FE} - \hat{\beta}_{RE} \neq 0$$

Este es el llamado contraste de Hausman.

Contraste de Hausman

- Notemos que el estimador RE es MCG, entonces es el de menor varianza.
- Por lo tanto $\text{var}(\hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})$ es una matriz definida positiva.
- Se puede probar que $\text{var}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = \text{var}(\hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})$.
- Entonces, definamos

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' [\text{var}(\hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})] (\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

tiene una distribución χ_K^2 bajo la nula.

Contraste de Mundlak

(Hsiao, 2003, sec. 3.2; Wooldridge, 2012, sec. 10.7.3)

- Mundlak (1978) propone usar un modelo de regresión que contenga los promedios de individuos i de las variables que varían en i y t :

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \beta_3 \bar{x}_i + \mu_i + v_{it}$$

- Entonces un contraste de $H_0 : \beta_3 = 0$ es un contraste por la validez de RE.

¿Cómo implementar paneles en STATA?

- Los datos hay que organizarlos:

id	tiempo	YVAR	XVAR
1	1	y ₁₁	x ₁₁
1	2	y ₁₂	x ₁₂
1	3	y ₁₃	x ₁₃
2	1	y ₂₁	x ₂₁
2	2	y ₂₂	x ₂₂
2	3	y ₂₃	x ₂₃

Es muy importante que no haya valores repetidos.

- Primero STATA tiene que identificar que se trata de datos en paneles. Para eso se necesita una variable numérica, ej. `id`, que identifica el individuo. Luego, `iis id`
- Pero si se tiene una muestra longitudinal con una estructura de series de tiempo, con una variable `tiempo`, `tsset id tiempo`
- Nota: la variable de tiempo tiene que ser en números discretos consecutivos. O sea, $t = -2, -1, 0, 1, \dots$

¿Cómo implementar paneles en STATA?

- Entonces estamos listos para usar datos en paneles:
- `reg D.y D.x1 D.x2 D.x3` (**modelo en diferencias**)
- `xtreg y x1 x2 x3, fe` (**modelo de efectos fijos**)
- `xi: reg y x1 x2 x3 i.id` (**lo mismo pero implementado "a mano" con dummies para id**) [Nota: comparar con una regresión de las variables transformadas within. ¿Cuál sería el problema con este modelo?]
- `xtreg y x1 x2 x3, re` (**modelo de efectos aleatorios**)
- `xtreg y x1 x2 x3, be` (**modelo between**)
- `xi: xtreg y x1 x2 x3 i.tiempo, fe` (**modelo de efectos fijos, two-way**)

¿Cómo implementar paneles en STATA?

- Contraste de Hausman test
xtreg y x1 x2 x3, fe
est store fe
xtreg y x1 x2 x3, re
est store re
hausman fe re

¿Cómo implementar paneles en STATA?

- <http://www.ats.ucla.edu/stat/stata/examples/eacspd/chapter10.htm>
- <http://www.wiley.com/legacy/wileychi/baltagi/datasets.html>
- `textttwebuse grunfeld`