Datos en panel

Gabriel Montes-Rojas

3 tipos de estructuras de datos

- Corte transversal (Cross section) Muestra de individuos, hogares, firmas, países, etc. que se toman en un momento dado del tiempo.

  $\{y_i, x_i\}_{i=1}^{n}$, donde i representa individuos.

  Ej.: EPH de 2013, datos de PBI en 2013 para muchos países.
- Series de tiempo Muestra por varios periodos del mismo individuo, pais, firma, etc.

  $\{y_t, x_t\}_{t=1}^{T}$, donde t es tiempo.

  Ej.: Inflación en la Argentina.
- Datos en panel Combinación de las dos anteriores.

  $\{y_{it}, x_{it}\}_{i=1,t=1}^{n,T}$, donde i representa individuos y t tiempo.

  1. **Cortes transversales independientes**
  2. **Muestra longitudinal**

Modelo de componentes del error en una dirección

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

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

$i = 1, 2, ..., N$ es el índice de individuos (firmas, familias, hogares, países),
$t = 1, 2, ..., T$ 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:

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

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},
- ν_{it}: errores idiosincráticos o shocks.

Modelo de componentes del error en dos direcciones

Consideremos el modelo

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

$i = 1, 2, ..., N$ es el índice de individuos (firmas, familias, hogares, países),
$t = 1, 2, ..., T$ es el índice de tiempo,
α es un escalar, β es $K \times 1$, X_{it} es la observación itde las K variables explicativas.

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

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},
- ν_{it}: errores idiosincráticos o shocks.

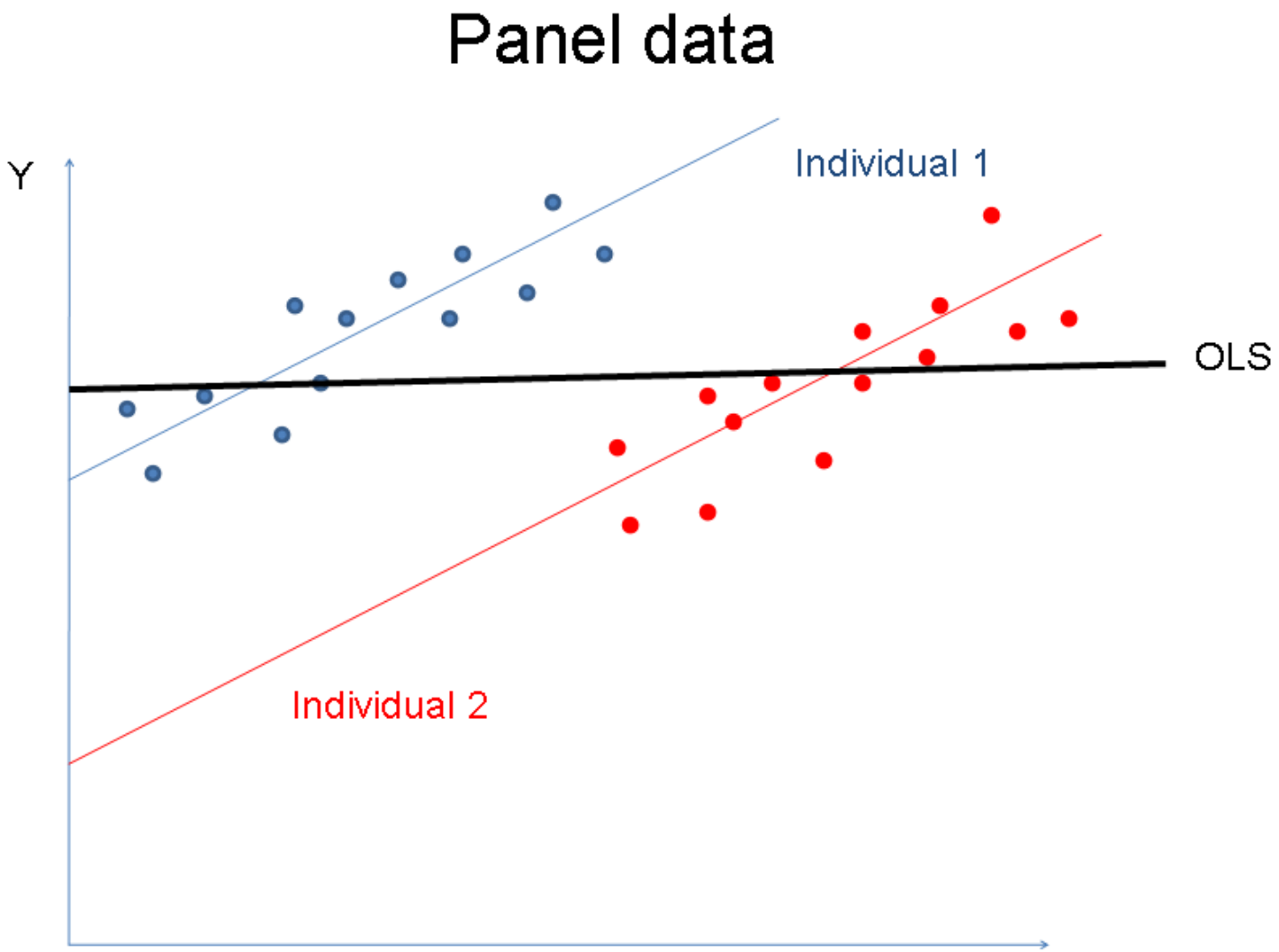
Inconsistencia de OLS

Si $cov(x_{it}, \mu_i) \neq 0 \Rightarrow$ OLS es sesgado e inconsistente. ¿Por qué?
Supongamos un modelo de regresión simple:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \mu_i + \nu_{it} \; i = 1, 2, ..., N, t = 1, 2, ..., T,$$

donde $E[\nu_{it}|x_{it}] = 0, \forall i, t$.
Probar:
1. $\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x})(y_{it}-\bar{y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x})^2} \xrightarrow{p} \beta_1 + \frac{cov(x_{it},\mu_i)}{var x_{it}}$ cuando $N, T \to \infty$.
2. $E[\hat{\beta}_1^{OLS}|X] = \beta_1 + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} E[\mu_i|X](x_{it}-\bar{x})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x})^2}$.

Inconsistencia de OLS



Los siguientes estimadores se proponen como soluciones a este problema:

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

Estimador en primeras diferencias (first differences,FD)

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

$$y_{it-1} = \beta_0 + \beta_1 x_{it-1} + \mu_i + \nu_{it-1}$$

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

donde Δ es el operador de diferencias, $\Delta y_{it} = y_{it} - y_{it-1}$.
Definamos $\hat{\beta}_1^{FD} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta x_{it}\Delta y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta x_{it}^2}$. Probar que es un estimador insesgado y consistente cuando $N, T \to \infty$.

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

Efectos fijos (fixed-effects,FE)

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

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

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

donde ˜ es una transformación que se aplica a cada invididuo (se usa en inglés **within** transformation), $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\bar{y}_{it} = T^{-1} \sum_{t=1}^{T} y_{it}$.

Definamos $\hat{\beta}_1^{FE} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}^2}$. Probar que es un estimador insesgado y consistente cuando $N, T \to \infty$.

¡Otra vez deseparece 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* $\nu_{it} \sim IID(0, \sigma_\nu^2)$. $x_{it} \perp\!\!\!\perp \nu_{it}, \forall i, t$.
Otro modo de ver los modelos FE es $E(\nu_{it}|x_{it}) \neq 0$ *pero que* $E(\nu_{it}|x_{it}, \mu_i) = 0$.
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**. Supongamos el modelo

$$\boldsymbol{y} = \boldsymbol{X_1}\beta_1 + \boldsymbol{X_2}\beta_2 + \boldsymbol{u}$$

con estimadores OLS $\hat{\beta} = [\hat{\beta_1}\hat{\beta_2}]$.
El teorema muestra que

$$\hat{\beta}_1 = (\boldsymbol{X}_1'\boldsymbol{M_2}\boldsymbol{X_1})^{-1}\boldsymbol{X}_1'\boldsymbol{M_2}\boldsymbol{y}$$

Si tomamos que X_2 es el conjunto de las **dummies** por individuo y X_1 son las variables de control, se puede demostrar que la matriz $\boldsymbol{M_2}$ realiza la transformación **within** de las variables. Entonces nos queda una regresión MCO de $\boldsymbol{M_2}\boldsymbol{y}$ en $\boldsymbol{M_2}\boldsymbol{X_1}$.

Efectos fijos (fixed-effects,FE)

- Podemos plantear $\hat{\mu}_i = \bar{y}_i - \hat{\beta}_{FE}\bar{x}_i$ como el estimador de los "efectos fijos".
- Si $T \to \infty$, $(\hat{\mu}_{FE}, \hat{\beta}_{FE})$ son estimadores consistentes e insesgados.
- Pero si T está fijo y $N \to \infty$, 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 presición.
- 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

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

En este caso hay **correlación serial**:
$Cov(u_{it}, u_{ij}) = Cov(\mu_i + \nu_{i,t}, \mu_i + \nu_{i,j}) = Var(\mu_i) \neq 0$. El estimador de RE tiene en cuanta 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:

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

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

Efectos aleatorios (random-effects,RE)

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

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

Ω tiene esta estructura:

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

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

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

Contraste de Hausman

El estimador de efectos fijos es siempre **consistente**. Sin embargo, el de efectos aleatorios es válido si las Xs no estan 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 $var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})$ es una matriz definida positiva.
- Se puede probar que $var(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})$.
- Entonces, definamos

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

tiene una distribución χ^2_K bajo la nula.

¿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 nececita 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, $t = -2, -1, 0, 1, ...$

¿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. ¿Cúal 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
```

Ejemplos en la web

- http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge13.html
- http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge14.html
- https://www.stata.com/manuals13/xtxtreg.pdf