INTRODUCCIÓN

Una sección cruzada puede expresarse matricialmente, conteniendo en columnas información sobre características de individuos, familias, empresas, regiones, etcétera, que quedan recogidos en las filas. Todos estos datos están referidos a un mismo momento del tiempo, o a una misma muestra. Cuando disponemos de dos o más de estas matrices, tomadas de forma aleatoria e *independiente* en distintos momentos del tiempo podemos tener errores incorrelados, pero las observaciones no estarán idénticamente distribuidas, por lo que hay que considerar que existen diferencias entre uno y otro conjunto de datos al estimar (con cambios en la constante o en los coeficientes de las variables).

Un panel de datos, como dos o más secciones cruzadas fusionadas, puede tener una dimensión temporal, pero en el caso del panel las unidades objeto de la muestra son las mismas, es decir, los atributos, para distintos momentos del tiempo, se refieren a las mismas personas, regiones, familias. Como es obvio, las observaciones ya no serán independientes en el tiempo, aunque podamos asumir que están idénticamente distribuidas. Para tener esto en cuenta se introducen variables ficticias (*dummies*) temporales, y a veces interacciones en los coeficientes con esas mismas *variables ficticias*.

***

¿Qué ventajas tienen los datos de sección cruzada fusionados? La más evidente es que multiplican el número de observaciones a nuestra disposición. Dado que el supuesto de idéntica distribución ya no se cumple. Veamos cómo se introducen las variables ficticias (*dummies*). Partimos de una especificación normal:

$$\ln y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

Ahora introducimos *las ficticias* temporales, incluidas interacciones para alguno de los coediciones:

$$\ln y = \alpha_0 + \alpha_1 d_1 + \beta_1 x_1 + \delta_1 d_1 x_1 + \beta_2 x_2 + u$$

Estamos suponiendo que disponemos de dos secciones cruzadas independientes, cada una de un período de tiempo distinto. La *dummy* d_1 toma valor 1 para las observaciones del segundo período, y 0 para el primero. El término constante para el primer conjunto de datos es α_0 y para el segundo $\alpha_0 + \alpha_1$. Hemos supuesto que el

coeficiente β_1 cambia con el tiempo, por lo que hemos introducido una interacción con la variable temporal *ficticia*. El coeficiente para el primer período temporal es β_1, y para el segundo período es $\beta_1 + \delta_1$, y δ_1 mide cómo el coeficiente β_1 *cambia* entre los dos períodos. Si interaccionamos todas las variables independientes con la *ficticia* temporal estamos haciendo lo mismo que estimar por separado el modelo original con los dos conjuntos de datos. En cualquier caso existe un test para el cambio estructural en el tiempo, conocido como *test de Chow*, que es un test F (por lo que requiere homocedasticidad, aunque hay una versión válida para casos de heterocedasticidad). Básicamente comprueba si una regresión múltiple difiere entre dos grupos de datos (dos submuestras normalmente). No obstante, existe otra forma de comprobar si hay cambios estructurales entre ambos períodos, más intuitiva. La idea es interaccionar todas las variables con una *ficticia* temporal para uno de los años y comprobar la significatividad conjunta de la *ficticia* temporal y de todas las interacciones. Otra posibilidad es, como hemos visto, introducir sólo algunas interacciones y comprobar que éstas son significativas una a una, mediante el estadístico *t*. Cuando el número de períodos es mayor que dos el número de interacciones se multiplica, pero el procedimiento que acabamos de ver es perfectamente generalizable.

Decíamos que una de las ventajas de las secciones cruzadas fusionadas es que nos permiten trabajar con más observaciones, aunque las interacciones introduzcan también más parámetros a estimar. Otra ventaja es que nos permiten estudiar los efectos de las políticas económicas. Mediante el procedimiento que hemos esbozado podemos saber si ha habido cambios significativos en los coeficientes después de introducida la medida de política económica, si esto tuvo lugar entre los dos períodos para los que se dispone de datos. Esto se conoce como experimento natural, o cuasi-experimento, que requiere un *grupo de control* y un *grupo de tratamiento*. Estos grupos no se definen a priori, antes del experimento, sino que quedan definidos en función de la política cuyos efectos se quiere estudiar (por eso se utiliza la partícula "quasi"). En nuestro caso, una muestra de la población que contiene al grupo de control y al de tratamiento antes y otra con los dos grupos después de la medida.

Si denotamos A al grupo de control y B al de tratamiento, creamos una variable *ficticia* dB que tiene valor 1 para los miembros del grupo de tratamiento y cero para los del grupo de control; además, d2 es una *ficticia* temporal que marca las observaciones del segundo período de tiempo, momento en que la medida de política económica ya ha tenido lugar. El modelo debería especificarse como

$$y = \alpha_0 + \delta_0 d2 + \alpha_1 dB + \delta_1 d2\ dB + \beta_1 x_1 + \beta_2 x_2 + \ldots + u$$

El efecto de la medida queda determinado por el coeficiente δ_1. Por otro lado, α_0 es la constante, referida al grupo de control en el primer período; δ_0 la constante para el grupo de control el segundo período; α_1 la constante para el grupo de tratamiento el primer período; y δ_1 el efecto medio de la política para el grupo de

tratamiento en el segundo período si no hay más variables en el modelo. Cuando hay más variables en el modelo, para tener en cuenta otros factores además de la medida de política económica en sí misma, la interpretación del coeficiente δ_1 es la misma, pero ya no se puede expresar como un estimador de diferencias-en-diferencias, es decir:

$$\hat{\delta}_1 = (\overline{y}_{2,B} - \overline{y}_{2,A}) - (\overline{y}_{1,B} - \overline{y}_{1,A})$$

A veces no hablamos de períodos de tiempo, sino de dos regiones, o dos hospitales, etc.

***

En este punto vamos a introducir un breve paréntesis. Vamos a tratar el problema de la omisión de variables. Cuando una variable explicativa queda excluida del modelo, por no disponerse de información sobre ella, la variable pasa a sumarse al error, lo que sesga los estimadores mínimo cuadráticos de los coeficientes. En efecto, si partimos de un modelo como este:

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

y no disponemos de información para la variable x_3, ésta pasará al error, y el modelo quedará reducido a

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

donde $u = \beta_3 x_3 + e$. Una forma de solución al problema que presenta la omisión de una variable relevante es encontrarle una variable *proxy*, a la que denotaremos con x_3^*. Suponemos que entre ambas existe una correlación significativa, de forma que

$$x_3 = \delta_0 + \delta_3 x_3^* + v$$

siendo δ_3 significativa como condición para que x_3^* sea una buena *proxy* de x_3. Simplemente, usamos x_3^* en lugar de x_3, pretendiendo con esto que ambas

variables son muy similares. Para que esta sustitución permita obtener estimadores de β_1 y β_2 consistentes es necesario que se cumplan dos condiciones. Primero, que el error u *no* esté correlacionado con x_1, x_2 y x_3, y también, separadamente, incorrelado con x_3^*, es decir, el valor esperado de u dadas esas cuatro variables debe ser cero, y también dadas las 3 primeras. La lógica es que, dado x_3, la variable x_3^* no añade nada, dado que ambas son similares y una reemplaza a la otra. La segunda condición es que el error v debe estar también incorrelado con x_1, x_2 y x_3, siendo la idea que una vez tenemos en cuenta a x3*, el valor esperado de x_3 no debe depender para nada de x_1 y x_2, o dicho de otra forma, la correlación entre x_3 y x_1 y x_2 es cero una vez tenemos en cuenta x_3^*, esto es, en la expresión que liga a x_3 y x_3^* no participan ni x_1 ni x_2.

Imaginemos por un momento qué pasaría si esos supuestos no se cumplen. Si tuviéramos

$$x_3 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3^* + v$$

siendo δ_1 y δ_2 positivos y significativos, y sustituimos ahora esa expresión en el modelo original, tendremos

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3^* + v) + e$$

y reordenando,

$$y = (\alpha_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3^* + u$$

Si hacemos $u = \beta_3 v + e$, y las propiedades de *e* y *v* garantizan que *u* está incorrelada con las variables explicativas. Como vemos, siendo δ_1 y δ_2 positivos y significativos, el coeficiente estimado no coincidirá con el real, y sólo cuando esos dos parámetros sean cero evitaremos el sesgo.

Si no conocemos nada de la variable omitida no será difícil escoger una buena *proxy* para ella. Lo que suele hacerse en estos casos es emplear la propia variable dependiente, retrasada un período, como variable explicativa, para introducir así en el modelo información sobre la variable omitida. La lógica que subyace a este planteamiento es que la variable omitida, de existir, estará influyendo sobre esa variable dependiente retrasada, y que al incluirla en el modelo estaremos incorporando, como variable explicativa, la variable omitida. Por ejemplo, los

índices de criminalidad de una ciudad pueden depender de variables no observables o desconocidas, pero estos factores pueden estar explicando en parte también los valores pasados de criminalidad, con lo que al incluir estos valores pasados estamos considerando como variables explicativas las variables inobservadas. El problema es que, muy probablemente, haya algún tipo de correlación entre esa variable dependiente retardada y el resto de variables explicativas, pero permite considerar más información en la estimación de nuestro modelo. Los paneles de datos permiten, precisamente, incorporar información de más de un período de tiempo para tratar el problema de la influencia de variables no observadas, como veremos.

Antes de eso vamos a ver otra posibilidad de tratar el problema de la omisión de variables con datos de sección cruzada. Como hemos visto, al omitir una variable explicativa ésta pasa a sumarse al error, y si había correlaciones entre las variables explicativas del modelo, al pasar al error el nuevo término romperá el supuesto de exogeneidad estricta. Para poder estimar un modelo en el que se han omitido variables explicativas significativas necesitamos alguna forma de información adicional. El procedimiento del que hablamos consiste en insertar en el modelo información proveniente de una *variable instrumental*, que debe tener *dos propiedades*: no debe estar correlacionada con el error (debe ser exógena); pero sí debe estar correlacionada con alguna de las variables explicativas observables. Veámoslo más despacio.

Como en el caso anterior, partimos del modelo:

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

y encontramos que no disponemos de información para la variable x_3, por lo que ésta pasará al error, y el modelo quedará reducido a

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Ahora surge el problema: es posible que x_3 y alguna de las variables explicativas observables tengan correlación entre ellas, por lo que *u* y dicha o dichas variables explicativas estarán correlacionadas también. Eso inutiliza el procedimiento de mínimos cuadrados, como sabemos. Ahora cambiaremos la notación del modelo anterior para distinguir las variables exógenas (que *no* tienen correlación con *u*) de las endógenas (que tienen correlación con *u*).

$$y_1 = \alpha_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Tenemos que y_2 es endógena, porque sospechamos tiene una correlación con u_1 dado que hay una variable en el error que está correlacionada positiva o negativamente con y_2. Mínimos cuadrados, en estas condiciones, nos daría estimadores sesgados e inconsistentes. Por tanto, la estrategia es buscar una *nueva* variable que sirva de instrumento para y_2. Esta variable instrumental debe ser exógena, como z_1, por lo que la llamaremos z_2. Además, el valor esperado de u_1 debe ser cero. Por tanto tendremos:

$$E(u_1) = 0 \qquad Cov(z_1,u_1) = 0 \qquad Cov(z_2,u_1) = 0$$

Si y_2 fuera también exógena e igual a z_2 tendríamos que los *estimadores de variables instrumentales* (que se derivan de las condiciones antedichas) equivaldrían exactamente a los mínimo cuadráticos.

La otra condición esencial es que z_2 (la variable instrumental) e y_2 estén correlacionadas, parcialmente correlacionadas, ya que hay una tercera variable interactuando con ellas (z_1). ¿Cómo expresamos esto? Digamos que podemos escribir y_2 en función de las variables exógenas z_1 y z_2, de forma que

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$$

donde, por definición, se cumplirá que

$$E(v_2) = 0 \qquad Cov(z_1,v_2) = 0 \qquad Cov(z_2,v_2) = 0$$

La clave aquí está en que $\pi_2 \neq 0$, es decir, ese parámetro no puede ser cero ya que exigimos que, aun considerando la influencia de z_1, las variables z_2 e y_2 estén correlacionadas. Una simple estimación mínimo cuadrática, robusta a la heterocedasticidad, de esa ecuación nos permite, a través de un estadístico t, hacer un test de ese supuesto.

Mientras que la ecuación que tiene variables endógenas entre las explicativas se conoce como *ecuación estructural*, la que pone una endógena en función de exógenas (la segunda) se conoce como *ecuación en forma reducida*. El *modelo de variables instrumentales* se puede generalizar fácilmente. Tenemos k variables explicativas en el modelo estructural:

$$y_1 = \alpha_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1$$

donde suponemos que y_2 está correlacionado con u_1. Buscamos una nueva variable z_k para incluir en el modelo, siendo también necesariamente exógena. Debe cumplirse entonces que

$$E(u_1) = 0 \qquad Cov(z_j, u_1) = 0 \qquad j = 1,..., k$$

La forma reducida de la ecuación sería

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2$$

Y aquí viene la segunda condición clave para el estimador de variables instrumentales:

$$\pi_k \neq 0$$

Bajo las dos condiciones (exogeneidad de z_k y existencia de una correlación parcial entre z_k y y_2), la variable z_k es un *instrumento* válido para y_2.

***

Con los datos de panel se da un cambio cualitativo muy importante, pues estaremos observando a *los mismos* individuos, familias, empresas, durante más de un período. El caso más sencillo, como en los datos de sección cruzada fusionados, cuenta con dos períodos temporales (T=2) y muchos individuos (N grande).

Con esta estructura de datos se consigue una gran ventaja para el tratamiento de *cierto tipo* de variables no observables, que no teníamos con las secciones cruzadas, y es que, al tratarse de los mismos individuos en uno y otro período, habrá variables explicativas que varíen en el tiempo, y otras que no. Estas variables que no varían en el tiempo son un problema si no son observadas, pero gracias al panel de datos es fácil eliminarlas o contrarrestarlas.

Una expresión sencilla del caso, donde *t* representa el tiempo e *i* a unidad muestral, es decir, la persona, familia, empresa, etcétera, sería

$$y_{it} = \alpha_0 + \delta_2 d2_t + \beta_1 x_{1it} + \beta_2 x_{2it} + ... + \beta_k x_{kit} + a_i + u_{it} \qquad t = 1,2$$

La variable a_i recoger los factores no observados que no cambian en el tiempo (por eso no tiene subíndice t) y que están afectando a y_{it}. Por ello se la conoce como *efecto no observado* o *efecto fijo*, términos que dan también nombre al modelo, o *heterogeneidad no observada*. Los factores no observados que sí cambian en el tiempo están recogidos en u_{it}, conocido como *error idiosincrático*, muy similar al que aparece en las series temporales.

¿Cómo podemos estimar la ecuación anterior? Si fusionamos los datos y aplicamos mínimos cuadrados podemos encontrarnos con problemas relacionados con la heterogeneidad no observada. Dado que no hay datos para a_i esa variable deberá sumarse necesariamente al error idiosincrático, que pasará a llamarse *error compuesto*. Si a_i está correlacionada con alguna de las variables explicativas observadas el error compuesto estará también correlacionado con ellas, aunque el error idiosincrático no lo esté, con lo que no se cumpliría uno de los supuestos esenciales de los mínimos cuadrados. Los coeficientes estimados estarían sesgados y serían inconsistentes, lo que se conoce como *sesgo de heterogeneidad*. Veamos cómo quedaría el modelo:

$$y_{it} = \alpha_0 + \delta_2 d2_t + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + v_{it} \qquad t = 1, 2$$

donde $v_{it} = a_i + u_{it}$ es el error compuesto. El problema que se nos presenta es cómo tratar el problema cuando no podemos asumir que a_i no esté correlacionada con las variables explicativas. Los paneles de datos aportan aquí una ventaja que una mera fusión de secciones cruzadas no puede igualar: al ofrecer observaciones repetidas *para las mismas unidades muestrales* podemos cancelar la presencia de la heterogeneidad no observada. Eso se hace, simplemente, restando las observaciones de dos años consecutivos.

Para el segundo período y la unidad muestral *i* tendremos que d2=1, y por tanto

$$y_{i2} = (\alpha_0 + \delta_2) + \beta_1 x_{1i2} + \beta_2 x_{2i2} + \ldots + a_i + u_{i2} \qquad t = 2$$

$$y_{i1} = \alpha_0 + \beta_1 x_{1i1} + \beta_2 x_{2i1} + \ldots + a_i + u_{i1} \qquad t = 1$$

Ahora sustraemos la primera ecuación de la segunda, y nos queda

$$(y_{i2} - y_{i1}) = \delta_2 + \beta_1 (x_{1i2} - x_{1i1}) + \beta_2 (x_{2i2} - x_{2i1}) - (u_{i2} - u_{i1})$$

Y en términos de incrementos, que denotamos con Δ, tendremos

$$\Delta y_i = \delta_2 + \beta_1 \Delta x_{1i} + \beta_2 \Delta x_{2i} + \ldots + \Delta u_i$$

Esta ecuación recibe el nombre de *ecuación en primeras diferencias*. Obsérvese que el término a_i ha sido eliminado en la sustracción, pues aparecía en ambas ecuaciones. El término constante δ_2 significa ahora otra cosa: el *cambio* en el término constante que se ha producido entre ambos períodos.

A nuestra ecuación en primeras diferencias se le puede aplicar mínimos cuadrados, siempre que se cumplan los supuestos, exactamente igual que con una ecuación basada en datos de sección cruzada. Pero veamos algunas particularidades que provienen de aplicar mínimos cuadrados a esta ecuación de "incrementos".

Un supuesto clave de los mínimos cuadrados es que el error no puede estar correlacionado con las variables explicativas (supuesto de exogeneidad estricta), y en nuestro caso, que Δu_i no puede estar correlacionado con las Δx_{ji}. ¿Cuándo se cumple esto? Cuando el error idiosincrático no está correlacionado con las variables explicativas en cada período de tiempo por separado, es decir, cada u_{it} no puede estar correlacionado con las x_{jit} de su mismo período. No obstante, gracias al panel de datos, éstas sí pueden estar correlacionadas con factores no observables constantes en el tiempo, como hemos visto. El supuesto impide que una de las variables explicativas sea la variable dependiente de un período anterior, ya que esto introduciría automáticamente correlaciones indeseadas. La exogeneidad estricta tampoco se da si hemos omitido una variable que cambia en el tiempo. El estimador de los coeficientes de las variables explicativas se conoce como *estimador de primeras diferencias*.

Como vemos, los paneles de datos son muy interesantes para estimar modelos en los que hay variables omitidas o inobservadas que no cambian con el tiempo. Pero cuando aplicamos la sustracción que nos lleva a las primeras diferencias para eliminar a_i nos encontramos con que la variabilidad en las variables explicativas se puede ver severamente reducida. Cuando esto ocurre los errores mínimo cuadráticos pueden ser grandes. A veces es mejor sustraer dos períodos más separados entre sí, en vez de restar dos contiguos.

Como en el caso de las secciones cruzadas fusionadas, el estudio de los efectos de una política económica en un cuasi-experimento es posible a partir de un panel de datos. Pero vuelve a aparecer aquí la diferencia esencial de los paneles de datos: las unidades muestrales son las mismas en todos los períodos. Veamos separadamente el caso.

Si la variable y_{it} representa el resultado que tratamos de medir, y la variable $prog_{it}$ es la variable que señala si la unidad muestral ha participado o no en el programa de política económica, o en el experimento, tendremos que el modelo sería

$$y_{it} = \alpha_0 + \delta_2 d2_t + \beta_1 prog_{it} + a_i + u_{it}$$

Suponemos ahora que la participación en el programa tiene lugar sólo para algunas unidades muestrales en el segundo período. La ecuación en primeras diferencias sería:

$$\Delta y_i = \delta_2 + \beta_1 \, prog_{it} + \Delta u_i$$

donde

$$\hat{\beta}_1 = \Delta\bar{y}_{trat} - \Delta\bar{y}_{contr}$$

Que es la diferencia entre los cambios medios en la variable *y* entre los dos períodos para los grupos de tratamiento y control. Se trata de la versión para datos de panel del estimador de diferencias-en-diferencias.

Si hay más variables explicativas, que varían con el tiempo, no cambia nada de lo anterior.

Otra posible generalización se da cuando disponemos de más de dos períodos. En vez del modelo para dos períodos:

$$y_{it} = \alpha_0 + \delta_2 d2_t + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + a_i + u_{it} \qquad t = 1,2$$

… tendremos un modelo para N unidades muestrales pero para más períodos (supongamos que T = 3):

$$y_{it} = \alpha_0 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + a_i + u_{it} \qquad t = 1,2,3$$

donde, como sabemos d2=1 si la observación pertenece al segundo período, d3=1 si pertenece al tercero, etc. Con lo que la constante para el tercer período será ($\alpha_0 + \delta_2 + \delta_3$), para el segundo ($\alpha_0 + \delta_2$) y para el primero α_0. Como en el caso de dos períodos, si a_i está correlacionado con cualquiera de las variables explicativas que varían en el tiempo, las secciones cruzadas fusionadas nos llevarán a estimaciones sesgadas e inconsistentes. Por tanto, debe cumplirse que $Cov(x_{kit}, u_{it}) = 0$ para todo i, k y t. Como ya hemos visto, este supuesto de exogeneidad estricta, típico de la estimación de mínimos cuadrados, impide que podamos usar como variable explicativa una variable dependiente de un período anterior (por ejemplo, y_{it-1}).

Este supuesto tampoco se cumple si hemos omitido una variable que cambia en el tiempo.

Hasta aquí, lo que ya hemos visto, sólo que aplicado a más variables. Dado que disponemos ahora de 3 períodos, pasamos a restar el primero del segundo, y el segundo del tercero, de forma que:

$$y_{i3} = (\alpha_0 + \delta_2 + \delta_3) + \beta_1 x_{1i3} + \beta_2 x_{2i3} + \ldots + \beta_k x_{ki3} + a_i + u_{i3} \qquad t = 3$$

$$y_{i2} = (\alpha_0 + \delta_3) + \beta_1 x_{1i2} + \beta_2 x_{2i2} + \ldots + \beta_k x_{ki2} + a_i + u_{i2} \qquad t = 2$$

$$y_{i1} = \alpha_0 + \beta_1 x_{1i1} + \beta_2 x_{2i1} + \ldots + \beta_k x_{ki1} + a_i + u_{i1} \qquad t = 1$$

Ahora sustraemos la primera ecuación de la segunda, y nos queda

$$(y_{i2} - y_{i1}) = \delta_2 + \beta_1 (x_{1i2} - x_{1i1}) + \beta_2 (x_{2i2} - x_{2i1}) + \ldots + \beta_k (x_{ki2} - x_{ki1}) + (u_{i2} - u_{i1})$$

y la segunda de la tercera:

$$(y_{i3} - y_{i2}) = \delta_3 + \beta_1 (x_{1i3} - x_{1i2}) + \beta_2 (x_{2i3} - x_{2i2}) + \ldots + \beta_k (x_{ki3} - x_{ki2}) + (u_{i3} - u_{i2})$$

Dos ecuaciones que pueden representarse de forma compacta como:

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \ldots + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it} \qquad t = 2,3$$

por tanto, tenemos ahora *dos* ecuaciones en diferencias para cada unidad muestral i. Si los supuestos básicos de la estimación por mínimos cuadrados se cumplen, podremos fusionar los datos y obtener estimadores insesgados. Es clave el supuesto de exogeneidad, es decir, que los Δu_{it} estén incorrelados con los Δx_{jit} para todo j (variable) y t (períodos temporales involucrados).

Es importante señalar que para t = 2 tendremos que $\Delta d2 = 1$ y $\Delta d3 = 0$, mientras que para t = 3 tendremos que $\Delta d2 = -1$ y $\Delta d3 = 1$. Por tanto, el modelo con tres períodos, y dos diferencias entre períodos, no tiene constante, y esto afecta a la posibilidad de calcular un R cuadrado. Para evitar este inconveniente se hace lo siguiente: se incluye en el modelo original una sola *dummy* temporal, normalmente en el tercer período, y una constante. Tendríamos una forma modificada del tipo:

$$\Delta y_{it} = \alpha_0 + \delta_3 \Delta d3_t + \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \ldots + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it} \qquad t = 2,3$$

siendo las estimaciones para los β_j idénticas a las del modelo anterior, sólo que ahora hay una constante. La obtención de esta forma es menos elegante que la anterior, pero con ella podemos calcular un R cuadrado. Salvo que tengamos un interés especial en las *dummies* temporales originales, esta segunda forma es más práctica.

En general, para un *panel equilibrado* (*balanced panel*), en el que todas las unidades muestrales están presentes todos los períodos, tendremos un T (periodos) relativamente pequeño con respecto a N (unidades en la muestra) debería incluirse una variable ficticia para cada período, de forma que los cambios seculares queden recogidos en ella. Si generalizamos el modelo adaptado (para incluir una constante) que acabamos de ver, tendremos:

$$\Delta y_{it} = \alpha_0 + \delta_3 \Delta d3_t + \delta_4 \Delta d4_t + \ldots + \delta_T \Delta dT_t + \beta_1 \Delta x_{1it} + \beta_2 \Delta x_{2it} + \ldots + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it}$$
$$t = 2,3, \ldots,T$$

Como puede verse, es una versión generalizada del modelo anterior, donde se continúa añadiendo variables a partir del tercer período. Tendremos T-1 períodos para cada unidad i, siendo el número total de observaciones N(T-1). La estimación se hace mediante mínimos cuadrados, siempre que se cumplan los supuestos claves, y siempre que el panel esté correctamente organizado para la fusión de datos, de forma que al calcular primeras diferencias el panel quede bien ordenado. Para ello se suelen colocar primero las T observaciones originales de la primera unidad muestral (i=1), a continuación las T observaciones de la segunda unidad (i=2), etcétera. Hay que tener cuidado de no restar a la primera observación (T=1) del individuo i-ésimo la última (T=t) del anterior individuo. Programas como Stata permiten una preparación de los datos originales correcta.

Los supuestos que permiten emplear mínimos cuadrados son los usuales, pero hay que ser cuidadoso una vez más con ellos. Recordamos algunas reglas básicas. Primero, debemos asumir que los Δu_{it} están incorrelacionados en el tiempo. Sin embargo, cuando los errores idiosincráticos originales, los u_{it}, están incorrelacionados en el tiempo, no podremos asumir que los Δu_{it} también lo están. Por ejemplo, si los u_{it} están incorrelacionados serialmente con varianza constante, entonces la correlación entre Δu_{it} y Δu_{it+1} es -0,5; si los u_{it} siguen un proceso AR(1) estable –cada error viene explicado por una función lineal del error del período anterior, más un término de error– entonces los Δu_{it} están serialmente correlacionados. De hecho, sólo cuando los u_{it} siguen un *paseo aleatorio* los Δu_{it} estarán incorrelacionados. Es fácil comprobar si hay correlación serial en los términos Δu_{it}. En efecto, si éstos siguieran un modelo AR(1) tendríamos que $\Delta u_{it} = \rho \Delta u_{it-1} + e_{it}$. Podemos por tanto estimar mediante mínimos cuadrados el modelo, tomar las estimaciones de los Δu_{it}, estimar el modelo autorregresivo con los residuos como variables y comprobar si se cumple la hipótesis nula según la cual

ρ=0 (podemos mirar el estadístico t para el coeficiente en cuestión). ¿Qué ocurre si detectamos la presencia de correlación serial? Hay que tratar el problema mediante una estimación basada en mínimos cuadrados generalizados, utilizando una transformación de Prais-Winstein. Los programas para estimación asumen que el proceso AR(1) se da entre observaciones, sin atender a las dos dimensiones de las que forman parte de un panel de datos (i y t), mientras que nosotros tenemos el problema sólo en la dimensión t (puesto que las observaciones son independientes entre unidades muestrales). Para un análisis de los procedimientos adecuados para estos casos hay que acudir a una referencia relativamente avanzada, como Wooldridge (2002, capítulo 10). Aquí basta con tomar nota de la naturaleza de los problemas que nos podemos encontrar al trabajar con paneles de datos estimando mediante modelos en primeras diferencias.

***

Tomar primeras diferencias, fusionar datos y aplicar mínimos cuadrados no es la única forma de tratar con los efectos fijos a_i. Existen otras formas de hacerlo, con sus particularidades específicas, y sus ventajas asociadas.

Uno de estos métodos se llama, precisamente, *estimación de efectos fijos*, o transformación de efectos fijos. Veamos la esencia de este procedimiento con un sencillo modelo de una sola variable explicativa:

$$y_{it} = \beta_1 x_{it} + a_i + u_{it} \qquad t = 1,2,\ldots,T$$

Ahora calculamos las medias temporales para cada unidad muestral i, de forma que tenemos

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

donde, como es lógico, hablamos de medias aritméticas, de forma que $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$. Dado que a_i permanece constante en el tiempo, aparece sin cambio en ambas ecuaciones. Ahora restamos a la primera ecuación la segunda, para cada período de tiempo, y obtenemos:

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i \qquad t = 1,2,\ldots,T$$

que podemos expresar como

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it} \qquad t = 1,2,\ldots,T$$

la variable dependiente representa los datos *en desviaciones con respecto a la media* (*time-demeaned*) relativos a *y*, y lo mismo pasa con los errores y la variable dependiente. A esta transformación de efectos fijos se la conoce también como *transformación intragrupos* (*within transformation*). Como puede verse, el objetivo se ha conseguido: la heterogeneidad inobservable ha desaparecido.

El procedimiento es, a partir de este punto, el de siempre: se fusionan todos los datos con la resta de los promedios temporales (*time-demeaned data*) y se estima por mínimos cuadrados. Al estimador mínimo cuadrático que opera sobre este tipo de datos se le conoce como *estimador de efectos fijos* o bien *estimador intragrupos* (*within estimator*) debido a que el estimador mínimo cuadrático está considerando la variación de *x* e *y* para cada observación de la sección cruzada. No podemos incluir en el modelo original variables explicativas que no varíen en el tiempo, ya que coincidirá con su media temporal y al restar desaparecerá. La constante que podamos incorporar al modelo original también desaparecerá.

El modelo se generaliza muy fácilmente para incluir más variables explicativas, incluidas *ficticias* temporales. Se opera exactamente igual: se restan las medias (también a las *ficticias* temporales), se fusionan los datos y se aplican mínimos cuadrados.

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + \ldots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it} \qquad t = 1,2,\ldots,T$$

El supuesto de exogeneidad estricta debe cumplirse también para garantizar que el estimador de efectos fijos es insesgado. En general, podemos decir que el error idiosincrático debe estar incorrelado con cada una de las variables explicativas de todos los períodos temporales. El estimador, sin embargo, permite cualquier correlación entre el efecto fijo a_i y las variables explicativas en cualquier período de tiempo, y no resulta afectado por ello. Los errores idiosincráticos u_{it} deben ser homocedásticos y deben estar serialmente incorrelados.

A la hora de computar los grados de libertad del estimador de efectos fijos hay que tener cuidado. Cuando estimamos la ecuación en desviaciones con respecto a la media (con las variables marcadas con dos puntitos) mediante mínimos cuadrados sobre datos fusionados tenemos un total de NT observaciones y k variables independientes, lo que nos puede llevar a pensar que disponemos de NT-k grados

de libertad. Esto es incorrecto. Para cada unidad muestral i perdemos un grado de libertad al restar la media temporal, dado que para cada i la suma temporal de los errores $ü_{it}$ debe ser cero (restricción que los errores idiosincráticos originales no soportan). El número correcto de grados de libertad es *gl* = NT – N – k = N(T-1) – k. Los programas informáticos hacen el cálculo de forma correcta.

No se pueden incluir variables constantes en el tiempo, como ya sabemos, al menos por sí mismas. Sí se pueden incorporar interaccionando con otras variables que sí varíen en el tiempo, y también con las *ficticias* temporales. Por ejemplo, en una ecuación de salarios en la que la educación no cambia para los individuos incluidos en la muestra durante el corto período de tiempo recogido en el panel, podemos interaccionar la educación con cada *ficticia* anual para ver si, en promedio, el rendimiento de la educación ha *cambiado* con los años. Pero no podemos calcular el rendimiento de la educación en el período base, y por tanto tampoco en los demás. El comportamiento de la variable que queremos introducir en el modelo puede dar problemas. Si la variable cambia en el tiempo, pero de forma constante, como hacen los años de experiencia, su efecto sobre la variable dependiente no se puede medir en presencia de *ficticias* temporales, que ya recogen el mismo efecto de cambio anual aplicable a todos los miembros en la muestra.

Hay un aspecto interesante que debemos considerar en el contexto de los modelos de efectos fijos. En teoría, el efecto fijo a_i para cada persona podría estimarse. Sólo tendríamos que incluir una variable ficticia (*dummy*) para cada unidad muestral en un modelo normal (sin restar las medias). El número de variables explicativas aumenta en N. El panel de datos, al proporcionar más observaciones (NT en vez de las N de una sección cruzada), podría permitir esta estimación si N + k < NT. Cuando N es un número muy elevado, la cosa se complica. Lo interesante del caso es que las estimaciones de los coeficientes β_j dan exactamente *lo mismo* en el modelo de efectos fijos que en esta *regresión con variables ficticias* (*dummy variable regression*). El R cuadrado es en este caso muy elevado, pues hemos introducido muchas variables en el modelo. Si comprobamos la significatividad conjunta de las ficticias (*dummies*) veremos que casi siempre son significativas. Los coeficientes estimados de estas *ficticias* pueden tener interés, sobre todo para ver cómo se distribuyen esos valores entre las unidades muestrales, o para ver si una en particular está por encima o por debajo de la media. Gracias a que los coeficientes β_j son los mismos que en el modelo de efectos fijos, podemos estimar éste y después calcular el efecto fijo estimado de cada unidad muestral, es decir, $\hat{a}_i$, haciendo

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \hat{\beta}_2 \bar{x}_{i2} - ... - \hat{\beta}_k \bar{x}_{ik} \qquad i = 1,...,N$$

donde las medias son temporales y los coeficientes beta son estimaciones a partir de los estimadores de efectos fijos. Con ese estadístico podemos saber, después de una estimación de un modelo de efectos fijos, si el efecto inobservado de una determinada unidad muestral está por encima de la media o no. Pensemos, por ejemplo, en salarios, o en criminalidad. Bajo los supuestos típicos de mínimos

cuadrados, la $\hat{a}_i$ es un estimador insesgado, pero no consistente con un T fijo conforme aumentamos N: cada nueva unidad muestral introduce una nueva variable, mientras que al permanecer T fijo no se incorpora nueva información en ella.

Si el programa econométrico con el que hacemos las estimaciones nos da un valor para la constante al estimar un modelo de efectos fijos, esa constante es la media de las $\hat{a}_i$. En general es mejor pensar en las a_i como variables omitidas que la transformación intragrupos permite eliminar.

¿Qué es preferible, tomar primeras diferencias o diferencias con respecto a la media? Cuando T = 2 todas las estimaciones y tests son idénticos en uno y otro caso. Primeras diferencias es un método más rápido y sencillo, y es fácil obtener estadísticos robustos a la heterocedasticidad para ese método.

Cuando $T \geq 3$ los estimadores ya no son iguales. Los dos son insesgados, y consistentes (para T fijo, conforme aumentamos N), ambos bajo los supuestos típicos de los mínimos cuadrados ordinarios. Por tanto, para N grande y T pequeño, la elección dependerá de la *eficiencia relativa* de ambos estimadores, y esto dependerá a su vez de la correlación serial de los errores idiosincráticos u_{it}.

Cuando los u_{it} están serialmente incorrelados, *el estimador de efectos fijos es más eficiente* que el de primeras diferencias, y sus errores estándar son válidos. Dado que implícitamente se supone que dicha condición se cumple, el estimador de efectos fijos se usa más a menudo que el de primeras diferencias. No obstante, debemos tener cuidado, pues cuando la incorrelación serial no se de la valoración debe cambiar. Si los u_{it} siguen un paseo aleatorio (que implica una sustancial correlación serial positiva) los Δu_{it} están serialmente incorrelados, y *el estimador de primeras diferencias es mejor*. Hay otros muchos casos intermedios en los que los u_{it} muestran alguna correlación serial positiva, pero no tanta como para que el estimador de primeras diferencias sea claramente superior en eficiencia.

Lógicamente, lo anterior plantea el problema de cómo saber si los u_{it} están correlacionados serialmente, o en qué grado. Obviamente no podemos estimar los errores idiosincráticos. Sí podemos comprobar, como vimos, si los errores Δu_{it} están serialmente incorrelados. Si lo están, el estimador de primeras diferencias puede usarse. Si hay una correlación serial negativa sustancial lo más probable es que el estimador de efectos fijos sea el mejor. En términos prácticos, es mejor emplear los dos, porque si los resultados no difieren sustancialmente elegir uno u otro no es un problema.

Cuando T es grande, y sobre todo cuando N no es muy grande, hay que tener cuidado con el estimador de efectos fijos, pues la inferencia puede verse severamente alterada por la violación de alguno de los supuestos bajo esas condiciones. El estimador de primeras diferencias es más robusto en los casos en los que T es grande y N relativamente pequeño (por ejemplo T = 30 y N = 20). También es más sensible el estimador de efectos fijos a la no normalidad de los términos aleatorios, a la heterocedasticidad y a la correlación serial de los errores idiosincráticos.

Sin embargo, y por otro lado, el estimador de efectos fijos es menos sensible a la violación del supuesto de estricta exogeneidad, especialmente cuando T es grande. Hay quien recomienda incluso estimar un modelo de efectos fijos con variables dependientes retardadas, cosa que viola directamente el supuesto de exogeneidad estricta.

Cuando los resultados de uno y otro estimador difieren fuertemente es difícil elegir uno, y se recomienda mostrar ambos y tratar de averiguar por qué difieren.

El modelo de efectos fijos funciona igual con *paneles incompletos* (*unbalanced panels*), que son aquellos para los que a algunas unidades muestrales les falta información en alguno de los períodos (a esto se lo conoce como *desgaste*, *attrition* en inglés). Puede haber problemas con este tipo de paneles dependiendo de la causa por la que esa información no aparece en el panel. Si la causa está correlacionada con el error idiosincrático u_{it} (factores inobservados que cambian con el tiempo) tendremos un problema de selección de la muestra que puede llevar a un problema de sesgo en los estimadores. Sin embargo, el modelo de efectos fijos tiene una interesante ventaja, y es que permite que el *desgaste* esté correlacionado con a_i. La idea es que en la muestra inicial algunas unidades muestrales tendrán desde el principio mayor probabilidad de desaparecer en sucesivas muestras, y esa predisposición a desaparecer, individual, sí queda recogida en la variable a_i. Sin embargo, un tratamiento de los problemas de *desgaste* en datos de panel es ciertamente complicado, y puede verse en Wooldridge (2002, capítulo 17).

***

Además de primeras diferencias y efectos fijos hay una tercera vía para tratar los efectos fijos inobservados en paneles de datos. Se trata de los *modelos de efectos aleatorios* (*random effects models*).

Empezamos con un modelo de efectos fijos inobservados, como siempre, sólo que añadiendo una constante que nos permite asumir que la media de a_i es cero:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + a_i + u_{it}$$

Se pueden añadir *dummies* temporales sin problemas a ese modelo básico. Cuando usamos primeras diferencias o efectos fijos la idea es eliminar a_i, ya que esa variable, inobservada, estará correlacionada con algunas de las restantes variables explicativas que sí varían en el tiempo, las x_{jit}. Sin embargo, ¿qué pasaría si a_i estuviera incorrelacionada con cada una de las variables explicativas en todos los períodos? Si hacemos una transformación para eliminar a_i en ese caso implicaría unos estimadores ineficientes.

Cuando suponemos que a_i está incorrelacionada con las variables explicativas (con cada una de ellas, para todos los períodos) la expresión anterior será un modelo de efectos aleatorios. En efecto, la condición que debe darse es:

$$\text{Cov}(x_{jit}, a_i) = 0 \qquad t = 1,2,..., T;\ j = 1,2,...,k$$

Bajo ese supuesto, que se añade a los que hemos visto para primeras diferencias y efectos fijos, podemos emplear un modelo de efectos aleatorios.

Si a_i está incorrelacionada con las variables dependientes no necesitamos para nada un panel de datos, y los coeficientes β_j pueden estimarse con una sola sección cruzada de forma consistente. Estaremos, eso sí, desaprovechando mucha información contenida en los demás períodos del panel. Podríamos hacer una estimación con mínimos cuadrados sobre los datos fusionados (con algunas *dummies* temporales añadidas). De esta forma también tendremos estimadores consistentes para los coeficientes β_j si se cumplen los supuestos del modelo de efectos aleatorios (básicamente, los propios de mínimos cuadrados más el de incorrelación de a_i con las demás explicativas).

No obstante, antes de cerrar conclusiones, es oportuno fijar nuestra atención en una característica interesante del modelo. Si definimos el *término de error compuesto* como $v_{it} = a_i + u_{it}$ podremos escribir el modelo como

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + v_{it}$$

Debido a que a_i forma parte del error compuesto en cada período, la variable v_{it} presentará correlación serial en el tiempo. De hecho, bajo los supuestos del modelo de efectos aleatorios,

$$\text{Corr}(v_{it}, v_{is}) = \sigma^2_a/(\sigma^2_a + \sigma^2_u), \qquad t \neq s$$

donde $\sigma^2_a = \text{Var}(a_i)$ y $\sigma^2_u = \text{Var}(u_{it})$. Esto implica una sustancial correlación serial positiva en los términos de error necesariamente. La consecuencia es que *no se pueden aplicar unos mínimos cuadrados sobre datos fusionados*.

Obviamente, se pueden aplicar mínimos cuadrados generalizados para estimar un modelo con correlación serial (véase Wooldridge 2002, capítulo 10), pero es necesario que N sea grande y T relativamente pequeña. El método funciona con paneles completos, o incompletos o desequilibrados (*balanced and unbalanced panels*).

Los *mínimos cuadrados generalizados* requieren una transformación. Definimos

$$\lambda = 1 - [\sigma^2_u/(\sigma^2_u + T\sigma^2_a)]^{1/2}$$

siendo $0 < \lambda < 1$. Transformamos ahora la ecuación original de forma que tenemos:

$$y_{it} - \lambda\bar{y}_i = \beta_0(1-\lambda) + \beta_1(x_{1it} - \lambda\bar{x}_{1i}) + ... + \beta_k(x_{kit} - \lambda\bar{x}_{ki}) + (v_{it} - \lambda\bar{v}_i)$$

donde las medias se calculan para la variable a lo largo de todos los períodos. En cierta forma, los datos están expresados en una especie de desviaciones con respecto a la media. Se está restando *una fracción* de la media temporal donde la fracción depende de λ, que depende a su vez de σ^2_u, σ^2_a y de T. El estimador de mínimos cuadrados generalizados es simplemente el estimador mínimo cuadrático para datos fusionados después de la transformación que acabamos de ver. Los errores del modelo ya no están correlacionados serialmente, aunque la demostración no es sencilla.

La transformación permite que persistan variables explicativas constantes en el tiempo, cosa que no ocurría con el modelo de efectos fijos ni con el modelo de primeras diferencias, ya que el modelo de efectos aleatorios asume incorrelación de a_i con el resto de variables explicativas, sean fijas en el tiempo o no. Gracias a esto podemos usar variables como la educación en una ecuación de salarios, eso sí, siempre y cuando asumamos que la educación está incorrelacionada con a_i, que en este caso podría contener cosas como la habilidad o el entorno familiar. En muchos casos, se emplean paneles de datos sólo para poder permitir que el efecto inobservado pueda estar correlacionado con las variables explicativas.

El parámetro λ no se puede conocer, pero se puede estimar, y hay distintas formas de hacerlo, a partir de mínimos cuadrados fusionados o de efectos fijos (de sus residuos más bien, que se emplean para estimar las varianzas que aparecen en la expresión de λ). Para ver cómo son estos estimadores de las varianzas puede verse Wooldridge (2002), capítulo 10.

Los programas informáticos computan alguna versión del parámetro λ. El estimador de mínimos cuadrados generalizados que usa esa estimación de λ recibe el nombre de *estimador de efectos aleatorios*. Este estimador, bajo ciertos supuestos, es consistente (insesgado) y asintóticamente normal conforme N crece para un T fijo (si N es pequeño y T relativamente grande las propiedades del estimador son en gran medida desconocidas).

En cierto sentido, el modelo de efectos aleatorios incluye como casos particulares a los mínimos cuadrados fusionados y a los efectos fijos. Cuando $\lambda = 0$ tenemos mínimos cuadrados fusionados, y cuando $\lambda = 1$ tenemos efectos fijos. En la práctica la *estimación* de λ nunca es cero o uno, pero si se acerca mucho a 0 las estimaciones del modelo de efectos aleatorios se parecerán mucho a las de los

mínimos cuadrados fusionados, y si se acerca mucho a 1 se parecerán a las del modelo de efectos fijos. Cuando λ se acerca mucho a 0 el efecto inobservado a_i tendrá una importancia relativamente escasa, ya que su varianza será muy pequeña relativa a σ^2_u. Será más usual que σ^2_a sea grande relativa a σ^2_u, y en ese caso la estimación de λ estará más cerca de 1. Otra relación interesante: conforme T crece la estimación de λ tiende a 1, y las estimaciones a través de efectos fijos y efectos aleatorios tienden a parecerse mucho.

Podemos ver más relaciones mediante una pequeña transformación. Escribimos el error compuesto de la siguiente forma

$$v_{it} - \lambda\bar{v}_i = (1-\lambda)a_i + u_{it} - \lambda\bar{u}_i$$

Podemos ver claramente en esa expresión cómo el error compuesto de la ecuación transformada que se usa para la estimación de efectos aleatorios pondera los efectos inobservados con un factor (1-λ). Una correlación entre a_i y una o más de las x_{jit} hace que la estimación de efectos aleatorios sea inconsistente, pero la correlación se atenúa por el factor (1-λ). Conforme λ tiende a 1 el sesgo tiende a cero, como es lógico si tenemos en cuenta que el estimador de efectos aleatorios tiende al estimador de efectos fijos. Sin embargo, si λ tiende a 0 estaremos dejando una porción mucho mayor del efecto inobservado dentro del término de error (compuesto), y por tanto el sesgo asintótico del estimador de efectos aleatorios será mayor.

¿Qué tipo de estimación es preferible? En la literatura sobre el tema se dice que cuando a_i es un parámetro a estimar debe usarse efectos fijos, y cuando es resultado de una variable aleatoria debe usarse efectos aleatorios. Por ejemplo, si las observaciones no son muestras aleatorias de una gran población, como cuando tenemos estados, regiones o provincias como unidades muestrales, es más lógico pensar que a_i es un parámetro a estimar, y debe usarse efectos fijos, que implica permitir una constante distinta para cada unidad muestral como se recordará. Si decidimos tratar a_i como si fuera una variable aleatoria tendremos que pensar si ésta está incorrelada o no con el resto de variables explicativas. Si asumimos que lo está, podemos usar efectos aleatorios, pero sólo en ese caso. Si a_i está, o pensamos que está, correlacionada con alguna de las variables explicativas, entonces tenemos que emplear también efectos fijos, o bien primeras diferencias. Comparar las estimaciones de efectos fijos y efectos aleatorios puede ser un buen test para ver si hay correlación entre a_i y las x_{jit}, siempre que los errores idiosincráticos y las variables explicativas estén incorrelados para todos los períodos (supuesto de exogeneidad). Este test fue propuesto por Hausman (1978) y muchos programas informáticos lo computan. Puede verse en detalle en Wooldridge (2002), capítulo 10.

***

Para acabar esta introducción es conveniente revisar aplicaciones de datos de panel a casos ligeramente a los descritos hasta aquí. Los métodos que hemos visto pueden aplicarse a casos en los que no hay variación temporal, es decir, a casos en los que T no indica sucesivos períodos de tiempo. Un ejemplo es el uso de datos de hermanos (t=1,2) para poder aplicar primeras diferencias y eliminar así factores no observables que afectan a ambos, y que están relacionados con el entorno familiar, como hacen Geronimus y Korenman (1992) al estudiar el impacto de los embarazos en adolescentes en la futura situación económica de la familia; o Ashenfelter y Krueger (1994) al emplear datos de gemelos para estimar el rendimiento de la educación eliminando los factores no observados relacionados con la habilidad innata. Este tipo de muestras con dos elementos para cada unidad muestral se conoce como *muestras pareadas*. Las muestras cluster tiene una serie de unidades muestrales que se agrupan de forma determinada, como una muestra de individuos que pertenecen a distintas familias, o de alumnos que pertenecen a distintos colegios, etcétera. Las familias y los colegios son clusters. Es muy probable que haya correlaciones dentro de cada cluster, y efectos fijos relacionados con el cluster. El tamaño de cada cluster suele diferir (familias de distintos tamaños, colegios con distinto número de alumnos) por lo que estamos ante un panel incompleto o desequilibrado en el que no todas las unidades muestrales tienen el mismo valor de T.