Introducción al análisis de datos de panel en R

Estudio longitudinal social de Chile (ELSOC)

Enero, 2021: sesión 2

- Observaciones repetidas de "individuos" en el tiempo.
- Unidades: Personas, firmas, naciones, etc.
- 3 dimensiones:
  - Unidades i=1,. . . ,n.
  - Variables: v=1,. . . , V (constantes y variables en el tiempo).
  - Mediciones en el tiempo (olas): t=1,. . . , T.
- ¿Cómo poner ésto en matrices de datos de dos dimensiones?

Wide format

ID	Kids84	Kids85	Educ84	Educ85
1	0	0	12	12
2	2	2	9	9
3	0	1	10	11
4	1	2	8	8
5	3	3	13	13
6	2	2	15	15
7	0	1	9	10
...	...	...	...	...

- N individuos, T mediciones.
- Tamaño de la matriz de datos: n filas y v*t columnas.

Long format

ID	Jahr	Kids	Educ
1	1984	0	12
1	1985	0	12
...	...	...	...
2	1984	2	9
2	1985	2	9
...	...	...	...
3	1984	0	10
3	1985	1	11
...	...	...	...
4	1984	1	8
4	1985	2	8
...	...	...	...
5	1984	3	13
5	1985	3	13
...	...	...	...
6	1984	2	15
6	1985	2	15
...	...	...	...
7	1984	0	9
7	1985	1	10
...	...	...	...
7	2000	2	13

- N individuos, T mediciones.
- Tamaño de la matriz de datos: 1) N=n*T filas; 2) V columnas.

Ejemplo 1: Y es continua

Log hourly wage: original data						
Year	n	Mean		Sd	Serial correlation	
		log(y)	y		(t, t-1)	(t, t=1)
1980	545	1.393	4.03	0.558		
1981	545	1.513	4.54	0.531	0.454	0.454
1982	545	1.572	4.81	0.497	0.611	0.432
1983	545	1.619	5.05	0.481	0.690	0.408
1984	545	1.690	5.42	0.524	0.675	0.316
1985	545	1.739	5.69	0.523	0.664	0.356
1986	545	1.800	6.05	0.515	0.632	0.297
1987	545	1.866	6.47	0.467	0.693	0.310

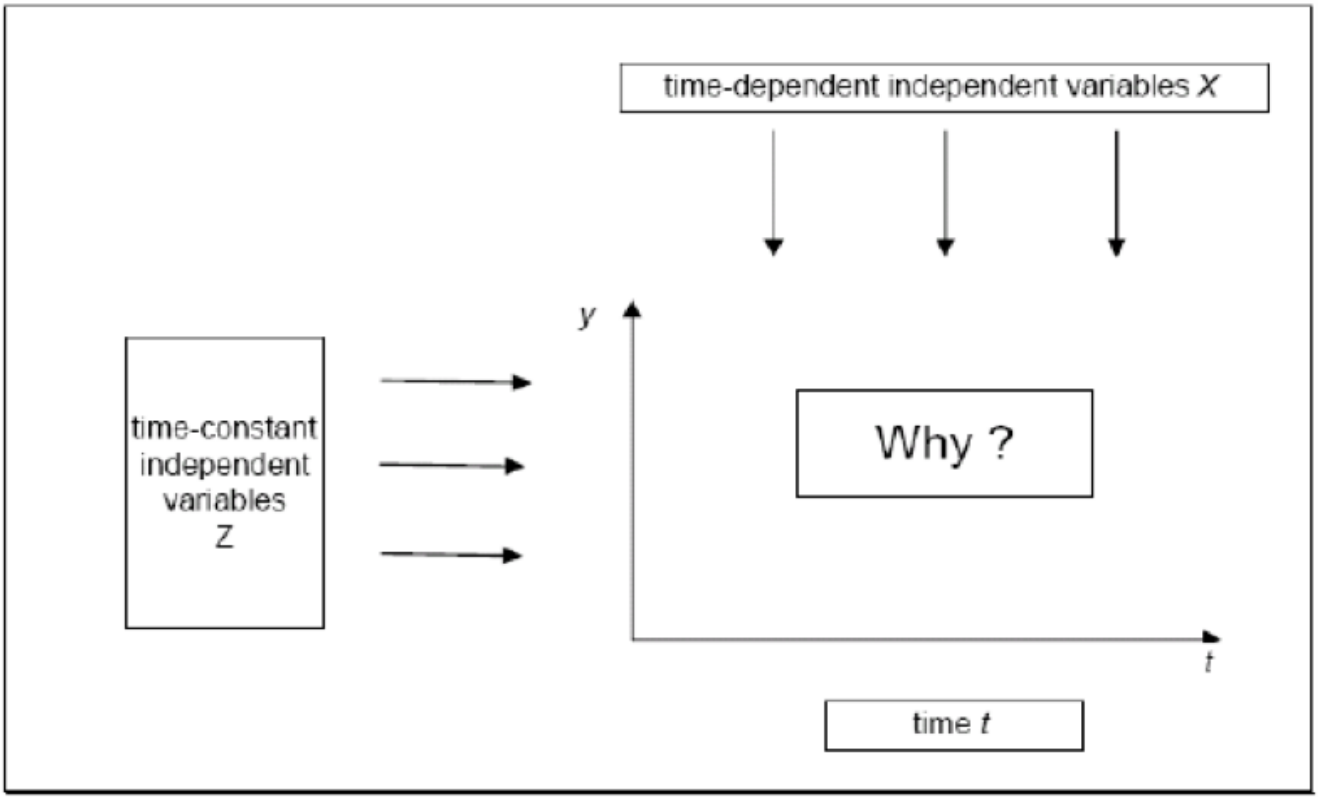
- n=545, T=8, y= log hourly wage.
- Alta correlación serial. Sin embargo, ésta disminuye con time-lags entre las mediciones.

Consecuencias

- Métodos estadísticos convencionales asumen observaciones independientes.
- En consecuencia
  - Errores estándares son subestimados.
  - Test estadísticos son muy altos.
  - Valores p son muy bajos.
  - Test de significación conducen a conclusiones erradas.

1. Manejo de datos de panel
   - Representación de datos de panel
   - Independencia de las observaciones
2. Modelamiento de datos de panel
   - Introducción
   - Variable dependiente continua: primeras diferencias (FD)
3. Efectos fijos (FE) para variables dependientes continuas
   - Formulación modelo FE
   - Supuestos FE
4. Resumen y literatura

¿Por qué hay diferentes trayectorias de ingresos?



- Explicación de dependencia serial
  - Variables constantes en el tiempo Z: No son un problema.
  - Variables X "time-dependent": X frecuentemente tiene valores similares en próximos años.
  - Y (variable dependiente) es influenciado por valores anteriores de Y.
- Términos técnicos: "state dependence", "modelos dinámicos".

- Y es una variable contínua para el individuo i en el tiempo t. Y puede ser modelada en términos de **niveles** o **cambios**.
- Si nos interesa los niveles, un modelo general para datos de panel se puede escribir del siguiente modo

$$E(y_{it}) = \beta_0 + \underbrace{\beta_1 x_{1it} + ... + \beta_k x_{kit}}_{\text{Depende del tiempo}} + \underbrace{\pi_1 z_{1i} + ... + \pi_j z_{ji}}_{\text{No varía en el tiempo}} + \mu_i + e_{it}. \quad (1)$$

- ¿Qué significa el término error?
- Los métodos de estimación para niveles y cambios son distintos.

$$E(y_{it}) = \beta_0 + \underbrace{\beta_1 x_{1it} + ... + \beta_k x_{kit}}_{\text{Depende del tiempo}} + \underbrace{\pi_1 z_{1i} + ... + \pi_j z_{ji}}_{\text{No varía en el tiempo}} + \mu_i + e_{it}. \quad (1)$$

- μ_i: característica desconocidas de las unidades i que son constantes en el tiempo. También conocido como "heterogeneidad no observada".

- e_{it}: error de medición y características de las unidades i no observadas que varían en el tiempo. También conocido como "error idiosincrático".

Supuestos del error: error = $\mu_i + e_{it}$

- Varianzas constantes: $(\sigma^2_\mu, \sigma^2_e)$.
- Independencia entre los errores: $\text{corr}(\mu_i, e_{it}) = 0$.
- No hay correlación serial: $\text{corr}(e_{it}, e_{is}) = 0$.
- Exogeneidad estricta: para cada punto en el tiempo t, el valor esperado del error idiosincrático condicionando por heterogeneidad no observada y por las covariantes X **en todos los puntos en el tiempo** es cero, en otras palabras, $E(e_{it}|X, \mu_i) = 0$.
- Exogeneidad estricta es especialmente importante para el estimador de efectos fijos.

model	$y_{it} = \beta_0 + \beta_1 x_{1it} + \gamma_1 z_{1i} + u_i + e_{it}$
$t = 3$	$y_{i3} = \beta_0 + \beta_1 x_{1i3} + \gamma_1 z_{1i} + u_i + e_{i3}$
$t = 2$	$y_{i2} = \beta_0 + \beta_1 x_{1i2} + \gamma_1 z_{1i} + u_i + e_{i2}$
$t = 1$	$y_{i1} = \beta_0 + \beta_1 x_{1i1} + \gamma_1 z_{1i} + u_i + e_{i1}$
$t_3 - t_2$	$y_{i3} - y_{i2} = \beta_1 (x_{1i3} - x_{1i2}) + (e_{i3} - e_{i2})$
$t_2 - t_1$	$y_{i2} - y_{i1} = \beta_1 (x_{1i2} - x_{1i1}) + (e_{i2} - e_{i1})$
in general for any T	$\Delta y_{it} = \beta_1 \Delta x_{1it} + \Delta e_{it}$

model	$y_{it} = \beta_0 + \beta_1 x_{1it} + \gamma_1 z_{1i} + u_i + e_{it}$
$t = 3$	$y_{i3} = \beta_0 + \beta_1 x_{1i3} + \gamma_1 z_{1i} + u_i + e_{i3}$
$t = 2$	$y_{i2} = \beta_0 + \beta_1 x_{1i2} + \gamma_1 z_{1i} + u_i + e_{i2}$
$t = 1$	$y_{i1} = \beta_0 + \beta_1 x_{1i1} + \gamma_1 z_{1i} + u_i + e_{i1}$
$t_3 - t_2$	$y_{i3} - y_{i2} = \beta_1 (x_{1i3} - x_{1i2}) + (e_{i3} - e_{i2})$
$t_2 - t_1$	$y_{i2} - y_{i1} = \beta_1 (x_{1i2} - x_{1i1}) + (e_{i2} - e_{i1})$
in general for any T	$\Delta y_{it} = \beta_1 \Delta x_{1it} + \Delta e_{it}$

- ¿Qué sucedió con μ_i?
- ¿Qué tipo de variables representa μ_i

- Primeras diferencias (FD) es un pooled OLS que usa datos *diferenciados*
  - No especificamos una constante.
  - Diferenciación elimina μ_i.
- Dado que t-1 es missing, FD trabaja con una muestra más pequeña
  - Hay que corregir los grados de libertad.
  - Stata o R hace la corrección de modo automático.

- Problemas de FD
  - No hay suficiente variación en las variables que cambian en el tiempo.
  - El supuesto de exogeneidad estricta no se cumple.
  - Puede incrementar el error de medición.
  - Reduce el tamaño de la muestra.

- FE es un pooled OLS que usa *time-demeaned data*
  - No espeficamos una constante.
  - El procedimiento de time-demeaned elimina μ_i.

- FE es un pooled OLS que usa *time-demeaned data*
  - No especificamos una constante.
  - El procedimiento de time-demeaned elimina μ_i.
- Número de observaciones por unidad no se modifica
  - Se pierden grados de libertad por razón de la transformación.
  - Stata o R lo hace automáticamente.

model	$y_{it} = \beta_0 + \beta_1 x_{1it} + \gamma_1 z_{1i} + u_i + e_{it}$
$t = 3$	$y_{i3} = \beta_0 + \beta_1 x_{1i3} + \gamma_1 z_{1i} + u_i + e_{i3}$
$t = 2$	$y_{i2} = \beta_0 + \beta_1 x_{1i2} + \gamma_1 z_{1i} + u_i + e_{i2}$
$t = 1$	$y_{i1} = \beta_0 + \beta_1 x_{1i1} + \gamma_1 z_{1i} + u_i + e_{i1}$
arithmetic mean	$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \gamma_1 z_{1i} + u_i + \bar{e}_i$

arithmetic mean	$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \gamma_1 z_{1i} + u_i + \bar{e}_i$
$t = 3$	$y_{i3} = \beta_0 + \beta_1 x_{1i3} + \gamma_1 z_{1i} + u_i + e_{i3}$
$t = 2$	$y_{i2} = \beta_0 + \beta_1 x_{1i2} + \gamma_1 z_{1i} + u_i + e_{i2}$
$t = 1$	$y_{i1} = \beta_0 + \beta_1 x_{1i1} + \gamma_1 z_{1i} + u_i + e_{i1}$
$t_3 - \bar{t}$	$y_{i3} - \bar{y}_i = \beta_1 (x_{1i3} - \bar{x}_{1i}) + (e_{i3} - \bar{e}_i)$
$t_2 - \bar{t}$	$y_{i2} - \bar{y}_i = \beta_1 (x_{1i2} - \bar{x}_{1i}) + (e_{i2} - \bar{e}_i)$
$t_1 - \bar{t}$	$y_{i1} - \bar{y}_i = \beta_1 (x_{1i1} - \bar{x}_{1i}) + (e_{i1} - \bar{e}_i)$
general	$\ddot{y}_{it} = \beta_1 \ddot{x}_{1it} + \ddot{e}_{it}$

- Variables constantes en el tiempo son eliminadas
  - Sin embargo, no eliminamos la interacción con variables que varían en el tiempo.
  - El procedimiento de time-demeaned elimina μ_i.
- Problemas
  - Ineficiente si no hay suficiente variación *within*.
  - Del mismo modo que FD, si no cumple el supuesto de exogeneidad estricta.
  - A diferencia de FD, puede sufrir de correlación serial.

- Para estimar FE, una alternativa al procedimiento de time-demeaned es el estimador least-squares-dummy-variables-estimator (LSDV): usar N-1 dummies por unidad como variables independientes.
- Efecto de variables que son constantes con el tiempo no puede ser estimado porque son linealmente dependientes de las dummies.
- Test de restricciones múltiples (F) de todas las dummies nos dice si hay heterogeneidad no observada.
- Procedimiento no es práctico cuando N es grande.

Supuestos para pooled OLS con FE: insesgamiento

1. Para cada unidad i el modelo es

$$y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + \mu_i + e_{it}, \ \ t = 1, ..., T. \tag{2}$$

2. Tenemos una muestra aleatoria para cada sección cruzada

3. Cada variable independiente cambia en el tiempo y no existe colinealidad perfecta entre ellas.

4. para cada t, el valor esperado del error idiosincrático, condicionando por las covariantes en todos los periodos de tiempo y por la heterogeneidad no observada es cero:

$$E(e_{it}|X_i, \mu_i) = 0. \tag{3}$$

del mismo modo que con FD, si (3) es válido, decimos que las variables x_{itj} son *estríctamente exógenas si condicionamos por el efecto no observado* μ_i.

Supuestos para pooled OLS con FE: eficiencia

5. La varianza de los errores diferenciados, condicionados por todas las variables independientes, es constante:

$$Var(e_{it}|X_i, \mu_i) = \sigma_e^2, \quad t = 1, ..., T. \tag{4}$$

6. Para todo $t \neq s$, las diferencias en los errores ideosincráticos no están correlacionadas (condicionando por todas las variables independientes)

$$Cov(e_{it}, e_{is}|X_i, \mu_i) = 0, \quad t \neq s. \tag{5}$$

7. Condicionando por X_i y μ_i, e_{it} están distribuidos de un modo normal, $N(0, \sigma_e^2)$.

- $E(e_{it}|X_i, \mu_i) = 0$ signica que:
  - x_{it} es estrícatmente exógena condicional sobre el efecto no observado: si condicionamos por μ_i, no hay correlación entre x_{is} y el error idiosincrático e_{it} **para todo s y t**.
  - Formulación alternativa 1: $E(e_{it}|\mu_i, x_{i1}, ..., x_{iT}) = 0$.
  - Esto implica $Corr(x_{is}, e_{it}) = 0, \quad s, t = 1, ..., T$.
  - Implicancia: valores futuros de los regresores no pueden estar correlacionados con el término error, por ejemplo, feedback entre y_{it} y x_{is}.

1. Manejo de datos de panel
   - Representación de datos de panel
   - Independencia de las observaciones
2. Modelamiento de datos de panel
   - Introducción
   - Variable dependiente continua: primeras diferencias (FD)
3. Efectos fijos (FE) para variables dependientes continuas
   - Formulación modelo FE
   - Supuestos FE
4. Resumen y literatura

- Modelamiento de datos de panel.
- FD y FE.

- Literatura efectos fijos
  - Angrist, J. D. y J. Pischke (2009). *Mostly Harmless Econometrics: an Empiricist's Companion*. Cap. 5.
  - Wooldridge, Jeffrey M. (2001). *Introducción a la econometría: un enfoque moderno*. Australia: Thomson. Caps. 13 y 14.
- Literatura datos de panel
  - Andress, H. J., K. Golsch y A. W. Schmidt (2013). *Applied Panel Data Analysis for Economic and Social Surveys*. Berlin, Heildelberg: Springer.