# TEMA 3: MODELOS DE DATOS FUSIONADOS Y MODELOS ESTÁTICOS PARA DATOS DE PANEL

M. Camacho

con la colaboración de A. Beyaert, M. González, A. Quesada
Departamento de Métodos Cuantitativos para la Economía y la Empresa

UNIVERSIDAD DE
MURCIA

# Lo que estudiaremos en este tema:

UNIVERSIDAD DE
MURCIA



*Bibliografía básica:*


Wooldridge, J. 2007. *Introducción a la econometría: un enfoque moderno*. Ed. Thomson,cap.13 y 14

# 1. Introducción

UNIVERSIDAD DE MURCIA

- Hasta ahora, análisis de regresión con:
  - Datos de corte transversal
  - Datos de serie temporal

- En este tema, análisis de regresión con datos que combinan información de corte transversal y de serie temporal:
  - Datos fusionados de secciones cruzadas independientes
  - Datos de panel o longitudinales estáticos

## Datos fusionados: estudio de efectos temporales

- Modelo de Fertilidad de las mujeres

$$y_i = \beta_0 + \delta_0 F74_i + \delta_1 F76_i + \delta_2 F78_i + \delta_3 F78_i + \delta_4 F80_i + \delta_5 F82_i + \delta_6 F84_i + \alpha' X_i + \varepsilon_i,$$
$$i = 1, \cdots, 1129 \qquad t=1972,1974,1976,1978,1980,1982,1984$$
$$N^* = 1129 \text{ observaciones}$$

y: número total de niños que ha tenido una mujer

F74,…, F84: variables ficticias temporales

X: vector de variables explicativas que incluye factores observados que afectan a la fertilidad (educación, edad, raza, región, entorno)

- Permite contestar a: una vez descontado el efecto de otros factores observables, ¿han cambiado los índices de fertilidad en el tiempo?

Nota: también se pueden incluir términos de interacción entre las ficticias anuales y alguna de las explicativas

UNIVERSIDAD DE MURCIA

## Datos de panel

- Modelo de crecimiento de Solow

$$\boxed{y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \eta_i + \varepsilon_{it}, \quad i = 1, \cdots, 86 \quad t = 60's, 70's, 80's}$$

$y_{it}$ : tasa de crecimiento de la renta del país i en la década t

$x_1$: log. renta inicial per cápita ($\beta_1$: tasa de convergencia)

$x_2$: tasa de crecimiento de la población

$x_3$: porcentaje del PIB destinado a inversión

$\eta_i$: características inobservables propias del país i, probablemente correlacionadas con su renta inicial, su inversión y su población

## Ventajas respecto de sección cruzada:

- **Tamaño muestral** más elevado (estimación más precisa)

- Permite **estudiar efectos temporales**: cambios en el tiempo del término constante o de los efectos de alguna de las explicativas
  - Disminución de la tasa de fertilidad en España
  - Mejora la nota de los universitarios en cursos más elevados

- Permite realizar **estudio de acontecimientos**: efectos causales de un determinado suceso o medida de política económica
  - Carnet por puntos => reducción de accidentes
  - Subida salario mínimo => aumento del desempleo
  - Programas de formación de trabajadores => mejora de la productividad
  - Apertura de discoteca => precio de vivienda (medición de efectos externos)
  - Estudios universitarios => mejoran el salario

- **Datos fusionados**: se obtienen realizando muestreo aleatorio de una población en diferentes momentos del tiempo
- En t se dispone de $N_t$ observaciones de una población, donde t=1,2,…,T y T≥ 2

$$y_i \qquad i=1,\ldots,N^*$$

el número total de observaciones es $N^* = N_1 + N_2 + \ldots + N_T$

Si T=2, el vector de observaciones es:

$$\left[\underbrace{y_1 \quad y_2 \quad \cdots \quad y_{N_1}}_{t=1} \quad \underbrace{y_{N_1+1} \quad y_{N_1+2} \quad \cdots \quad y_{N_1+N_2}}_{t=2}\right] \Rightarrow N^* = N_1 + N_2$$

- Provienen de observaciones muestrales independientes

$$y_i = \beta_0 + \delta_0 F2_i + \ldots \delta_{T-2} FT_i + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, \cdots, N \quad i \in \{t = 1, \ldots, t = T\}$$

- $F2_i, \ldots FT_i$ son (T-1) ficticias aditivas temporales (podrían haber multiplicativas)
- En notación matricial $Y = X\beta + \varepsilon$

$$\underbrace{\left[\begin{array}{c} y_1 \\ \vdots \\ y_{N_1} \\ \hline \vdots \\ \hline y_{N^*-N_T+1} \\ \vdots \\ y_{N^*} \end{array}\right]}_{Y} = \underbrace{\left[\begin{array}{ccccccc} 1 & F2_1 & \cdots & FT_1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & F2_{N_1} & \cdots & FT_{N_1} & x_{1N_1} & \cdots & x_{kN_1} \\ \hline \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \hline 1 & F2_{N^*-N_T+1} & \cdots & FT_{N^*-N_T+1} & x_{1(N^*-N_T+1)} & \cdots & x_{k(N^*-N_T+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & F2_{N^*} & \cdots & FT_{N^*} & x_{1N^*} & \cdots & x_{kN^*} \end{array}\right]}_{X} \cdot \underbrace{\left[\begin{array}{c} \beta_0 \\ \delta_0 \\ \vdots \\ \delta_{T-2} \\ \beta_1 \\ \vdots \\ \beta_k \end{array}\right]}_{\beta} + \underbrace{\left[\begin{array}{c} \varepsilon_1 \\ \vdots \\ \varepsilon_{N_1} \\ \hline \vdots \\ \hline \varepsilon_{N^*-N_T+1} \\ \vdots \\ \varepsilon_{N^*} \end{array}\right]}_{\varepsilon}$$

- En notación semimatricial

$$y_i = X_i'\beta + \varepsilon_i \quad \forall i=1,\ldots,N \;\text{ donde }\; X_i' = \left(1 \quad F2_i \quad \ldots \quad FT_1\, x_i \quad \cdots \quad x_i\right)$$

UNIVERSIDAD DE MURCIA

## El modelo de datos fusionados

$$y_i = X_i'\beta + \varepsilon_i \qquad i = 1,\ldots,N_1, N_1+1,\ldots,N_1+N_2,\ldots,N^*$$

- Si el modelo cumple los supuestos de Gauss-Markov:
  - MCO es ELIO
  - Son válidos los procedimientos de inferencia habituales
- Si hay heteroscedasticidad la tratamos como en sección cruzada

Normalmente el modelo

$$y_i = X_i'\beta + \varepsilon_i \quad \forall i=1,\dots,N \ \text{donde} \ X_i' = \begin{pmatrix} 1 & F2_i & \dots & FT_i\, x_i & \cdots & x_i \end{pmatrix}$$

incluye ficticias temporales para:

- Ficticias aditivas: La fertilidad empeora con el tiempo

- Ficticias multiplicativas: fertilidad empeora a través de una explicativa como puede ser la renta

- Estudios de acontecimiento: análisis de causalidad (ejemplo siguiente)

# 2. Datos fusionados
## 2.4 Utilidad de estos modelos
### Estudio de acontecimiento

UNIVERSIDAD DE MURCIA

## Efecto causal

1. Si toda las observaciones fueran iguales y reaccionaran igual
   - T=1 cuando ocurre el acontecimiento y T=0 cuando no ocurre

$$y = \beta_0 + \beta_1 T \Rightarrow EC = (y / T = 1) - (y / T = 0) = \beta_1$$

   - Problema: las observaciones nunca son iguales

2. Solución: muestreos en experimentos
   - Experimentos aleatorios: aleatoriamente se asigna
     - Grupo Tratamiento (T=1)
     - Grupo de Control (T=0)

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

$$\bar{y}^C = E(y / \text{control}) = \beta_0$$

$$\bar{y}^T = E(y / \text{tratamiento}) = \beta_0 + \beta_1$$

$$\bar{y}^T = \beta_0 + \beta_1$$

$$y^C = \beta_0$$

   - Estimador en diferencias: Efecto causal medio

$$ECM = E(y / T = 1) - E(y / T = 0) = \bar{y}^T - \bar{y}^C = \beta_1$$

   - Problema: Son costosos y no siempre viables

3. Solución: experimento natural o cuasi experimento

- Acontecimiento (incinerador) genera GT y GC ***como si*** hubiese sido un experimento aleatorio
- Problema: parte de $\bar{y}^T - \bar{y}^C = \beta_1$ ya estaba antes del acontecimiento (no es aleatorio)
- Solución: estimador de la **diferencia en las diferencias** con **datos fusionados**
- Ejemplo: Efecto causal de un incinerador sobre el precio de la vivienda

$$\boxed{\begin{array}{l} y_i = \beta_0 + \delta_0 F81_i + \beta_1 cerca_i + \delta_1 F81_i \cdot cerca_i + \alpha' X_i + \varepsilon_i, \\ i=1,\dots,321 \quad t = 1978, 1981 \quad N_{1978}=179 \quad N_{1981}=142 \end{array}}$$

y: precio de la vivienda en términos reales

F81: variable ficticia temporal que toma valor 1 si t=1981

T = cerca: ficticia que toma valor uno si la casa está cerca del incinerador

X: variables explicativas de las características observables de las viviendas (años, superficie, nº habitaciones….)

- $\delta_1$ es el coeficiente que interesa. Indica el efecto causal del incinerador en la caída del precio de la vivienda. Su valor estimado se obtendrá usando el **estimador de la diferencia en las diferencias**

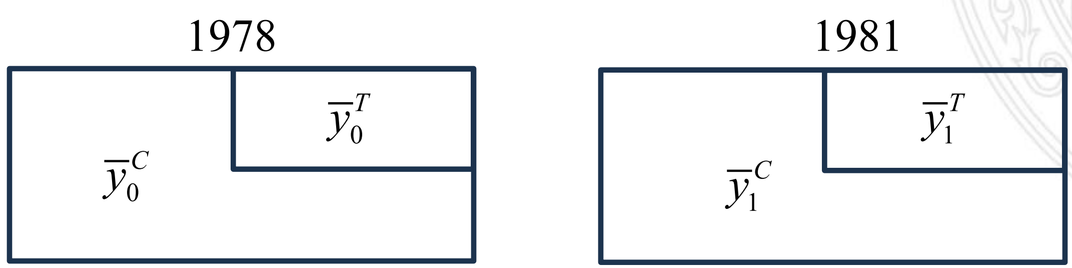
# Datos fusionados: estudio de acontecimientos

- $\delta_1$ y el **estimador de la diferencia en las diferencias**
  - Sin perder generalidad, supongamos $X_i=0$

$$y_i = \beta_0 + \delta_0 F81_i + \beta_1 cerca_i + \delta_1 F81_i \cdot cerca_i + \varepsilon_i$$

  - Llamemos

$$\bar{y}_0^C = E(y / control)_{78} = \beta_0$$
$$\bar{y}_0^T = E(y / tratamiento)_{78} = \beta_0 + \beta_1$$
$$\bar{y}_1^C = E(y / control)_{81} = \beta_0 + \delta_0$$
$$\bar{y}_1^T = E(y / tratamiento)_{81} = \beta_0 + \delta_0 + \beta_1 + \delta_1$$



- Tres formas de encontrar $\delta_1$

$$y_i = \beta_0 + \delta_0 F81_i + \beta_1 cerca_i + \delta_1 F81_i \cdot cerca_i + \varepsilon_i$$

| | control | tratamiento |
|---|---|---|
| 78 | $\bar{y}_0^C = \beta_0$ | $\bar{y}_0^T = \beta_0 + \beta_1$ |
| 81 | $\bar{y}_1^C = \beta_0 + \delta_0$ | $\bar{y}_1^T = \beta_0 + \delta_0 + \beta_1 + \delta_1$ |

$\Rightarrow \bar{y}_0^T - \bar{y}_0^C = \beta_1$

$\Rightarrow \bar{y}_1^T - \bar{y}_1^C = \beta_1 + \delta_1$

1 $$\left(\bar{y}_1^T - \bar{y}_1^C\right) - \left(\bar{y}_0^T - \bar{y}_0^C\right) = \delta_1$$

$$\underbrace{\bar{y}_1^C - \bar{y}_0^C = \delta_0 \qquad \bar{y}_1^T - \bar{y}_0^T = \delta_0 + \delta_1}$$

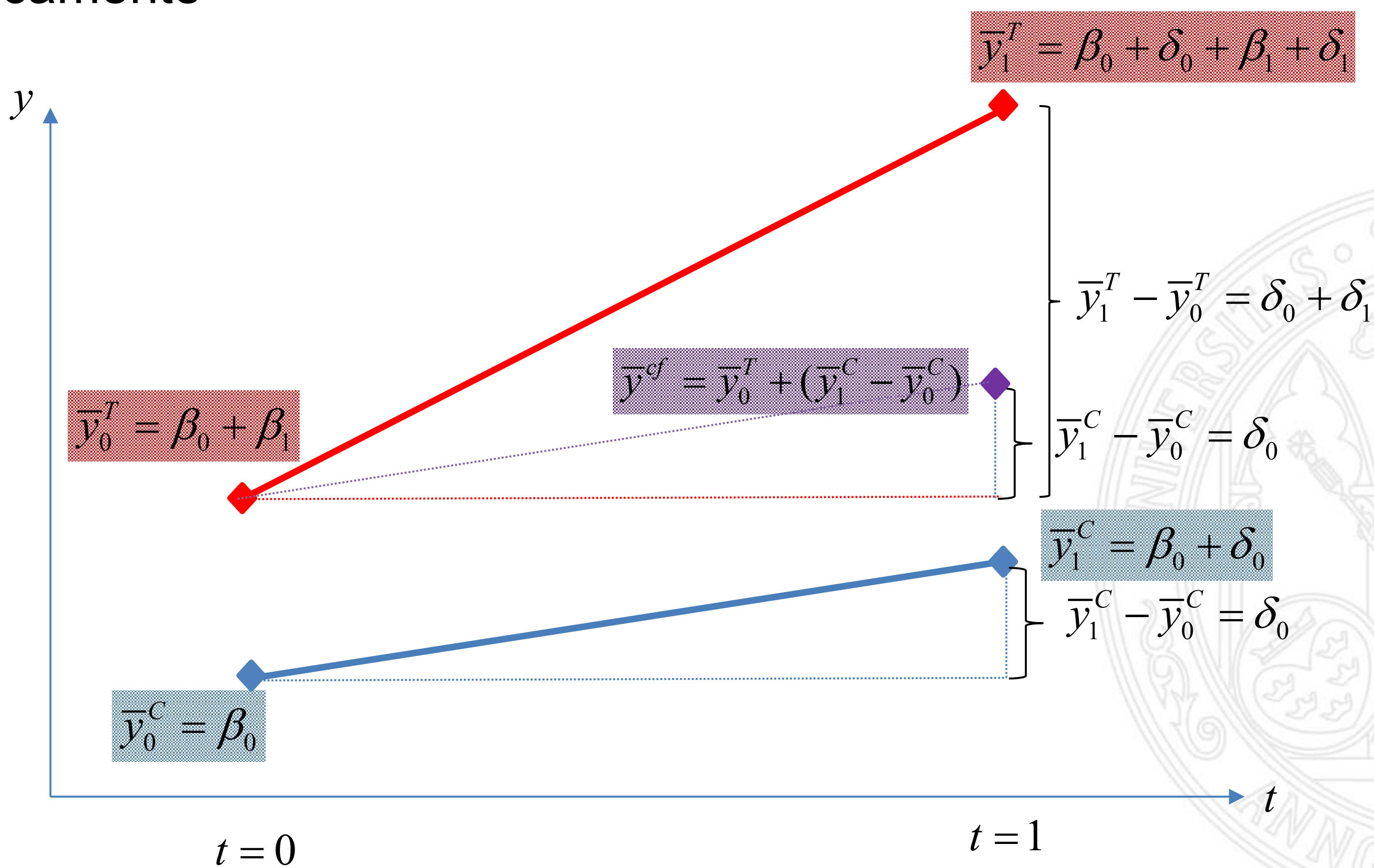
2 $$(\bar{y}_1^T - \bar{y}_0^T) - (\bar{y}_1^C - \bar{y}_0^C) = \delta_1$$

3 $$\bar{y}_1^T - \underbrace{[\bar{y}_0^T + (\bar{y}_1^C - \bar{y}_0^C)]}_{\bar{y}^{cf}} = (\beta_0 + \delta_0 + \beta_1 + \delta_1) - [\beta_0 + \beta_1 + (\beta_0 + \delta_0 - \beta_0)] = \delta_1$$

$\bar{y}^{cf}$ = contrafactual

UNIVERSIDAD DE MURCIA

- Gráficamente



$$y_i = \beta_0 + \delta_0 F81_i + \beta_1 cerca_i + \delta_1 F81_i \cdot cerca_i + \varepsilon_i$$

- El supuesto de tendencias paralelas
  - En los cálculos anteriores hemos supuesto
    - La diferencia entre T y C, medida por $\beta_1$, se mantiene en el tiempo $\Rightarrow$ aceptable
    - El incremento temporal, medido por $\delta_0$, afecta igual a T y C $\Rightarrow$ dudoso
  - Genera inconsistencia
    - Supongamos
      - Incremento temporal de C es $\delta_0$
      - Incremento temporal de T es $\delta_0 + \Delta$
    - En este caso el dif-dif queda

$$(\bar{y}_1^T - \bar{y}_0^T) - (\bar{y}_1^C - \bar{y}_0^C) = (\delta_0 + \Delta + \delta_1) - \delta_0 = \delta_1 + \Delta$$

  - Econométricamente ocurre cuando

$$\text{cov}(\varepsilon_i, F81_i cerca_i) \neq 0$$

UNIVERSIDAD DE MURCIA

- **Datos de panel** (o longitudinales): observaciones repetidas en el tiempo para una muestra de sección cruzada (individuos, familias, empresas, regiones, países, etc.)
- Son las **mismas unidades de sección cruzada** las que son observadas a través del tiempo

$$y_{it};\ i=1,\ldots,N \quad t=1,\ldots T \Rightarrow \text{el número total de observaciones es NT}$$

Si T=2, el vector de observaciones es:

$$\left[\underbrace{y_{11} \quad y_{21} \quad \cdots \quad y_{1N}}_{t=1} \quad \underbrace{y_{12} \quad y_{22} \quad \cdots \quad y_{2N}}_{t=2}\right] \Rightarrow NT = 2N$$

- Las observaciones no se distribuyen de forma independiente a lo largo del tiempo

- Resulta útil visualizar los datos de la siguiente manera:

Serie temporal individuo 2

Sección cruzada año 2

| t\i | 1 | 2 | 3 | … | N |
|---|---|---|---|---|---|
| 1 | $y_{11}$ | $y_{21}$ | $y_{31}$ | … | $y_{N1}$ |
| 2 | $y_{12}$ | $y_{22}$ | $y_{32}$ | … | $y_{N2}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ |
| T | $y_{1T}$ | $y_{2T}$ | $y_{3T}$ | … | $y_{NT}$ |

## Dimensiones del panel de datos

- N grande y T pequeño: *micropaneles* o *paneles cortos*

"Datos micro" observados en distintos momentos del tiempo (ej. datos de encuestas familiares o del balance de las empresas):

  - PSID (Panel Study of Income Dynamics)
  - ECHP (European Community Household Income and Wealth)
  - EPA (Encuesta de Población Activa)
  - ESEE (Encuesta sobre Estrategias Empresariales)
  - ECPF (Encuesta Continua de Presupuestos Familiares)

- N pequeño y T grande: *paneles largos*

"Datos macro" observados a lo largo del tiempo (ej. bases de datos internacionales):

  - Estadísticas de la OCDE, Eurostat, FMI, etc.
  - Penn World Table

- Los paneles de datos pueden ser:
  - Completos o equilibrados: $T_i = T \quad \forall i$

    No falta ninguna observación temporal para ninguna unidad de sección cruzada
  - Incompletos: $T_i \neq T \quad \text{para algún i}$

    Falta alguna observación temporal para alguna unidad

- En este tema consideramos:
  - Modelos lineales y sin endógena retardada
  - Paneles completos
  - “Datos micro”: N grande y T pequeño

    ⇨ Para las aproximaciones asintóticas suponemos T fijo y N→∞

UNIVERSIDAD DE MURCIA

- Hay una única ecuación para las N unidades de sección cruzada

$$y_{it} = \beta_0 + \delta_0 F2_t + \ldots \delta_{T-2} FT_t + \beta_1 x_{1,it} + \ldots + \beta_k x_{k,it} + \eta_i + \varepsilon_{it},$$
$$i = 1, \cdots, N \quad t = 1, \cdots, T$$

- $F2_t, \ldots FT_t$ : son (T-1) variables ficticias aditivas temporales (también podría haber multiplicativas)

- $\eta_i$ es el "efecto individual" o "efecto inobservable" o "efecto permanente": refleja diferencias inobservables constantes en el tiempo entre las unidades de sección cruzada

# 3. Datos de panel

## 3.2 El modelo de regresión

- Normalmente los datos se organizan mediante unidades de sección cruzada

$$\underbrace{\left[\begin{array}{c} y_{11} \\ \vdots \\ y_{1T} \\ \hline \vdots \\ \hline y_{N1} \\ \vdots \\ y_{NT} \end{array}\right]}_{Y} = \underbrace{\left[\begin{array}{cccccccc} 1 & F2_1 & \dots & FT_1 & x_{1,11} & x_{2,11} & \cdots & x_{k,11} \\ & & \ddots & & & \vdots & \ddots & \vdots \\ 1 & F2_T & \dots & FT_T & x_{1,1T} & x_{2,1T} & \cdots & x_{k,1T} \\ \hline & & \vdots & & & \vdots & \ddots & \vdots \\ \hline 1 & F2_1 & \dots & FT_1 & x_{1,N1} & x_{2,N1} & \cdots & x_{k,N1} \\ & & \vdots & & & \vdots & \ddots & \vdots \\ 1 & F2_T & \dots & FT_T & x_{1,NT} & x_{2,NT} & \cdots & x_{k,NT} \end{array}\right]}_{X} \cdot \underbrace{\left[\begin{array}{c} \beta_0 \\ \delta_0 \\ \vdots \\ \delta_{T-2} \\ \beta_1 \\ \vdots \\ \beta_k \end{array}\right]}_{\beta} + \underbrace{\left[\begin{array}{c} \eta_1 \\ \vdots \\ \eta_1 \\ \hline \vdots \\ \hline \eta_N \\ \vdots \\ \eta_N \end{array}\right]}_{\eta} + \underbrace{\left[\begin{array}{c} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1T} \\ \hline \vdots \\ \hline \varepsilon_{N1} \\ \vdots \\ \varepsilon_{NT} \end{array}\right]}_{\varepsilon}$$

$\Rightarrow$ notación matricial: $Y = X\beta + \eta + \varepsilon$

$\Rightarrow$ notación semimatricial: $y_{it} = X'_{it}\beta + \eta_i + \varepsilon_{it}$ $\quad i=1,\dots,N \quad t=1,\dots,T$

donde $X'_{it} = \left(1 \quad F2_t \quad \dots \quad FT_t \; x_{1,it} \quad \cdots \quad x_{k,it}\right)$

Características adicionales de datos de panel:

- Permite modelizar **respuestas dinámicas** en las poblaciones de sección cruzada (introduciendo variables retardadas).

- **Permite controlar las diferencias inobservables entre las unidades de sección cruzada.**

- Tenemos el siguiente modelo con **efectos individuales**

$$y_{it} = X'_{it}\beta + \eta_i + \varepsilon_{it} \qquad i=1,\ldots,N \quad t=1,\ldots,T$$

- **Objetivo**: estimar $\beta$ para cuantificar las relaciones causales entre las variables

- Problema: $\eta_i$ no se observa
  $\Rightarrow$ omisión de variables relevantes

UNIVERSIDAD DE MURCIA

$$y_{it} = X'_{it}\beta + \eta_i + \varepsilon_{it} \qquad i=1,\ldots,N \quad t=1,\ldots,T$$

- Si **Corr($X_{it}$ ,$\eta_i$)=0** la estimación MCO es consistente pero no óptima

  Con datos de panel la estimación más adecuada es **el estimador de efectos aleatorios** $\Rightarrow$ Modelo de efectos aleatorios

- Si **Corr($X_{it}$ ,$\eta_i$) ≠0** la estimación MCO no es consistente

  Con datos de panel la estimación más adecuada es **el estimador de efectos fijos** $\Rightarrow$ Modelo de efectos fijos

## Datos de panel: el control de la heterogeneidad inobservable constante

- Modelo de crecimiento de Solow

$$\boxed{y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \eta_i + \varepsilon_{it}, \quad i = 1, \cdots, 86 \quad t = 60's, 70's, 80's}$$

$y_{it}$ : tasa de crecimiento de la renta del país i en la década t

$x_1$: log. renta inicial per cápita ($\beta_1$: tasa de convergencia)

$x_2$: tasa de crecimiento de la población

$x_3$: porcentaje del PIB destinado a inversión

$\eta_i$: características inobservables propias del país i, probablemente correlacionadas con su renta inicial, su inversión y su población

$\Rightarrow$ La estimación MCO (y sobre $x_1$, $x_2$ y $x_3$) será sesgada e inconsistente

- Incluso cuando no hay datos sobre $\eta_i$, los datos de panel permiten la estimación consistente de $\beta_1$, $\beta_2$ y $\beta_3$

$$y_{it} = X'_{it}\beta + \underbrace{\eta_i + \varepsilon_{it}}_{u_{it}} \qquad i=1,\ldots,N \quad t=1,\ldots,T$$

- Supuestos, además de $\mathrm{Corr}(X_{it}, \eta_i) \neq 0$

  - S.1. Variables explicativas estrictamente exógenas: en todos los períodos, el término de error $\varepsilon_{it}$ está incorrelacionado con los valores presentes pasados y futuros de las explicativas

    $$E(\varepsilon_{it} | X) = 0, \quad \forall t$$

    ⇨ Excluye modelos con variables endógenas retardadas

  - S.2. Perturbaciones clásicas: Los errores son condicionalmente homocedásticos y no serialmente correlacionados

    $$V(\varepsilon_i | X) = \sigma^2 I_T \qquad \text{con } \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$$

El estimador de efectos fijos $y_{it} = X'_{it}\beta + \underbrace{\eta_i + \varepsilon_{it}}_{u_{it}}$

- Resulta de restar a cada observación su media individual
- Se obtiene:
  - 1º Calculando la ecuación para la media de cada grupo

$$\bar{y}_i = \bar{X}_i{}'\beta + \eta_i + \bar{\varepsilon}_i, \quad i = 1, \cdots, N \qquad \text{con} \quad \bar{\varepsilon}_i = \frac{\sum_{t=1}^{T} \varepsilon_{it}}{T}$$

  - 2º Restando a la ecuación original la ecuación de la media

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)'\beta + \varepsilon_{it} - \bar{\varepsilon}_i, \quad i = 1, \cdots, N \qquad t = 1, \cdots, T$$

- El modelo en desviaciones de la media:

$$\tilde{y}_{it} = \tilde{X}'_{it}\beta + \tilde{\varepsilon}_{it}, \quad i = 1, \cdots, N \qquad t = 1, \cdots, T$$

$$\tilde{y}_{it} = \left(y_{it} - \bar{y}_i\right) \qquad \tilde{X}_{it} = \left(X_{it} - \bar{X}_i\right) \qquad \tilde{\varepsilon}_{it} = \left(\varepsilon_{it} - \bar{\varepsilon}_i\right)$$

## El estimador del modelo en desviaciones de la media

- El estimador MCO del modelo en desviaciones de la media es el **estimador de efectos fijos** o **estimador intra-grupos**

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

- Se denomina también estimador intra-grupos porque sólo utiliza la variación temporal dentro de cada sección cruzada

- Bajo S.1 y S.2, $\hat{\beta}_{EF}$ es **insesgado** y **óptimo**. También es consistente cuando $N \rightarrow \infty$.

UNIVERSIDAD DE MURCIA

- Podemos obtener el efecto inobservable: $\hat{\eta}_i = \bar{y}_i - \bar{X}_i'\hat{\beta}_{EF}, \quad i = 1,\dots,N$ , pero esta estimación no es consistente cuando N→∞, para T fijo.

- Si alguno de los regresores es constante en el tiempo, no se podrá distinguir su efecto del de $\eta_i$

## Limitación del modelo de efectos fijos

- El modelo de efectos fijos deja parte de la regresión sin modelizar (la que sólo tiene variación de sección cruzada). Si esta parte explica mucha variabilidad de la variable dependiente, el modelo estimado tendrá muy poca capacidad predictiva

- La posibilidad de controlar la heterogeneidad inobservable solo es una ventaja cuando el objetivo del análisis econométrico es la medición de relaciones estructurales, no la predicción.

- El coste del modelo de efectos fijos es que excluimos los efectos de cualquier variable observada constante en el tiempo.

UNIVERSIDAD DE MURCIA

$$y_{it} = X_{it}'\beta + \eta_i + \varepsilon_{it}, \qquad i = 1, \cdots, N \quad t = 1, \cdots, T$$

- En el modelo de efectos aleatorios $\eta_i$ es una variable aleatoria inobservable **no correlacionada** con las variables explicativas de modelo, Corr($X_{it}$, $\eta_i$)=0, t= 1,2,…,T :

$$y_{it} = X_{it}'\beta + \underbrace{\eta_i + \varepsilon_{it}}_{u_{it}}, \qquad i = 1, \cdots, N \quad t = 1, \cdots, T$$

$$y_{it} = X_{it}'\beta + u_{it}, \quad i = 1, \cdots, N \quad t = 1, \cdots, T$$

- Ahora el error del modelo es un término de error compuesto **no correlacionado** con las variables explicativas del modelo

$$y_{it} = X'_{it}\beta + u_{it}, \quad i = 1, \cdots, N \quad t = 1, \cdots, T$$

$$u_{it} = \eta_i + \varepsilon_{it}, \qquad \mathrm{Corr}(X_{it}, \eta_i) = 0$$

- Bajo los supuestos:

$$\begin{cases} \text{S.1}'. \ \ \eta_i \left| X_i \sim \text{iid}(0, \sigma_\eta^2) \right. \\ \text{S.2}'. \ \ \varepsilon_{it} \left| X_i \sim \text{iid}(0, \sigma^2) \right. \\ \text{S.3}'. \ \ \eta_i \ \text{y} \ \varepsilon_{it} \ \text{son independientes} \end{cases}$$

- El modelo de efectos aleatorios presenta:
  - Exogeneidad estricta: $E\left(u_{it} \middle| X\right) = 0, \quad \forall t$
  - **Autocorrelación**: $E\left(u_{it}u_{is}\right) = \begin{cases} \sigma^2 + \sigma_\eta^2, & t = s \\ \sigma_\eta^2, & t \neq s \end{cases}$

$\Rightarrow$ El estimador MCO es insesgado y consistente, pero no óptimo

UNIVERSIDAD DE MURCIA

La autocorrelación en el modelo de efectos aleatorios

- La matriz de covarianzas de los T valores correspondientes al individuo i-ésimo es:

$$V(u_i)_{T\times T} = E\left(u_i u_i'\right) = V\begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix} = \begin{bmatrix} \left(\sigma_\eta^2+\sigma^2\right) & \sigma_\eta^2 & \dots & \sigma_\eta^2 \\ \sigma_\eta^2 & \left(\sigma_\eta^2+\sigma^2\right) & \dots & \vdots \\ \vdots & \vdots & \ddots & \sigma_\eta^2 \\ \sigma_\eta^2 & \dots & \sigma_\eta^2 & \left(\sigma_\eta^2+\sigma^2\right) \end{bmatrix} = v$$

⇨ La correlación entre dos observaciones provenientes de un mismo individuo es constante y no desaparece en el tiempo

- La matriz de covarianzas para toda la muestra es diagonal por bloques:

$$V(u)_{TN\times TN}=E\left(uu'\right)=\begin{bmatrix} v_{T\times T} & 0 & \cdots & 0 \\ 0 & v_{T\times T} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{T\times T} \end{bmatrix}=V$$

- La matriz de covarianzas no es una matriz diagonal

$\Longrightarrow$ para buscar optimalidad hay que estimar por MCG

UNIVERSIDAD DE MURCIA

- El estimador óptimo del modelo de efectos aleatorios sería el estimador MCG:

$$\hat{\beta}_{MCG} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}Y$$

pero V depende de $\sigma^2$ y $\sigma_\eta^2$ que son desconocidos

⇨ hay que hallar el estimador **MCG factible**

- MCG factible es el **estimador de efectos aleatorios**

$$\hat{\beta}_{EA} = \left(X'\hat{V}^{-1}X\right)^{-1} X'\hat{V}^{-1}Y$$

UNIVERSIDAD DE MURCIA

- Para calcular $\hat{V}$ necesitamos estimaciones consistentes de $\sigma^2$ y $\sigma_\eta^2$ :

$$\hat{\sigma}^2 = \frac{1}{N(T-1)-k}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \tilde{X}_{it}'\hat{\beta}_{EF}\right)^2$$

$$\hat{\sigma}_\eta^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{y}_i - \bar{X}_i\hat{\beta}_{EG}\right)^2 - \frac{\hat{\sigma}^2}{T}$$

donde $\hat{\beta}_{EG}$ es el “estimador entre-grupos” resultante de la estimación MCO de:

$$\bar{y}_i = \bar{X}_i{}'\beta + \eta_i + \bar{\varepsilon}_i, \quad i = 1, \cdots, N$$

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

## Conclusión

- Bajo el supuesto de efectos individuales incorrelacionados con las explicativas, el **estimador de efectos aleatorios** es **consistente** y **asintóticamente óptimo**.

- Ojo!! sería inconsistente si los efectos individuales están correlacionados con las explicativas.

Recordemos:

- $\hat{\beta}_{EA}$ es:
  - consistente y asintóticamente óptimo si $\mathrm{Corr}(X_{it}, \eta_i) = 0$
  - inconsistente si $\mathrm{Corr}(X_{it}, \eta_i) \neq 0$
- $\hat{\beta}_{EF}$ es siempre consistente

$\Rightarrow$ Si $\mathrm{Corr}(X_{it}, \eta_i) = 0$ ambas estimaciones serían muy similares

- Necesitamos elegir entre $\hat{\beta}_{EA}$ y $\hat{\beta}_{EF}$

- El contraste de Hausman (1978) contrasta la correlación entre los efectos individuales y las explicativas comparando $\hat{\beta}_{EF}$ y $\hat{\beta}_{EA}$

$$H_0 : \eta_i \text{ y } X_i \text{ no correlacionados } (\hat{\beta}_{EF} \approx \hat{\beta}_{EA})$$
$$H_A : \eta_i \text{ y } X_i \text{ correlacionados}$$

- El estadístico para el contraste es:

$$H = \left(\hat{\beta}_{EA} - \hat{\beta}_{EF}\right)' \left(\hat{V}(\hat{\beta}_{EF}) - \hat{V}(\hat{\beta}_{EA})\right)^{-1} \left(\hat{\beta}_{EA} - \hat{\beta}_{EF}\right) \underset{H_0}{\overset{a}{\sim}} \chi^2_k$$

donde k es el número de parámetros del modelo

# Lo que hemos aprendido:



- Qué son datos fusionados y qué son datos de panel
- Estimación con datos fusionados: bajo los supuestos clásicos MCO es ELIO
- Estimación con datos de panel:
  - Si el efecto individual está correlacionado con las explicativas:
    - MCO sesgado e inconsistente
    - El estimador de efectos fijos es insesgado y óptimo
  - Si el efecto individual no está correlacionado con las explicativas:
    - MCO insesgado pero no óptimo
    - El estimador de efectos aleatorios es insesgado y asintóticamente óptimo
  - Contraste de Hausman para elegir entre el estimador de efectos fijos y el estimador de efectos aleatorios