¿Cómo citar?: Montero. R (2011): *Efectos fijos o aleatorios: test de especificación.* Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España

Efectos fijos o aleatorios: test de especificación

Roberto Montero Granados
Universidad de Granada

junio de 2011

1. Introducción

Los datos de panel[1] combinan cortes transversales (información de varios individuos en un momento dado) durante varios períodos de tiempo. El disponer de datos de panel constituye una ventaja y un inconveniente:

- ventaja porque disponemos de más datos y se puede hacer un seguimiento de cada individuo.

- inconveniente porque si todas las cualidades relevantes del individuo NO son observables entonces los errores individuales estarán correlacionados con las observaciones y los MCO serán inconsistentes.

Supongamos que el modelo que pretendemos estimar es el siguiente:

$$y_{it} = X_{it}\beta + \varepsilon_{it}$$

si no se disponen de todas las variables de influencia entonces $Cov(X_{it}, \varepsilon_{it}) \neq 0$, es decir los residuos no son independientes de las observaciones por lo que MCO estará sesgado. Para solucionarlo se proponen modelos alternativos a la regresión agrupada (pooled) mediante el anidamiento de los datos: el de efectos fijos y el de efectos aleatorios.

2. Regresión agrupada (pooled)

Este modelo es el elemental. Estima el siguiente modelo:

$$y_{it} = \alpha + \beta X_{it} + u_{it} \qquad (1)$$

Como se ha mencionado, es posible que $Cov(X_{it} ; u_{it}) \neq 0$, entonces la regresión agrupada estará sesgada. Muchas veces dicha correlación es debida a un error de especificación por la ausencia de alguna variable relevante o la existencia de cualidades

[1] Suponemos un panel balanceado (con todos sus datos completos). Un panel no balanceado es un panel en el que faltan algunas observaciones que se excluyen del cálculo. En este caso el sesgo también puede venir dado por la calidad de las variables observadas y la razón de que se omitan algunas observaciones.

inobservables de cada individuo. Este problema puede solucionarse con una regresión de datos anidados.

3. Efectos fijos

Los modelos de regresión de datos anidados, realizan distintas hipótesis sobre el coportamiento de los residuos, el más elemental y el más consistente es el de Efectos Fijos. Este modelo es el que implica menos suposiciones sobre el comportamiento de los residuos. Supone que el modelo a estimar es ahora:

$$y_{it} = \alpha_i + \beta X_{it} + u_{it} \tag{2}$$

Donde $\alpha_i = \alpha + v_i$, luego reemplazando en (2) queda:

$$y_{it} = \alpha + \beta X_{it} + v_i + u_{it} \tag{3}$$

es decir supone que el error (ε_{it}) puede descomponerse en dos una parte fija, constante para cada individuo (v_i) y otra aleatoria que cumple los requisitos MCO (u_{it}) ($\varepsilon_{it} = v_i + u_{it}$), lo que es equivalente a obtener una tendencia general por regresión dando a cada individuo un punto de origen (ordenadas) distinto. Esta operación puede realizarse de varias formas, una de ellas es introduciendo una dummy por cada individuo (eliminando una de ellas por motivos estadísticos) y estimando por MCO. Otra es calculando las diferencias. Así, si (3) es cierto, también es cierto que:

$$\bar{y}_{it} = \alpha + \bar{X}_{it}\beta + v_i + \bar{u}_i \tag{4}$$

y también la diferencia (3) – (4):

$$(y_{it} - \bar{y}_{it}) = (X_{it} - \bar{X}_{it})\beta + (u_{it} - \bar{u}_i) \tag{5}$$

(5) puede resolverse fácilmente por MCO. Los programas informáticos (i.e. *stata*) la estiman generalmente con este segundo método, descomponiendo, además la varianza en dos: intro y entre grupos.

3. Efectos aleatorios

El modelo de efectos aleatorios tiene la misma especificación que el de efectos fijos con la salvedad de que v_i, en lugar de ser un valor fijo para cada individuo y constante a lo largo del tiempo para cada individuo, es una variable aleatoria con un valor medio v_i y una varianza $Var(v_i) \neq 0$. Es decir la especificación del modelo es igual a (3)

$$y_{it} = \alpha + \beta X_{it} + v_i + u_{it} \tag{6}$$

salvo que ahora v_i es una variable aleatoria. Este modelo es más eficiente (la varianza de la estimación es menor) pero menos consistente que el de efectos fijos, es decir es más exacto en el cálculo del valor del parámetro pero este puede estar más sesgado que el de efectos fijos.

¿Qué significa que v_i es una variable aleatoria? Significa que no estamos seguros del valor exacto en el origen que pueda tener cada individuo sino que pensamos que este, probablemente gravitará en torno a un valor central. Eso suele ocurrir cuando tomamos una muestra de un gran universo de individuos. Por ejemplo sabemos que los niños aprueban más si estudian más y sabemos que hay niños más inteligentes que otros entonces supondremos que cada niño parte de un punto de origen distinto (probablemen superior para los individuos más inteligentes) y, a partir de ahí, existe una relación entre trabajo y calificaciones. Sin embargo no podemos evaluar a todos los niños del mundo sino sólo una muestra. En este caso es evidente que, es posible que, si en lugar de escoger esa muestra hubiésemos elegido otra los resultados del origen y de la pendiente fuesen distintos, es decir no estamos seguros del origen del que parten lo niños en función de su coeficiente intelectual, pues ¡ala! ya tenemos una v_i aleatoria.

4. Pruebas de especificación

Surgen entonces dos dudas: ¿Cuando debemos aplicar un MCO Pooled y cuando un modelo de datos anidados y, en este último caso, de entre los dos posibles cual de ambos es más procedente? Para solucionarlas debemos responder a varias preguntas:

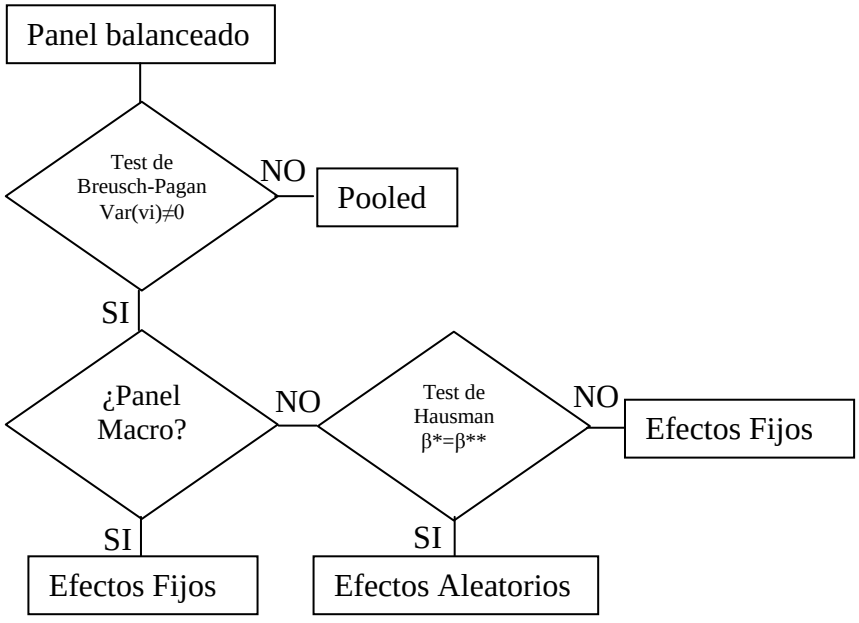
a) ¿la varianza de v_i es significativamente distinta de cero? Si la respuesta es afirmativa implica que efectivamente existe un componente inobservable de la varianza asociada a cada individuo y que MCO estará sesgado. Es decir el test de regresión anidada versus regresión agrupada (pooled) consiste en estimar si cada individuo tiene un origen en ordenadas distinto mediante la estimación de si (v_i) tiene una distribución distinta de cero. Nótese que tanto en el caso de efectos fijos (donde v_i tiene un valor constante para cada individuo pero una distribución para toda la muestra) como en el caso de efectos variables (donde v_i tiene una distribución para cada individuo) v_i siempre tiene que tener una cierta distribución (un valor y una desviación). Ojo lo importan no es que tenga un valor, ya que el valor fijo se estima en la constante del modelo, sino que lo relevante es que tenga una varianza, una distribución, significativamente distinta de cero.

b) Si la respuesta anterior es afirmativa, la siguiente es ¿Tenemos un panel en el que están TODOS los individuos del universo? En caso afirmativo se tienen que aplicar efectos fijos, si, por el contrario tenemos una muestra, más o menos representativa tendremos que pasar a la siguiente cuestión. El panel en el que están todos los individuos del universo (por ej.: todas las provincias del país. todas las empresas de conservas del mercado, etc.) también se suele llamar (de forma no muy correcta) panel macro.

c) Si la respuesta anterior es negativa, la siguiente pregunta es ¿las estimaciones consistentes (efectos fijos) y las eficientes (efectos aleatorios) son significativamente distintas? Una respuesta afirmativa implica que es mejor escoger el estimador que consideramos más consistente (el de efectos fijos), por el contrario si son ortogonalmente iguales se deberá escoger la estimación más eficiente, la de efectos aleatorios.

A la primera pregunta responde el test de Breusch-Pagan, también denominado del Multiplicador de Lagrange[2]. La prueba consiste en realizar la regresión auxiliar $indep_{it}=dep_{it}+u_i+e_{it}$ .La hipótesis nula es $Var(u_i)=0$ con una χ^2 de contraste. Si el valor del test es bajo (p-valor mayor de 0.95) la hipótesis nula se confirma y es mejor MCO. Si el valor del test es alto (p-valor menor de 0.05) la hipótesis nula se rechaza y es mejor elegir un modelo anidado.

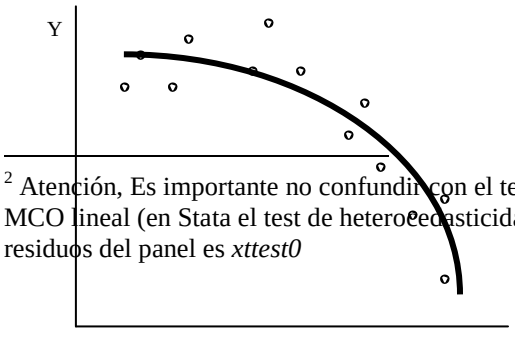
A la tercera pregunta responde el test de Hausman. El mismo compara las estimaciones del modelo de efectos fijos y el de efectos aleatorios. Si encuentra diferencias sistemáticas (se rechaza la hipótesis nula de igualdad, es decir se obtiene un valor de la prueba alto y un p-valor bajo, menor de 0.05) y siempre que estemos medianamente seguros de la especificación, podremos entender que continúa existiendo correlación entre el error y los regresores ($Cov(X_{it},u_{it}) \neq 0$) y es preferible elegir el modelo de efectos fijos.



5. Ejemplos

a) Base de datos no anidados: $Cov(X_{it}, v_i)=0$

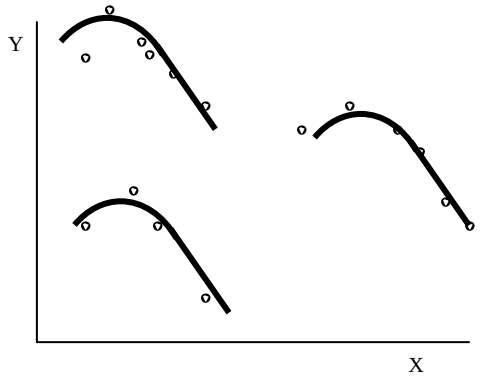
(aconsejable MCO)



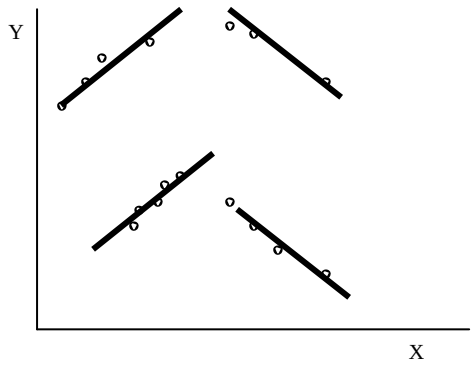
[2] Atención, Es importante no confundir con el test Breusch-Pagan de Heterocedasticidad para la regresión MCO lineal (en Stata el test de heteroscedasticidad es *estat hettest* y el que calcula la varianza de los residuos del panel es *xttest0*

b) Bases de datos anidados: $Cov(X_{it}, v_i) \neq 0$

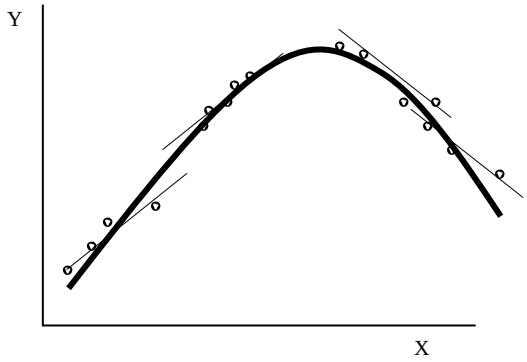
(aconsejable efectos fijos)



(aconsejable efectos aleatorios)



c) Base de datos anidados pero $Var(v_i) = 0$
(aconsejable MCO pooled)



6. Bibliografía

Breusch, T., Pagan, A. (1980): "The Lagrange multiplier and its applications to model specificatión in econometrics" *Review of Economics Studies*. 47, 239-253.

Hausman, J.A. (1978): "Specification test in econometrics". *Econometrica*. 46: 1251-1271.

Hausman, J., McFadden, C. (1984): "Specification test in econometrics", *Econometrica*, 52, 1219-1240.

Stata (2005) *Reference manual A-J.* Stata Pres. Texas, 441-448.

Stata (2005) *Longitudinal panel data.* Stata Pres. Texas, 304-310