Tema 10. Modelos de datos panel

Gustavo A. García

ggarci24@eafit.edu.co

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides formato html

Link slides formato PDF

En este tema

- Generalidades
- Tipos de paneles
- Técnicas de estimación
- Modelo de MCO agrupados (*Pooling*)
- Modelo efectos fijos: MCO con variable dicótoma (MCVD)
- Modelo efectos fijos: estimador intragrupal (*Within*)
- Modelo de efectos aleatorios
- FE vs RE: algunos lineamientos
- FE vs RE: qué dice Wooldridge
- Ejercicio aplicado en R: función de costos de líneas de aviación

Lecturas

- Wooldridge, J. (2013). *Introducción a la econometría. Un enfoque moderno*. 5a edición. Cenagage Learning Cap. 13
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 16

Generalidades

¿Por qué datos panel? ¿Cuáles son las ventajas de los datos panel?

- Como los datos de panel se refieren a individuos, empresas, estados, países, etc., a lo largo del tiempo, lo más seguro es la presencia de heterogeneidad en las unidades. Las técnicas de estimación de datos de panel toman en cuenta de manera explicita tal heterogeneidad, al permitir la existencia de variables específicas por sujeto

- Al combinar las series de tiempo de las observaciones de corte transversal, los datos de panel proporcionan una mayor cantidad de datos informativos, más variabilidad, menos colinealidad entre variables, más grados de libertad y una mayor eficiencia

Generalidades

¿Por qué datos panel? ¿Cuáles son las ventajas de los datos panel?

- Al estudiar las observaciones en unidades de corte transversal repetidas, los datos de panel resultan más adecuados para estudiar la dinámica del cambio. Los conjuntos de datos respecto del desempleo, la rotación en el trabajo y la movilidad laboral se estudian mejor con datos de panel

- Los datos de panel detectan y miden mejor los efectos que sencillamente ni siquiera se observan en datos puramente de corte transversal o de series de tiempo. Por ejemplo, los efectos de las leyes concernientes al salario mínimo sobre el empleo y los salarios, se estudian mejor si incluimos oleadas sucesivas de incrementos en los salarios mínimos estatales y/o federales

Tipos de paneles

- Panel balanceado: se dice que un panel es balanceado si cada sujeto (empresa, individuos, etc.) tiene el mismo número de observaciones
- Panel desbalanceado: si cada unidad tiene un número diferentes de observaciones
- Panel corto: el número de unidades de corte transversal, N, es mayor que el número de períodos, T
- Panel largo: el número de unidades de corte transversal, N, es menor que el número de períodos, T

Las técnicas de estimación dependen de que se cuente con un panel corto o uno largo

Técnicas de estimación

Para ilustrar este tema asumamos que tenemos información sobre los costos de 6 líneas de aviación comercial ($N = 6$) de 1970 a 1984 ($T = 15$), para un total de 90 observaciones de datos panel. Con esto en mente, las técnicas de estimación son las siguientes:

- Modelo de MCO agrupados (*Pooling*)
  Se agrupan las 90 observaciones y se estima una sola regresión, sin tener en cuenta ni la parte de corte transversal ni la parte de series de tiempo
- Modelo efectos fijos: MCO con variables dicótomas (MCVD)
  Aquí se agrupan las 90 observaciones, pero se permite que cada unidad de corte transversal (cada aerolínea) tenga su propia variable dicótoma (intercepto)
- Modelo efectos fijos: estimador intragrupal (*Within*)
  En este caso también se agrupan las 90 observaciones, pero por cada aerolínea expresamos cada variable como una desviación de su valor medio y luego estimamos una regresión de MCO sobre los valores corregidos por la media
- Modelo de efectos aleatorios
  A diferencia del modelo MCVD, en el que se permite que cada aerolínea tenga su propio valor de intercepto (fijo), suponemos que los valores del intercepto son una extracción aleatoria de una población mucho mayor de aerolíneas

Modelo de MCO agrupados (*Pooling*)

Considere el siguiente modelo a estimar, el cual representa una función de costos:

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

C: costos totales (en miles de dólares); Q: producción (ingresos por milla por pasajero); PF: precio del combustible; LF: factor de carga (la utilización promedio de la capacidad de la flotilla)

En este modelo:

- Se supone que los coeficientes de regresión son iguales para todas las aerolínas $\Longrightarrow$ no hay distinción entre aerolíneas: una aerolínea es tan buena como la otra
- No distingue entre las diferentes aerolíneas ni indica si la respuesta del costo total a las variables explicativas a través del tiempo es la misma para todas las aerolíneas $\Longrightarrow$ si agrupamos diferentes aerolíneas en diferentes períodos se oculta la heterogeneidad (individualidad o singularidad) que existen entre las unidades
- La heterogeneidad individual se subsume en u_{it}, con lo cual el término de error se correlacionará con algunos regresores, en este caso los coeficientes estimados pueden estar sesgados y ser inconsistentes

Modelo de MCO agrupados (*Pooling*)

Miremos cómo el término de error se correlaciona con los regresores. Considere el siguiente modelo:

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + \beta_5 M_i + u_{it}$$

donde M es la filosofía de la administración o calidad de la administración

Dos cosas sobre esta nueva variable

- es invariante o constante en el tiempo
- no puede observarse directamente

Una forma indirecta de medir el efecto de M es a partir del siguiente cambio en el modelo

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + \alpha_i + u_{it}$$

donde α_i representa el efecto no observado o de heterogeneidad no observable

Esta heterogeneidad no observable puede estar asociada a otras variables, como por ejemplo el grado de habilidad del gerente, si el gerente es hombre o mujer, etc.

Modelo de MCO agrupados (*Pooling*)

Una pregunta que puede surgir es si el término α_i no es observable, ¿por qué no considerarlo aleatorio e incluirlo en el término de error u_{it}? Entonces el modelo queda de la forma

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + v_{it}$$

donde $v_{it} = \alpha_i + u_{it}$

Si α_i esta correlacionado con cualquiera de los regresesores del modelo, entonces v_{it} estará correlacionado con los regresores y estaría violando el supuesto de exogeneidad. Esto implica que los estimadores por MCO son sesgados e inconsistentes

La pregunta entonces es

¿Cómo se toman en cuenta los efectos no observables o heterogeneidad no observable, para obtener estimaciones consistentes y eficientes de las variables de interés primordial?

Se debe tener en cuenta que el interés primordial no se centra en obtener el efecto de la heterogeneidad no observable, ya que esta no cambia para una unidad dada. Por esta razón, la heterogeneidad no observable se llaman parámetros incómodos ¿Cómo proceder entonces?

Modelo efectos fijos: MCO con variables dicótomas (MCVD)

El MCVD toma en cuenta la heterogeneidad entre las unidades analizadas ya que permite que cada unidad tenga su propio valor del intercepto. El modelo queda de la forma

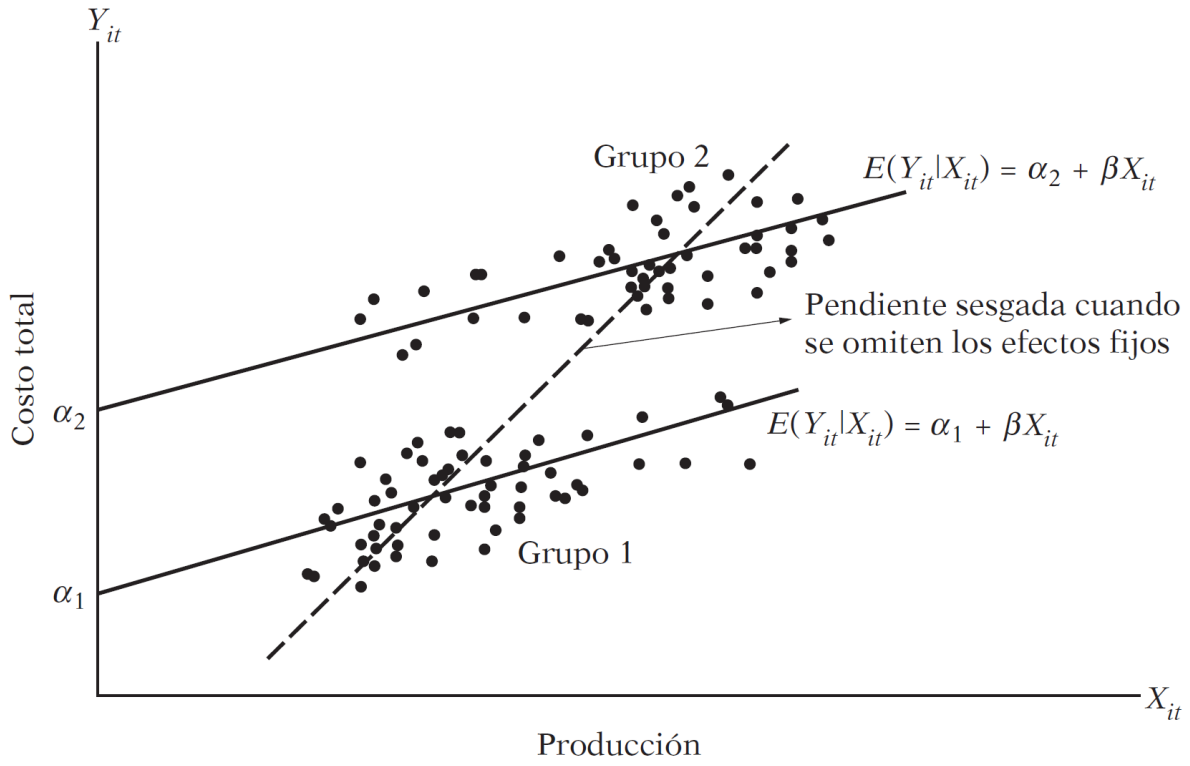
$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

Las diferencias entre unidades quizá se deban características especiales de cada aerolínea, como el estilo de administración, la filosofía de la empresa o el tipo de mercado que atiende cada aerolínea

El término de efectos fijos se debe a que, aunque el intercepto puede diferir entre las unidades, el intercepto de cada una de éstas no varía con el tiempo, es decir, es invariante en el tiempo

Modelo efectos fijos: MCO con variable dicótoma (MCVD)

Comparando el modelo *Pooling* con el MCVD, visualmente sería



La regresión agrupada sesga la estimación de la pendiente

Modelo efectos fijos: MCO con variable dicótoma (MCVD)

¿Cómo se permite en realidad que el intercepto (de efecto fijo) varíe entre unidades?

Se realiza con la técnica de variables dicótomas, en particular las variables dicótomas con intercepto diferencial. El modelo queda de la forma

$$C_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \alpha_6 D_{6i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

$D_{2i} = 1$ si corresponde a la aerolínea 2, y 0 en otro caso; y así para las otras dummies

Si los coeficientes de los interceptos diferenciales son estadísticamente significativos, indicaría que las seis aerolíneas son heterogéneas

Modelo efectos fijos: MCO con variable dicótoma (MCVD)

Es posible proporcionar una prueba para comparar el modelo *Pooling* y el de efectos fijos. Note que el modelo *Pooling* es un modelo restringido, pues se impone un intercepto común para todas las aerolíneas. Se puede utilizar la prueba F restringida

$$F = \frac{(R^2_{NR} - R^2_R)/m}{(1 - R^2_{NR})/(n-k)}$$

m: número de restricciones lineales
k: número de parámetros en la regresión no restringida
n: número de observaciones
R^2_{NR} y R^2_R: R^2 obtenidos de la regresión no restringida y restringida, respectivamente

La H_0 en este caso es que todos los interceptos diferenciales son iguales a cero. Si se rechaza H_0 indica que el modelo de efectos fijos es mejor que el modelo *Pooling*

Modelo efectos fijos: estimador intragrupal (*Within*)

Una forma de estimar una regresión agrupada es eliminar el efecto fijo, β_{1i}, expresando los valores de las variables dependiente y explicativas de cada unidad como desviaciones de sus respectivos valores medios en el tiempo.

Si para cada i se promedia la ecuación en el tiempo se obtiene

$$\bar{C}_i = \beta_{1i} + \beta_2 \bar{Q}_i + \beta_3 \bar{PF}_i + \beta_4 \bar{LF}_i + \bar{u}_i$$

donde $\bar{C}_i = \frac{\sum C_{it}}{T}$, y así sucesivamente. Si se resta la ecuación original con la anterior, se tiene

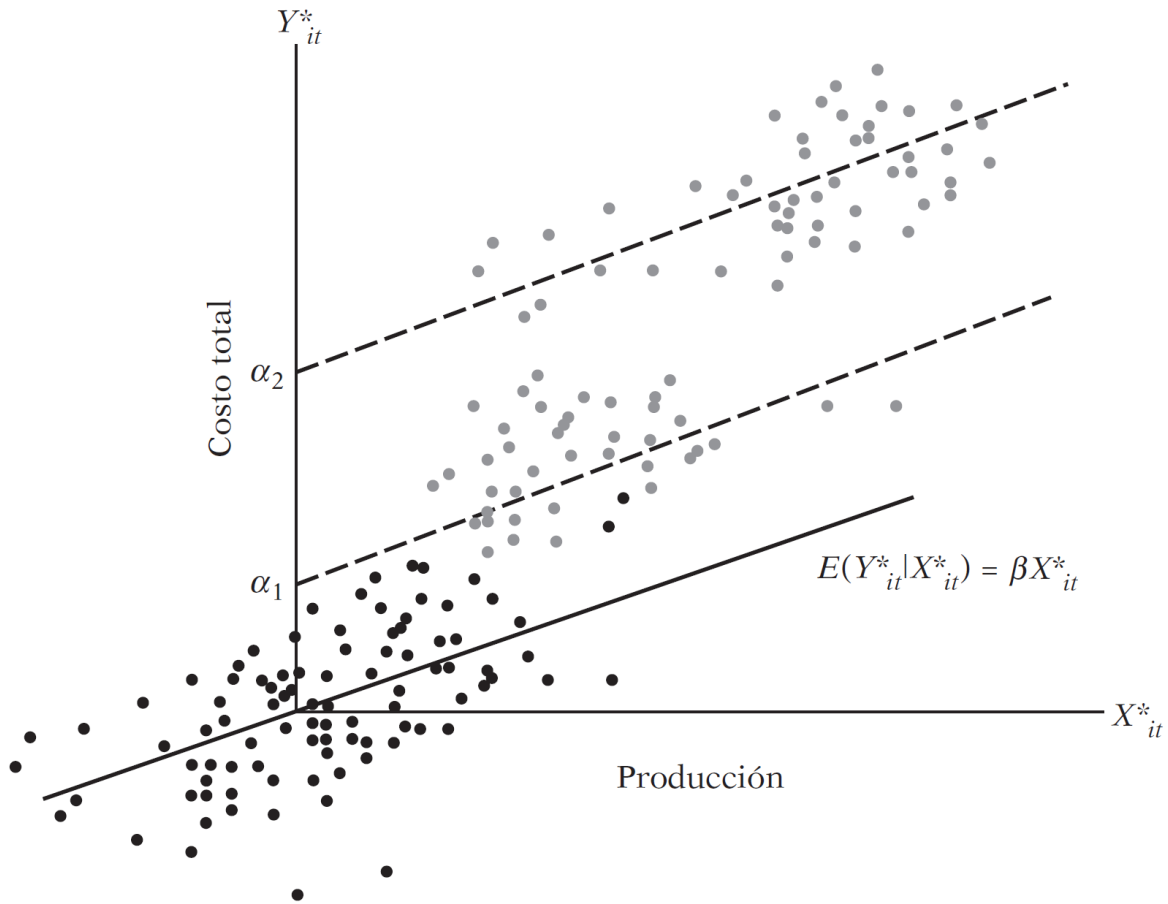
$$C_{it} - \bar{C}_i = \beta_2 (Q_{it} - \bar{Q}_i) + \beta_3 (PF_{it} - \bar{PF}_i) + \beta_4 (LF_{it} - \bar{LF}_i + (u_{it} - \bar{u}_i)$$

$$\ddot{C}_{it} = \beta_2 \ddot{Q}_{it} + \beta_3 \ddot{PF}_{it} + \beta_4 \ddot{LF}_{it} + \ddot{u}_{it}$$

Se nota que el efecto inobservable, β_{1i}, ha desaparecido

Modelo efectos fijos: estimador intragrupal (*Within*)

Gráficamente el estimador *Within* sería



Modelo efectos fijos: estimador intragrupal (*Within*)

- El estimador Within produce estimaciones consistentes de los coeficientes de pendiente, mientras que la regresión agrupada tal vez no
- Sin embargo, debe añadirse que los estimadores *Within*, aunque consistentes, son ineficientes (es decir, tienen varianzas grandes) en comparación a los *Pooling*
- Los estimadores *Within* arrojan estimaciones de las pendientes iguales a las estimaciones del MCVD, esto es porque matemáticamente los dos modelos son idénticos
- Una desventaja del modelo *Within* es que las variables invariantes en el tiempo se elimiarían del modelo, con lo cual no se sabría el efecto de la variable dependiente ante cambios en esas variables independientes invariantes en el tiempo. Pero es el precio que hay que pagar para evitar la correlación entre el término de error (α_i incluido en v_{it}) las variables explicativas
- Otra desventaja es que puede distorsionar los valores de los parámetros y desde luego eliminar los efectos de largo plazo

Modelo de efectos aleatorios

- Cuando se utilizan efectos fijos, el objetivo es eliminar β_{1i} porque se considera que está correlacionada con una o más de las X_{itj}
- Pero suponga que β_{1i} no está correlacionada con ninguna variable explicativa en todos los períodos. Entonces, el uso de una transformación para eliminar β_{1i} da como resultado estimadores ineficientes
- El modelo se vuelve el modelo de efectos aleatorios cuando se da por sentado que el efecto inobservable β_{1i} no se correlaciona con ninguna variable explicativa

Modelo de efectos aleatorios

La idea básica es comenzar con la ecuación

$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

En lugar de considerar fija a β_{1i}, suponemos que es una variable aleatoria con un valor medio igual a β_1. Además, el valor del intercepto para una empresa individual se expresa como:

$$\beta_{1i} = \beta_1 + \epsilon_i$$

donde ϵ_i es un termino de error aleatorio con valor medio igual a cero y varianza σ_ϵ^2

Lo que se afirma es que las seis empresas de la muestra se tomaron de un universo mucho más grande de este tipo de compañías, que tienen una media común para el intercepto ($= \beta_1$) y que las diferencias individuales en los valores del intercepto de cada empresa se reflejan en el término de error ϵ_i

Bajo el anterior supuesto el modelo queda de la forma

$$\begin{aligned} C_{it} =& \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + \epsilon_i + u_{it} \\ =& \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + w_{it} \end{aligned}$$

donde $w_{it} = \epsilon_i + u_{it}$

Modelo de efectos aleatorios

$$w_{it} = \epsilon_i + u_{it}$$

w_{it} presenta dos componentes

- ϵ_i: componente de error de corte transversal o error específico del individuo
- u_{it}: la combinación del componente de error de series de tiempo y corte transversal, y que a veces se denomina término idiosincrásico

Por estos dos componentes del término de error es que el modelo de efectos aleatorios también se llama Modelo de Componentes del Error (MCE)

Los supuestos comunes en los que se basa el MCE son:

- $\epsilon_i \sim N(0, \sigma_\epsilon^2)$
- $u_{it} \sim N(0, \sigma_u^2)$
- $E(\epsilon_i u_{it}) = 0$; $E(\epsilon_i \epsilon_j) = 0, i \neq j$
- $E(u_{it} u_{is}) = E(u_{ij} u_{ij}) = E(u_{it} u_{js}) = 0, i \neq j; t \neq s$

Modelo de efectos aleatorios

Diferencias entre el modelo de efectos fijo y el modelo de efectos aleatorios

En el modelo de efectos fijos, cada unidad de corte transversa tiene su propio valor (fijo) de intercepto. En el modelo de efectos aleatorios, el intercepto (común) representa el valor medio de todos los interceptos de (de corte transversal), y el componente de error ϵ_i significa la desviación (aleatoria) de intercepto individual respecto al valor medio

Como resultado de los supuestos establecidos, se deriva que

$$E(w_{it}) = 0$$

$$Var(w_{it}) = \sigma_\epsilon^2 + \sigma_u^2$$

Si $\sigma_\epsilon^2 = 0$ no hay diferencias entre el modelo *Pooling* y el modelo de efectos aleatorios, en cuyo caso se hará la regresión *Pooling*

Aunque w_{it} sea homocedástico puede demostrarse que w_{it} y w_{is} están correlacionados, se tiene entonces que

$$\rho = Corr(w_{it}, w_{is}) = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_u^2}; t \neq s$$

Si no tenemos en cuenta esta estructura de correlación y estimamos el modelo de efectos aleatorios mediante MCO, los estimadores resultantes serán ineficientes. El método más adecuado en este caso es el de Mínimos Cuadrados Generalizados (MCG)

Modelo de efectos aleatorios

La pregunta que surge ahora es cuál modelo estimar, el modelo de efectos fijos o el modelo de efectos aleatorios?

Test de Hausman

- En este test H_0 es que los estimadores del modelo de efectos fijos y el modelo de efectos aleatorios no difieren considerablemente
- El estadístico de prueba tiene una distribución asintótica χ^2
- Si se rechaza H_0, la conclusión es que el modelo de efectos aleatorios no es apropiado, ya que es probable que los efectos aleatorios estén correlacionados con una o más regresores. En este caso el modelo de efectos fijos se prefiere al modelo de efectos aleatorios

Test LM de Breusch-Pagan

- En este test H_0 es que no hay efectos aleatorios, es decir, que σ_ϵ^2 en $Var(w_{it}) = \sigma_\epsilon^2 + \sigma_u^2$ es cero
- El estadístico de prueba tiene una distribución asintótica χ^2 con 1 gdl, sólo 1 gdl ya que se esta probando una sola hipótesis
- Si no rechazamos la H_0 implica que el modelo de efectos aleatorios no es apropiado

FE vs RE: algunos lineamientos

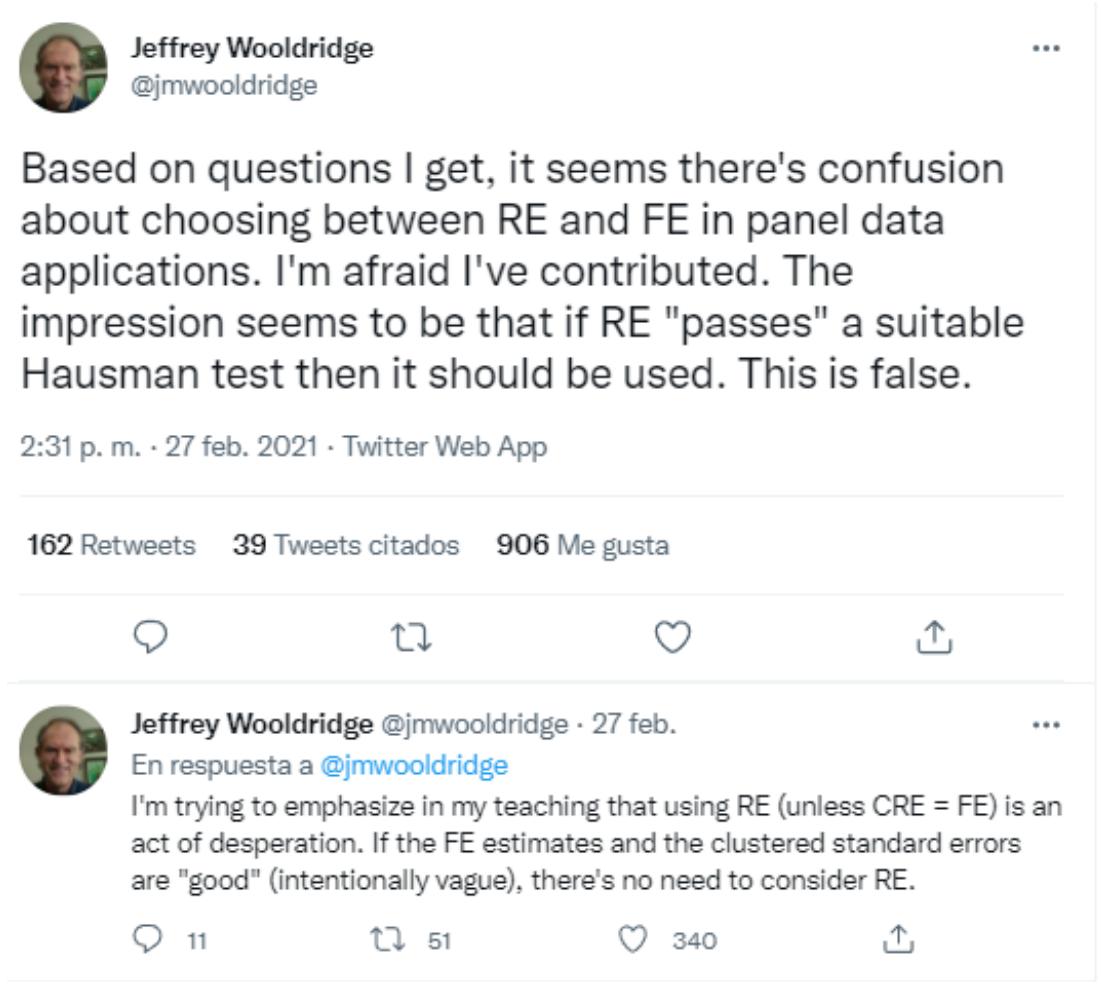
- La disyuntiva que enfrenta un investigador es: ¿qué modelo es mejor, modelo de efectos fijos o modelo de efectos aleatorios? La respuesta gira en torno del supuesto respecto de la probable correlación entre el componente de error individual, o específico de la unidad de corte transversal, ϵ_i, y las regresoras X

- Si se supone que ϵ_i y las X no están correlacionados, el modelo de efectos aleatorios puede resultar apropiado; pero si ϵ_i y las X están correlacionados, entonces el modelo de efectos fijos puede ser adecuado

- El supuesto en que se basa el modelo de efectos aleatorios es que ϵ_i representa una muestra aleatoria de una población mucho más grande, aunque no siempre es así

- Si T es grande y N es pequeño, es probable que haya muy poca diferencia entre los valores de los parámetros estimados mediante el modelo de efectos fijos y el modelo de efectos aleatorios. Por tanto, en este caso la elección se basa en la conveniencia de cálculo. Desde esta perspectiva, parece preferible el modelo de efectos fijos

FE vs RE: qué dice Wooldridge



Jeffrey Wooldridge
@jmwooldridge

Based on questions I get, it seems there's confusion about choosing between RE and FE in panel data applications. I'm afraid I've contributed. The impression seems to be that if RE "passes" a suitable Hausman test then it should be used. This is false.

2:31 p. m. · 27 feb. 2021 · Twitter Web App

162 Retweets **39** Tweets citados **906** Me gusta

Jeffrey Wooldridge @jmwooldridge · 27 feb.
En respuesta a @jmwooldridge
I'm trying to emphasize in my teaching that using RE (unless CRE = FE) is an act of desperation. If the FE estimates and the clustered standard errors are "good" (intentionally vague), there's no need to consider RE.

11 51 340

Link al tweet

Ejercicio aplicado: función de costos de líneas de aviación

En este ejercicio aplicado se va a estudiar la función de costos de 6 líneas de aviación comercial entre 1970 y 1984, para un total de 90 observaciones de datos de panel. La ecuación a estimar es la siguiente:

$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

i: identificación de la aerolínea; $t = 1970...1984$; C: costo total, en 1000 dólares; Q: producción, como ingresos por milla por pasajero (es un índice); PF: precio del combustible; LF: factor de carga, la utilización promedio de la capacidad de la flotilla

En los siguientes links se encuentran los datos y el código utilizado en R:

- Datos
- Código en R