Modelos lineales de datos panel

Gustavo A. García

ggarci24@eafit.edu.co

Econometría avanzada

Programa de Economía

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Motivación
- El problema de variables omitidas
- Algunas consideraciones
- Naturaleza de los efectos inobservables
- Estimando modelos de efectos inobservables
- Test de Hausman
- Qué dice Wooldridge entre RE y FE?
- Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Lecturas

- Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2a edición. MA: MIT Press. Cap 10

Motivación

- El uso de métodos de regresión de datos de panel se ha vuelto cada vez más popular a medida que la disponibilidad de datos longitudinales ha aumentado

- Los datos panel contienen observaciones repetidas de series de tiempo (T) para un gran número (N) de unidades transversales (por ejemplo, individuos, hogares o empresas)

- Una importante ventaja de utilizar datos panel es que permiten a los investigadores controlar la heterogeneidad no observable, esto es, las diferencias sistemáticas entre las unidades de sección transversal

- Omitiendo esta heterogeneidad no observable en los modelos de regresión que tiene parte temporal y transversal, la inferencia estadística podría ser sesgada

Motivación

- Cuando los datos de panel son disponibles, los modelos de error de componentes pueden ser usados para controla por estas diferencias individuales $\Longrightarrow$ estos modelos asumen que el término de error estocástico tiene dos componentes:
    - un efecto individual invariante en el tiempo que captura la heterogeneidad individual inobservable
    - un término de error usual
- Los efectos individuales invariantes en el tiempo son tratados como variables aleatorias, extraídas de la población junto con las variables explicativas, en oposición a la idea de parámetros a ser estimados
- Bajo este marco, la cuestión clave es si el efecto individual no observado está o no correlacionado con las variables explicativas
- Los modelos de datos panel también permiten mirar la dinámica de las relaciones, algo que no se puede en una sola sección cruzada

El problema de variables omitidas

- Cuando existen variables omitidas en un modelo de regresión, la estructura de datos panel puede ser usada para obtener estimadores consistentes

- El interés es estimar el efecto parcial de las variables explicativas observables x_j sobre la variable dependiente y, esto es

$$E(y|x_1, x_2, \ldots, x_K, c)$$

c es una variable aleatoria inobservable y nos gustará mantenerla constante cuando se obtienen los efectos parciales de las variables explicativas. Es importante resaltar que esta variable inobservable c es aleatoria y no un parámetro a estimar

- Asumiendo un modelo lineal, se tiene

$$E(y|\mathbf{x}, c) = \beta_0 + \mathbf{x}\boldsymbol{\beta} + c$$

- Si c no se encuentra correlacionado con cada x_j, entonces c será otro factor inobservable afectando y y cuyo efecto es de interés

- Si $Cov(x_j, c) \neq 0$ para alguna j, poniendo c en el término de error puede causar problemas y estimar inconsistentemente a $\boldsymbol{\beta}_{K\times 1}$

El problema de variables omitidas

- Cuando se tiene panel de datos es posible lidiar con $Cov(\mathbf{x}, c) \neq \mathbf{0}$
- Por ejemplo, supongamos que observamos y y $\mathbf{x}$ en dos periodo, con lo cual tenemos y_t y $\mathbf{x}_t$, y se supone que c no varia en le tiempo, entonces el modelo será

$$E(y_t|\mathbf{x}_t, c) = \beta_0 + \mathbf{x}_t\boldsymbol{\beta} + c, t = 1, 2$$

- c entonces es un efecto inobservable al tener el mismo efecto sobre y en cada periodo y ser constante a través del tiempo
- Este efecto inobservable es a menudo interpretado como características individuales inobservables, como habilidades cognitivas, motivación o educación familiar temprana

El problema de variables omitidas

Surge entonces un supuesto adicional para estimar $\boldsymbol{\beta}$. Reecribiendo el modelo tenemos

$$y_t = \beta_0 + \mathbf{x}_t\boldsymbol{\beta} + c + u_t$$

donde por definición el supuesto de estricta exogeneidad de las variables explicativas indica

$$E(u_t|\mathbf{x}_t, c) = 0, t = 1, 2$$

Lo que implica que

$$E(\mathbf{x}_t^{'}u_t) = \mathbf{0}, t = 1, 2$$

Dos consideraciones para estimar el model

- si se asume que $E(\mathbf{x}_t^{'}c) = \mathbf{0}$, se podrá aplicar *pooled OLS*
- si c está correlacionado con cualquier elemento de $\mathbf{x}_t$, entonces *pooled OLS* es sesgado e inconsistente $\implies$ es necesario métodos de estimación para eliminar el componente

Algunas consideraciones

- Se asume un panel balanceado: se tiene el mismo número de periodos en cada unidad de corte transversal. En paneles no-balanceados se debe tener cuidado el sesgo de selección y el *attrition*
- Nos centramos en las propiedades asintóticas de los estimadores, por tanto T es fijo y N crece sin límite, así $N \geq T$. Con un N grande es posible ver a las observaciones de seccón cruzada como independientes, idénticamente distribuidas tomadas de la población

Naturaleza de los efectos inobservables

Surge entonces una primera inquietud sobre la naturaleza de los efectos inobservables: efectos fijos o aleatorios?

El modelo de efectos inobservables puede plantearse de la siguiente forma

$$y_{it} = \mathbf{x}_{it}\beta + c_i + u_{it}, t = 1, 2, \ldots, T$$

c_i entonces representa el efecto individual o la heterogeneidad individual

La discusión entonces se centra en saber si c_i es tratado como un efecto aleatorio o como un efecto fijo $\Longrightarrow$ es una variable aleatoria o un parámetro a ser estimado

Bajo este enfoque, lo principal es saber si c_i está o no correlacionado con las variables explicativas $\mathbf{x}_{it}$

Naturaleza de los efectos inobservables

Efectos aleatorios

- $Cov(\mathbf{x}_{it}, c_i) = \mathbf{0}$
- En la literatura cuando c_i es referenciado como efecto aleatorio individual se está asumiendo que no se encuentra correlacionado con $\mathbf{x}_{it}$

Efectos fijos

- $Cov(\mathbf{x}_{it}, c_i) \neq \mathbf{0}$
- En este caso c_i es llamado efecto fijo individual

Estimando modelos de efectos inobservables

- *Pooled OLS*
- Modelo de efectos aleatorios
- Modelo de efectos fijos
- Modelo de variables dummy

Estimando modelos de efectos inobservables

Pooled OLS

Bajo ciertos supuestos, el estimador *Pooled OLS* puede ser usado para obtener estimadores consistentes de $\boldsymbol{\beta}$. Reescribiendo el modelo

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}$$

donde $v_{it} = c_i + u_{it}$, es lo que se llama los errores compuestos, que es la suma del efecto inobservable y un error idiosincrático

La estimación *Pooled OLS* es consistente si $E(\mathbf{x}_{it}^{'}v_{it}) = \mathbf{0}$, es decir si

$$E(\mathbf{x}_{it}^{'}u_{it}) = \mathbf{0}$$
$$E(\mathbf{x}_{it}^{'}c_i) = \mathbf{0}$$

Si los anteriores supuestos se cumplen, los errores compuestos serán serialmente correlacionados debido a la presencia de c_i en cada periodo de tiempo. Por tanto, la inferencia usando *Pooled OLS* requiere un estimador robusto de la matriz de varianzas y tests estadísticos robustos

Es importante tener un N grande y un T fijo cuando se utilice *Pooled OLS*, para evitar que la correlación serial afecte las estimaciones

Estimando modelos de efectos inobservables

Modelo de efectos aleatorios

Como en el caso de *Pooled OLS*, un análisis de efectos aleatorios pone a c_i en el término de error. Se imponen más supuestos que en el caso de *Pooled OLS*

Supuesto RE.1:
- $E(u_{it}|\mathbf{x}_i, c_i) = 0 \implies$ exogeneidad estricta ($\implies E(c_i u_{it}) = 0$, $E(\mathbf{x}'_{it} u_{it}) = 0$)
- $E(c_i|\mathbf{x}_i) = E(c_i) = 0 \implies$ ortogonalidad entre c_i y cada $\mathbf{x}_{it}$

La superioridad de un enfoque de efectos aleatorios sobre *Pooled OLS*, es que el primero tiene en cuenta la correlación serial en los errores compuestos, $v_{it} = c_i + u_{it}$, en un marco de mínimos cuadrados generalizados (GLS)

Escribiendo el modelo para todo T como

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i$$

$\mathbf{v}_i = c_i \mathbf{j}_T + \mathbf{u}_i$, donde $\mathbf{j}_T$ es un vector de unos de $T \times 1$

La matriz de varianza de $\mathbf{v}_i$ es

$$\boldsymbol{\Omega} = E(\mathbf{v}_i \mathbf{v}'_i)_{T \times T}$$

Para consistencia de los GLS, es necesario la usual condición de rango para GLS

Supuesto RE.2: rango $E(\mathbf{X}'_i \boldsymbol{\Omega}^{-1} \mathbf{X}_i) = K$

Estimando modelos de efectos inobservables

Modelo de efectos aleatorios

Un análisis general de mínimos cuadrados generalizados factibles (FGLS), usando un estimador de $\mathbf{\Omega}$ es consistente y asintóticamente normal cuando $N \to \infty$

Hasta este punto no se está explotando la estructura de efectos inobservables de v_{it}, así que es necesario adicionar supuestos sobre el error idiosincrático que da a $\mathbf{\Omega}$ una forma especial. Los supuestos son

- $E(u_{it}^2) = \sigma_u^2 \Longrightarrow$ Homoscedaticidad
- $E(u_{it}u_{is}) = 0 \Longrightarrow$ No autocorrelación

Bajo estos supuestos ya es posible construir la matriz de varianzas y covarianzas de $\mathbf{v}_i$ $(\mathbf{\Omega})$

Varianza: $E(v_{it}^2) = \sigma_c^2 + \sigma_u^2$
Covarianza $(t \neq s)$: $E(v_{it}^2 v_{is}^2) = \sigma_c^2$

$$\mathbf{\Omega} = E(\mathbf{v}_i\mathbf{v}_i^{'}) = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_u^2 \end{bmatrix}_{T\times T} = \sigma_u^2\mathbf{I}_T + \sigma_c^2\mathbf{j}_T\mathbf{j}_T^{'}$$

Estimando modelos de efectos inobservables

Modelo de efectos aleatorios

La correlación entre v_{is} y v_{it} es

$$Corr(v_{is}, v_{it}) = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2} \geq 0$$

Esta correlación es también el ratio de la varianza de c_i a la varianza del error compuesto, y es útil como una medida de la importancia relativa del efecto inobservable c_i

Un tercer supuesto que surge es

Supuesto RE.3:

- $E(\mathbf{u}_i\mathbf{u}_i^{'}|\mathbf{x}_{it}, c_i) = \sigma_u^2\mathbf{I}_T$
- $E(c_i^2|\mathbf{x}_i) = \sigma_c^2$

El primer supuesto es más fuerte que el supuesto visto de $e(u_{it}^2) = \sigma_u^2$ de homoscedasticidad, ya que asume que las varianzas condicionales son constantes y las covarianzas condicionales son cero

El segundo supuesto plantea que la $Var(c_i|\mathbf{x}_i) = Var(c_i)$, que es el supuesto de homoscedasticidad sobre el efecto inobservable c_i

Estimando modelos de efectos inobservables

Modelo de efectos aleatorios

Asumiendo que se tienen estimadores consistentes de σ_c^2 y σ_u^2, se tendrá un estimador para $\boldsymbol{\Omega}$

$$\widehat{\boldsymbol{\Omega}} = \widehat{\sigma}_u^2 \mathbf{I}_T + \widehat{\sigma}_c^2 \mathbf{j}_T \mathbf{j}_T'$$

El estimador FGLS que usa la anterior matriz de varianza es conocido como el estimador de efectos aleatorios

$$\widehat{\boldsymbol{\beta}}_{RE} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{y}_i \right)$$

- $\widehat{\boldsymbol{\beta}}_{RE}$ es claramente motivado por el supuesto RE.3
- $\widehat{\boldsymbol{\beta}}_{RE}$ es consistente si se cumple o no el supuesto RE.3
- Si los supuestos RE.1 y RE.2 se cumplen, $\widehat{\boldsymbol{\beta}}_{RE} \xrightarrow{p} \boldsymbol{\beta}$ cuando $N \to \infty$
- Bajo el supuesto RE.3, $\widehat{\boldsymbol{\beta}}_{RE}$ es eficiente

Estimando modelos de efectos inobservables

Modelo de efectos aleatorios

Con el fin de implementar el procedimiento de efectos aleatorios, es necesario obtener $\widehat{\sigma}_c^2$ y $\widehat{\sigma}_u^2$. Sin embargo, una estrategia más fácil es encontrar un estimador para σ_v^2, así que un estimador consistente es

$$\widehat{\sigma}_v^2 = \frac{1}{(NT-K)} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\widehat{v}}_{it}^2$$

donde $\widehat{\widehat{v}}_{it}^2$ son los residuales del *Pooled OLS*

Un estimador consistente para σ_c^2 es

$$\widehat{\sigma}_c^2 = \frac{1}{[NT(T-1)/2-K]} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \widehat{\widehat{v}}_{it}^2 \widehat{\widehat{v}}_{is}^2$$

Dado $\widehat{\sigma}_v^2$ y $\widehat{\sigma}_c^2$ se puede calcular $\widehat{\sigma}_u^2 = \widehat{\sigma}_v^2 - \widehat{\sigma}_c^2$

Estimando modelos de efectos inobservables

Modelo de efectos aleatorios

- $\widehat{\sigma}_c^2$ puede ser negativo, aunque en la mayorí de los ejercicios empíricos tiende a ser positivo. Implicaciones que sea negativo
  - correlación negativa en u_{it}, lo que significa que el primer supuesto en RE.3 se viola
  - otros supuestos también pueden ser violados
  - se deben incluir variables dummies de tiempo en el modelo si son significativas, su omisión puede llevar a correlación serial en u_{it}
  - FGLS no restringido puede ser utilizado

- Probando por la existencia de efectos inobservables
  - Si los supuestos del modelo de efectos aleatorios (RE.1-RE.3) se cumplen pero modelo contiene un efecto inobservable, el *Pooled OLS* es más eficiente
  - H_0: $\sigma_c^2 = 0$ vs H_a: $\sigma_c^2 \neq 0$
  - Si no rechazamos H$_0$ se concluye que efectos aleatorios no es apropiado. Esto es, no existe evidencia de diferencias significativas a través de las unidades de corte transversal, por tanto se puede correr un *Pooled OLS*
  - Existen dos test: Breusch-Pagan (1980) y Wooldridge (2010)

Estimando modelos de efectos inobservables

Modelo de efectos fijos

- En muchas aplicaciones el punto central al usar panel de datos es permitir que c_i este correlacionado con $\mathbf{x}_{it}$ y el modelo de efectos fijos permite esto

- El modelo de efectos fijos se escribe como

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + c_i\mathbf{j}_T + \mathbf{u}_i$$

  donde $\mathbf{j}_T$ es un vector de unos de $T \times 1$

- El primer supuesto del modelo de efectos fijos es
  Supuesto FE.1
  $E(u_{it}|\mathbf{x}_i, c_i) = 0 \Longrightarrow$ exogeneidad estricta ($\Longrightarrow E(c_i u_{it}) = 0, E(\mathbf{x}_{it}^{'} u_{it}) = 0$)

- Note que FE.1 es similar a RE.1, pero en el primero no se incluye el supuesto que $E(c_i|\mathbf{x}_i)$=0. Relajando este último supuesto (presencia de variable omitidas invariantes en el tiempo que se encuentran relacionadas con $\mathbf{x}_{it}$) se puede estimar consistentemente $\boldsymbol{\beta}$

- Entonces FE es más robusto que RE. Sin embargo, esta ventaja de FE tiene un costo: no se pueden incluir factores invariantes en el tiempo en $\mathbf{x}_{it}$ (género o raza)

Estimando modelos de efectos inobservables

Modelo de efectos fijos

Para estimar $\boldsymbol{\beta}$ bajo el supuesto FE.1 se debe transformar la ecuación para eliminar el efecto inobservable $c_i \implies$ la transformación *within* permite tal eliminación

La transformation *within*

- Promedie la ecuación $y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}$ sobre $t = 1, \ldots, T$ para obtener la ecuación de sección cruzada

$$\overline{y}_i = \overline{\mathbf{x}}_i\boldsymbol{\beta} + c_i + \overline{u}_i$$

donde $\overline{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, $\overline{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$ y $\overline{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$

- Restando el modelo original con este anterior en medias

$$y_{it} - \overline{y}_i = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\boldsymbol{\beta} + u_{it} - \overline{u}_i$$

o lo que es lo mismo

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}$$

Se observa que el efecto inobservable c_i se ha eliminado

La anterior ecuación podría ser estimada por *Pooled OLS*, sin embargo, es necesario determinar si en este modelo se cumple el supuesto $E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = \mathbf{0}$ para obtener estimadores consistentes

Estimando modelos de efectos inobservables

Modelo de efectos fijos

Entonces la pregunta es: ¿Es posible aplicar OLS al modelo *within* y obtener estimadores consistentes?

En otras palabras ¿se mantiene el supuesto $E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = \mathbf{0}$ bajo el supuesto FE.1, con lo cual es posible aplicar OLS al modelo *within* y obtener estimadores consistentes?

$$E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = E[(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)^{'}(\mathbf{u}_{it} - \overline{\mathbf{u}}_i)]$$

Bajo el supuesto FE.1 $E(u_{it}|\mathbf{x}_{it}, c_i) = 0$ se tiene que

$$E(\mathbf{x}_{it}^{'}\mathbf{u}_{it}) = 0$$
$$E(\mathbf{x}_{it}^{'}\overline{\mathbf{u}}_i) = 0$$
$$E(\overline{\mathbf{x}}_i^{'}\mathbf{u}_{it}) = 0$$
$$E(\overline{\mathbf{x}}_i^{'}\overline{\mathbf{u}}_i) = 0$$

Con lo cual

$$E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = E[(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)^{'}(\mathbf{u}_{it} - \overline{\mathbf{u}}_i)] = 0$$

Entonces aplicar OLS al modelo *within* genera estimadores consistentes de $\boldsymbol{\beta}$

Estimando modelos de efectos inobservables

Modelo de efectos fijos

De lo anterior hay otras dos implicaciones

- $E(\ddot{u}_{it}|\mathbf{x}_i) = E(u_{it}|\mathbf{x}_i) - E(\bar{u}_i|\mathbf{x}_i) = 0$
- $E(\ddot{u}_{it}|\ddot{\mathbf{x}}_{i1}, \ldots, \ddot{\mathbf{x}}_{iT}) = 0$

Lo que implica que $\ddot{\mathbf{x}}_{it}$ satisface la condición de exogeneidad estrictia y el estimador de efectos fijos o *within* de $\boldsymbol{\beta}$ será insesgado bajo el supuesto FE.1

En resumen el estimador de efectos fijos (FE) $\boldsymbol{\beta}_{FE}$ es el estimador *Pooled OLS* de la regresión

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}$$

Con el fin de asegurar que el estimador FE tenga un buen comportamiento en términos asintóticos, es necesario la condición rango esténdar sobre la matriz de variables explicatorias descontando la parte temporal, esto es

Supuesto FE.2: rango $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}^{'}\ddot{\mathbf{x}}_{it}) =$ rango $E(\ddot{\mathbf{X}}_i^{'}\ddot{\mathbf{X}}_i) = K$

Estimando modelos de efectos inobservables

Modelo de efectos fijos

El estimador de efectos fijos o estimador *within* (usa la variación temporal entre cada unidad de corte transversal) puede expresarse como

$$\boldsymbol{\beta}_{FE} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}^{'} \ddot{\mathbf{x}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}^{'} \ddot{\mathbf{y}}_{it} \right)$$

El siguiente supuesto asegura que el anterior estimador sea eficiente

Supuesto FE.3: $E(\mathbf{u}_i \mathbf{u}_i^{'} | \mathbf{x}_i, c_i) = \sigma_u^2 \mathbf{I}_T$

Estimando modelos de efectos inobservables

Modelo de variables *dummy*

- Hasta ahora se ha visto a c_i como una variable aleatoria inobservada. Sin embargo, en enfoques tradicionales c_i es un parámetro a estimar junto con $\boldsymbol{\beta}$. La pregunta que surge entonces es cómo estimar c_i?

- Una posibilidad es definir N variables *dummy*, una para cada para unidad de sección cruzada: $d_i = 1$ si $n = i$, $d_i = 0$ si $n \neq i$ y estimar una regresión *Pooled OLS* de la forma

$$y_{it} = d_i + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}$$

Entonces, por ejemplo, $\hat{c}_1$ es el coeficiente de d_1, y así se estiman los c_i. Recordar evitar la tramapa de las variables dummy excluyendo una d_i

Test de Hausman (1978)

La principal consideración para seleccionar entre el modelo de efectos aleatorios y el modelo de efectos fijos es determinar si c_i y $\mathbf{x}_{it}$ están correlacionados $\Longrightarrow$ El test de Hausman proporciona esta prueba entre estos dos modelos

La idea del test de Hausman es que, FE es consistente cuando c_i y $\mathbf{x}_{it}$ están correlacionados, pero RE es incosistente, así que si existe una diferencia estadísticamente significativa entre FE y RE es evidencia en contra del supuesto RE.1 $E(c_i|\mathbf{x}_{it}) = 0$

El estadístico de Hausman tiene la siguiente forma

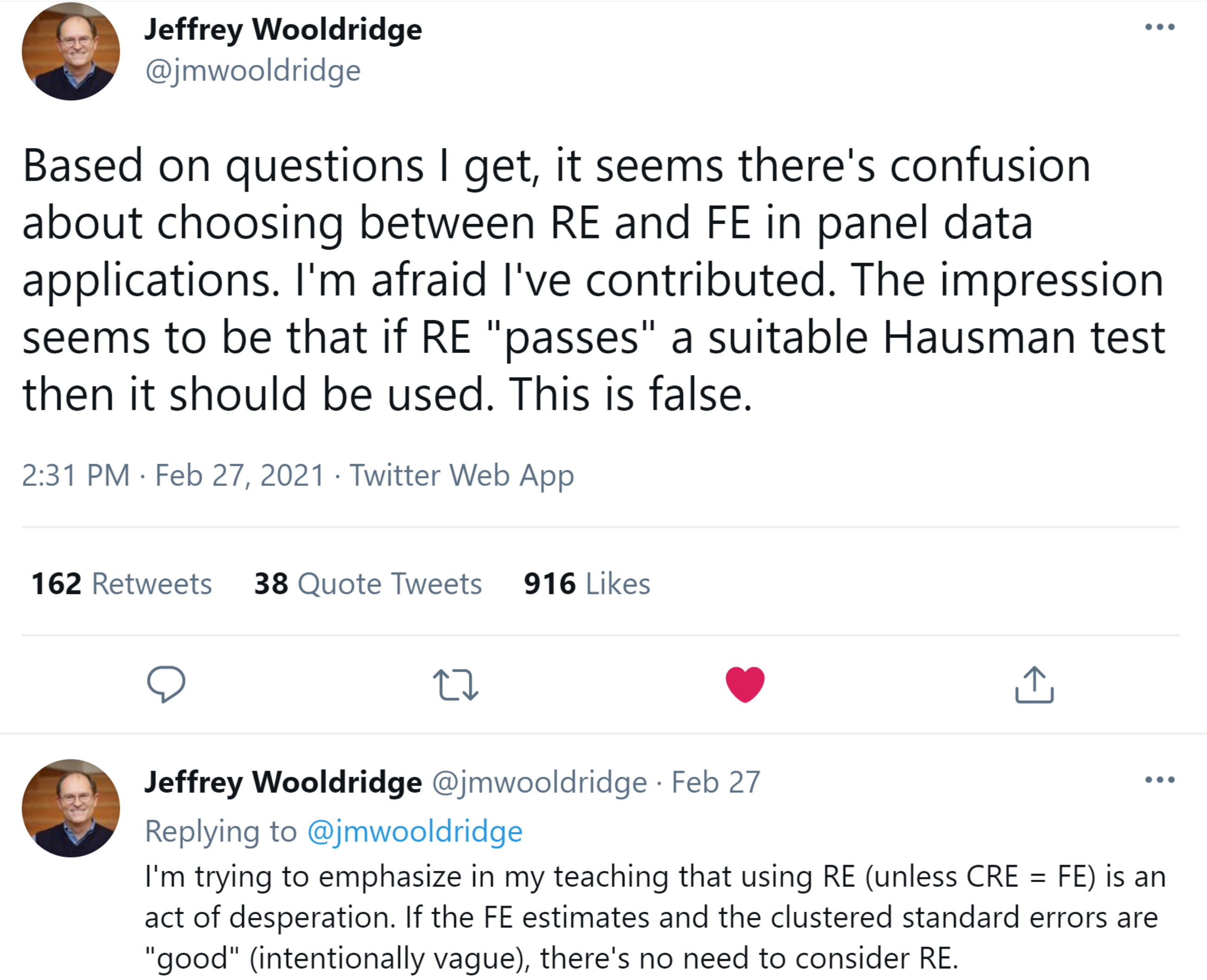
$$H = (\hat{\boldsymbol{\delta}}_{FE} - \hat{\boldsymbol{\delta}}_{RE})' [Av\widehat{a}r(\hat{\boldsymbol{\delta}}_{FE}) - Av\widehat{a}r(\hat{\boldsymbol{\delta}}_{RE})]^{-1} (\hat{\boldsymbol{\delta}}_{FE} - \hat{\boldsymbol{\delta}}_{RE}) \sim \chi^2_M$$

donde $\hat{\boldsymbol{\delta}}_{FE}$ es el vector de estimaciones del modelo de efectos fijos, $\hat{\boldsymbol{\delta}}_{RE}$ son las estimaciones del modelo de efectos aleatorios (ambos de $M \times 1$) y $Av\widehat{a}r(\hat{\boldsymbol{\delta}}_{FE})$ y $Av\widehat{a}r(\hat{\boldsymbol{\delta}}_{RE})$ son las varianzas asintóticas de los estimadores para cada modelo

H_0: RE $(Cov(c_i, \mathbf{x}_{it}) = 0)$

H_a: FE $(Cov(c_i, \mathbf{x}_{it}) \neq 0)$

Qué dice Wooldridge entre RE y FE?



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 PM · Feb 27, 2021 · Twitter Web App

162 Retweets **38** Quote Tweets **916** Likes

Jeffrey Wooldridge @jmwooldridge · Feb 27
Replying to @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.

Link al tweet

Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

En este ejercicio vamos a estudiar los derterminantes de los salarios teniendo en cuenta la heterogeneidad inobservable.

Los datos para este ejercicio proviene de la *National Longitudinal Survey of Young Working Women* de los Estados Unidos. En los siguientes links se encuentran los datos, la descripción dellada de los datos y el código utilizado en R:

- Datos
- Descripción de la información
- Código en R

Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Cargando las librerías

```r
library(haven); library(plm); library(tidyverse); library(summarytools)
library(gt); library(knitr); library(kableExtra); library(tibble); library(modelsummary)
```

Leyendo los datos y procesando la información

```r
nlswork <- read_dta("http://www.stata-press.com/data/r17/nlswork.dta") |> # leemos la base de datos
                select(idcode, year, ln_wage, age, not_smsa, south) # seleccionando variables

View(nlswork)
head(nlswork) # take a quick peak at the data
```

```
# A tibble: 6 x 6
  idcode  year ln_wage   age not_smsa south
   <dbl> <dbl>   <dbl> <dbl>    <dbl> <dbl>
1      1    70    1.45    18        0     0
2      1    71    1.03    19        0     0
3      1    72    1.59    20        0     0
4      1    73    1.78    21        0     0
5      1    75    1.78    23        0     0
6      1    77    1.78    25        0     0
```

```r
names(nlswork)
```

```
[1] "idcode"   "year"     "ln_wage"  "age"      "not_smsa" "south"
```

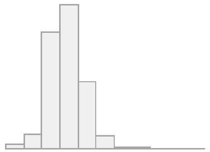
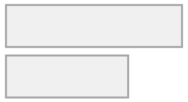
Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Estadísticas descriptivas

```r
summary(nlswork)
```

```
     idcode           year          ln_wage            age       
 Min.   :   1   Min.   :68.00   Min.   :0.000   Min.   :14.00  
 1st Qu.:1327   1st Qu.:72.00   1st Qu.:1.361   1st Qu.:23.00  
 Median :2606   Median :78.00   Median :1.641   Median :28.00  
 Mean   :2601   Mean   :77.96   Mean   :1.675   Mean   :29.05  
 3rd Qu.:3881   3rd Qu.:83.00   3rd Qu.:1.964   3rd Qu.:34.00  
 Max.   :5159   Max.   :88.00   Max.   :5.264   Max.   :46.00  
                                                NA's   :24     
    not_smsa          south       
 Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.0000  
 Mean   :0.2824   Mean   :0.4096  
 3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000  
 NA's   :8        NA's   :8       
```

```r
st_options(lang = "es", footnote=NA, headings = FALSE)
print(dfSummary(nlswork[,c("ln_wage","south")], valid.col = FALSE, silent=FALSE), method = "render", varnumbers=F)
```

Variable	Etiqueta	Estadísticas / Valores	Frec. (% sobre válidos)	Gráfico	Perdidos
ln_wage [numeric]	ln(wage/GNP deflator)	Media (d-s) : 1.7 (0.5) min < mediana < max: 0 < 1.6 < 5.3 RI (CV) : 0.6 (0.3)	8173 valores distintos		0 (0.0%)
south [numeric]	1 if south	Min : 0 Media : 0.4 Max : 1	0 : 16843 (59.0%) 1 : 11683 (41.0%)		8 (0.0%)

Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Número de idcode y años

```
length(unique(nlswork$idcode))
```

```
[1] 4711
```

```
length(unique(nlswork$year))
```

```
[1] 15
```

Determinando si el panel se encuentra balanceado

```
pdim(nlswork)$balanced
```

```
[1] FALSE
```

```
is.pbalanced(nlswork)
```

```
[1] FALSE
```

Balanceando el panel

```
nlswork_balanced <- make.pbalanced(nlswork,
                                   balance.type = "shared.individuals",
                                   index = c("idcode","year"))
pdim(nlswork_balanced)$balanced
```

```
[1] TRUE
```

Dando estructura panel a los datos

```
nlswork_balanced <-  pdata.frame(nlswork_balanced, c("idcode","year"))
```

Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Utilizamos el paquete `modelsummary` para generar tablas editadas (Word, tex, text, png, html...)

```
modelos <- list("Pool" = plm(ln_wage ~ age + I(age^2) + not_smsa + south + factor(year), data = nlswork_balanced, model = "pooling"),
                "RE"   = plm(ln_wage ~ age + I(age^2) + not_smsa + south + factor(year), data = nlswork_balanced, model = "random"),
                "FE"   = plm(ln_wage ~ age + I(age^2) + not_smsa + south, data = nlswork_balanced, model = "within", effect = "twoway"))
cm <- c('(Intercept)' = 'Constante', 'age' = 'Edad', 'I(age2)' = 'Edad2', 'not_smsa' = 'No SMSA (=1)', 'south' = 'Sur (=1)')
modelsummary(modelos, output = 'gt', coef_map = cm, stars = c('*'=.1, '**'=.05, '***'=.01), statistic = "std.error", title = 'Tabla 1. Determinantes d
    tab_style(style = cell_text(size = 'small'), locations = cells_body(rows = 1:8)) |>
    tab_style(style = cell_text(color = 'red'), locations = cells_body(rows = 1)) |>
    tab_source_note(source_note = "Nota: Errores estándar en paréntesis") |>
    tab_style(style = cell_text(color = "black", size = "x-small"), locations = cells_source_notes())
```

Tabla 1. Determinantes de los salarios

	Pool	RE	FE
Edad	0.026	0.031	0.020
	(0.023)	(0.020)	(0.036)
No SMSA (=1)	-0.201***	-0.091***	-0.055
	(0.022)	(0.033)	(0.037)
Sur (=1)	-0.157***	-0.119***	-0.090**
	(0.021)	(0.037)	(0.044)
Num.Obs.	1290	1290	1290
R2	0.233	0.298	0.007

* p < 0.1, ** p < 0.05, *** p < 0.01

Nota: Errores estándar en paréntesis

Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Los modelos

```
pool <- plm(ln_wage ~ age + I(age^2) + not_smsa + south + factor(year), data = nlswork_balanced, model = "pooling")
re <- plm(ln_wage ~ age + I(age^2) + not_smsa + south + factor(year), data = nlswork_balanced, model = "random")
fe <- plm(ln_wage ~ age + I(age^2) +  not_smsa + south, data = nlswork_balanced, model = "within", effect = "twoway")
```

Test de efectos inobservables

El test de efectos inobservables *a la* Wooldridge (ver Wooldridge (2010) 10.4.4), es un test semiparamétrico con $H_0 : \sigma^2_{c_i} = 0$, es decir que no existen efectos inobservables en los residuales

```
pwtest(ln_wage ~ age + I(age^2) + not_smsa + south, data = nlswork_balanced)
```

```
    Wooldridge's test for unobserved individual effects

data:  formula
z = 4.3892, p-value = 1.138e-05
alternative hypothesis: unobserved effect
```

Breusch-Pagan test

El test de BP es un test LM que ayuda a decidir entre RE y *pooled OLS*. La hipótesis nula es las varianzas a través de la unidades de sección cruzada son cero, esto es que no hay diferencias significativas enter unidades de corte transversal (es decir, no hay efectos panel)

```
plmtest(pool, type="bp")
```

```
    Lagrange Multiplier Test - (Breusch-Pagan) for balanced
    panels

data:  ln_wage ~ age + I(age^2) + not_smsa + south + factor(year)
chisq = 3498.5, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Ejercicio aplicado en R: determinantes de los salarios con un panel de datos

Test de *poolability*

H_0: todos los interceptos son iguales

```
pooltest(pool, fe)
```

```
    F statistic

data:  ln_wage ~ age + I(age^2) + not_smsa + south + factor(year)
F = 26.251, df1 = 85, df2 = 1186, p-value < 2.2e-16
alternative hypothesis: unstability
```

```
pFtest(fe, pool)
```

```
    F test for twoways effects

data:  ln_wage ~ age + I(age^2) + not_smsa + south
F = 26.251, df1 = 85, df2 = 1186, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Test de Hausman

```
phtest(fe, re)
```

```
    Hausman Test

data:  ln_wage ~ age + I(age^2) + not_smsa + south
chisq = 5.4579, df = 4, p-value = 0.2435
alternative hypothesis: one model is inconsistent
```