vv

Econometría I - EAE- 250-A

Datos de Panel

Ezequiel Garcia-Lembergman

Instituto de Economía - Pontificia Universidad Católica de Chile

2024

Datos de Panel

- A veces, no tenemos una variable instrumental para solucionar el sesgo de selección
- Ademas, sabemos que es dificil encontrar suficientes variables de control para no tener variables omitidas.
- Para eso, necesitamos datos de otra naturaleza que los que discutimos hasta ahora
  - Necesitamos que para un mismo grupo, tenemos varias observaciones.
- Eso nos conlleva a usar: **datos de panel**

Datos de panel

- Una base de datos de panel es una donde las unidades observadas tienen 2 dimensiones:
  - Observamos individuos/firmas/regiones/etc *i* varias veces (*j*).
- Lo más común es que *i* está observado varias veces en el tiempo (es decir j es tiempo *t*).
- Pero *j* puede ser otras cosas tambien:
- Si tenemos esto, vamos a poder eliminar cualquier sesgo que este relacionado con caracteristicas de *i* que sean constantes en *j*. Por ejemplo, si *j* es tiempo, caracteristicas de *i* que no cambien a lo largo de los anios (e.g: habilidad innata, inteligencia, genes, etc.). Es como si pudieramos controlar por todas variables que son invariantes en el tiempo, aun no teniendolas en la base de datos!!

Ejemplo datos de corte transversal

Ejemplo: consumo de productos de supermercado

Datos de corte transversal (hasta hoy):

product	year	lnconsumo	treatment
1	2018	5.806405	1
2	2018	5.977645	1
3	2018	5.599422	0
4	2018	5.4149	0
5	2018	5.257495	0

Ejemplo datos de panel

Ejemplo: consumo de productos de supermercado

Datos de panel: a cada producto *i* lo observamos en dos anios diferentes (2014, 2018).

product	year	lnconsumo	treatment
1	2014	6.318968	1
1	2018	5.806405	1
2	2014	6.308098	1
2	2018	5.977645	1
3	2014	5.609472	0
3	2018	5.599422	0
4	2014	5.42495	0
4	2018	5.4149	0
5	2014	5.298317	0
5	2018	5.257495	0

Tener este tipo de datos, nos va a permitir controlar por TODAS las caracteristicas de los productos que no varian en el tiempo.

Ejemplo: efecto de ineficiencia judicial en reducir el crimen?

- i: regiones
- t: tiempo
- y: crimen
- x: ineficiencia del sistema judicial
- Cortes transversal, solo vemos las regiones en un momento del tiempo:
  - Por lo tanto, estimamos por MCO: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Critica: sesgo por variables omitidas. Ejemplo: hay regiones que tienen mayor poblacion que otras. Estas regiones suelen tener mas ineficiencia juridica y mas crimines.
- Vamos a explotar datos de panel (ver datos de la misma region en muchos momentos del tiempo) para solucionar algunas de esas cosas

Detalles básicos

- Imaginan que tenemos un modelo con datos de panel tal que:

$$y_{it} = \alpha + X'_{it}\beta + \varepsilon_{it}, i = 1, ..., N; t = 1, ..., T \quad (1)$$

- Los índices de las variables denotan las 2 dimensiones de los datos que todos modelos con datos de panel va a tener
- Supongan que el termino de error se puede dividir en dos partes:
  - error compuesto: $\varepsilon_{it} = \mu_i + \nu_{it}$ donde μ_i representa el efecto específico inobservable correspondiente al individuo i.
  - Noten que μ_i no varia en el tiempo. Es algo especifico a i que es igual en todo t. Ejemplo: la inseguridad promedio historica del barrio.
  - ν_{it}, en cambio, es el resto del residuo
  - Tener datos de panel nos va a permitir solucionar variables omitidas del estilo μ_i.

Modelar la parte fija del error

- Podemos pensar que el componente fijo del error se puede estimar
  - Eso se llama “efectos fijos”

Efectos fijos

- Se pueden estimar los efectos fijos poniendo una variable binaria para cada una de las unidades *i*.

$$y_{it} = \alpha + x_{it}\beta + \alpha_i + v_{it},$$

- donde α_i es una dummy que toma valor 1 para el individuo *i* y 0 para cualquier otro individuo.
- Es decir, definir *N* variables dummy para cada *i*, α_i, que toma valor 1 para la region *i* y 0 para todas las demas.
- Eso implica que si tenemos $N * T$ observaciones, *k* otras variables de interés y una constante, tendremos que estimar $k + N$ parámetros.
  - Se necesita suficiente $T > 1$ para poder hacer eso
- En Stata “xtreg, fe” nos entrega la estimación
- O, simplemente, agregar a una regresion una dummy por cada *i*.

Equivalencia

- Estimar el modelo de efectos fijos agregando una dummy para cada i es equivalente a restar las medias de las varibles:

$$\text{efectos fijos: } y_{it} = \alpha + x_{it}\beta + \alpha_i + v_{it}$$

- Para cada i tomen su promedio de y en el tiempo (definiendo $\bar{y}_i = \sum_t \frac{1}{T} y_{it}, \bar{x}_i = \sum_t \frac{1}{T} x_{it}$):

$$\bar{y}_i = \alpha + \bar{x}_i\beta + \alpha_i + \bar{v}_i$$

- Se puede demostrar que incluir una dummy en el modelo de efectos fijos es equivalente a restarle a y_{it} su promedio y hacer la regresion:

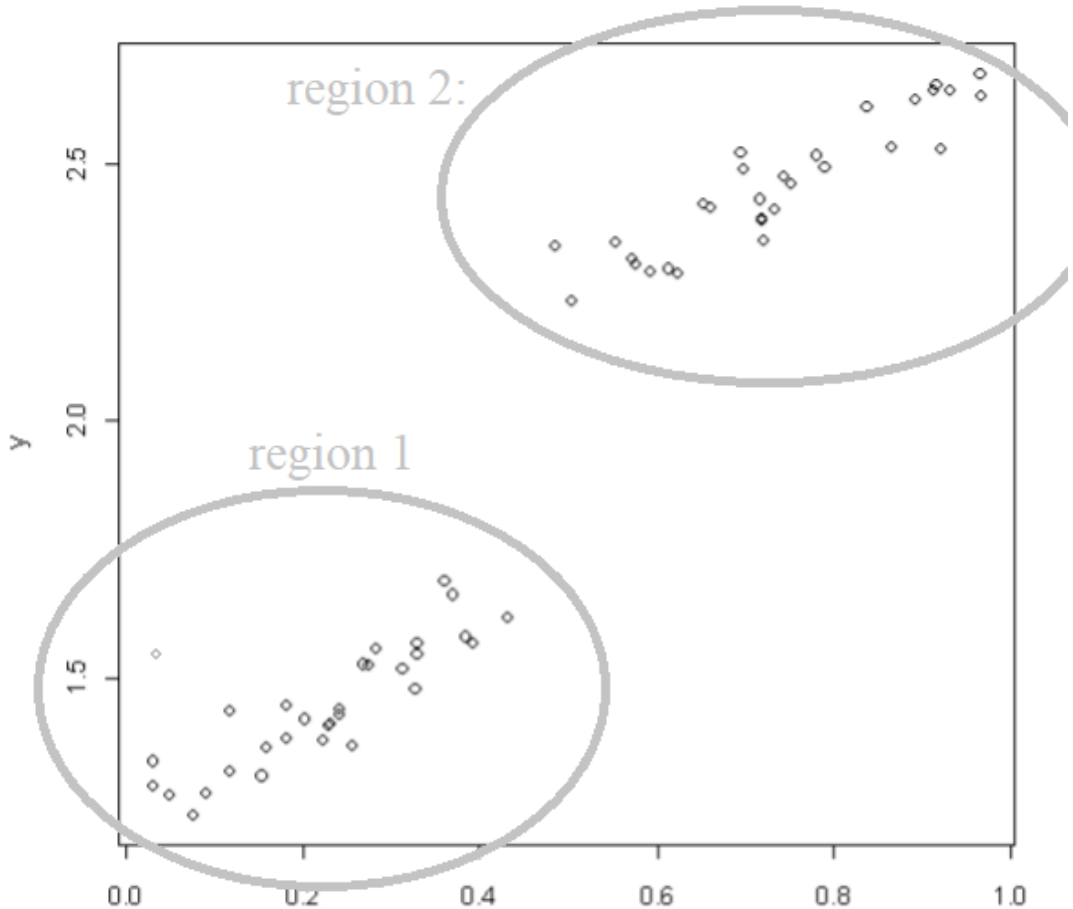
$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (\alpha_i - \alpha_i) + (v_{it} - \bar{v})$$

- Noten que el termino en rojo se va! Es decir que cuando estimamos efectos fijos, estamos eliminando el sesgo potencial por cosas de i que no cambian en el tiempo!!
- Nos sacamos de encima un monton de problemas de endogeneidad.

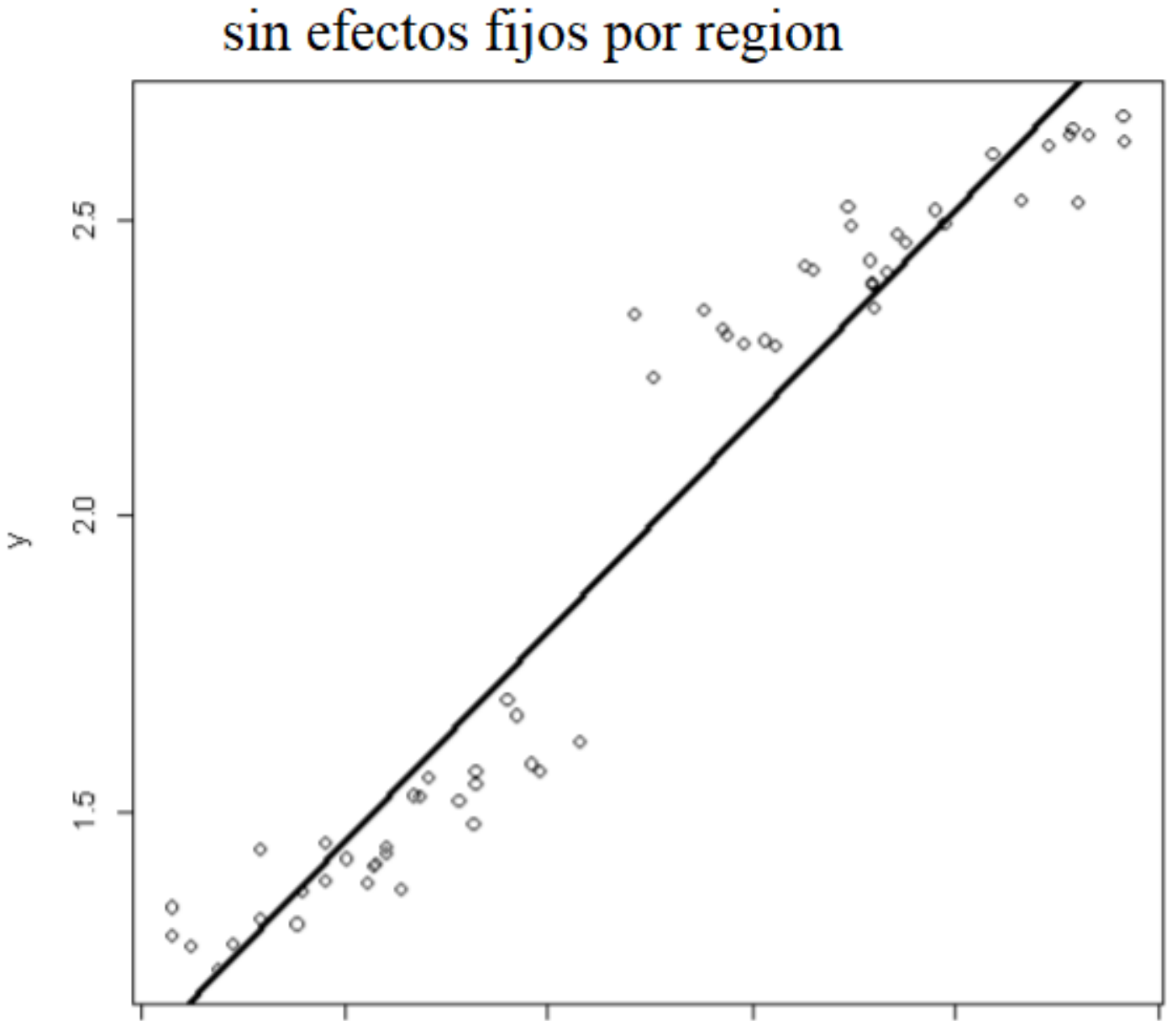
Que hace efectos fijos: ejemplo grafico

Imaginen que queremos estudiar si la ineficiencia juridica (x) de las regiones afecta el crimen (y). tenemos $i = 2$ regiones que las vemos en muchos t.

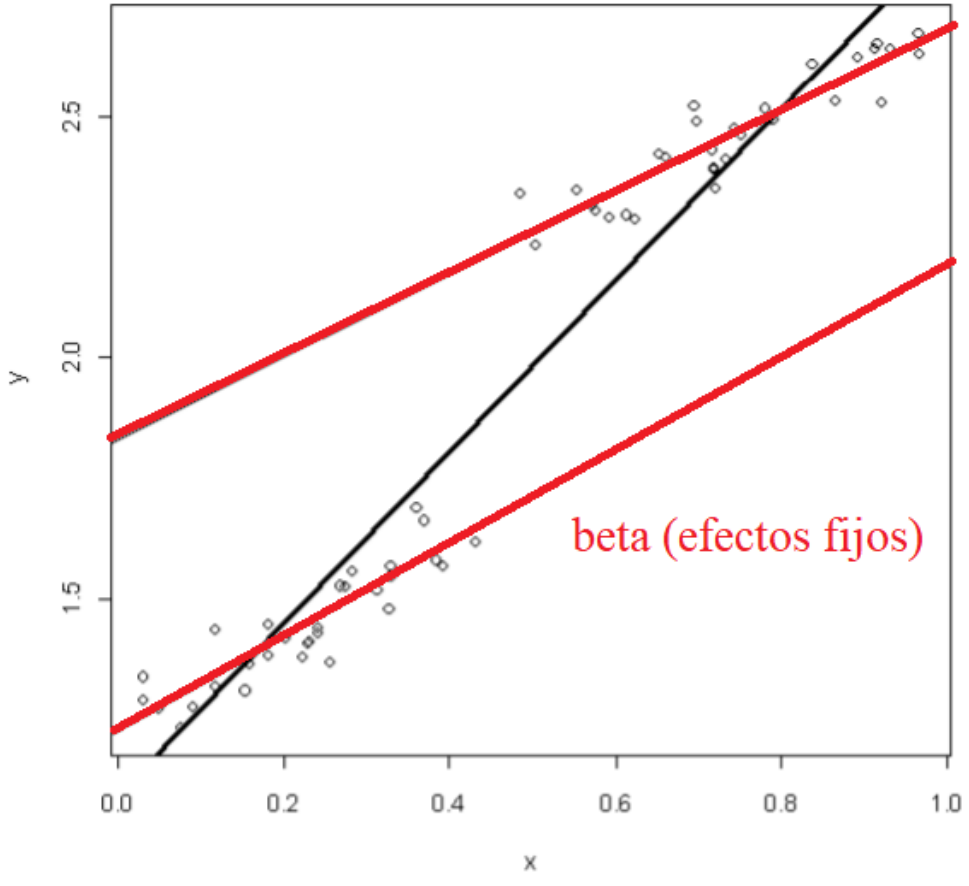
La data en grafico se ve asi:



Si estimamos el modelo sin efectos fijos



Incluyendo efectos fijos



- Con efectos fijos, controlamos por el hecho de que cada *i* tenia valores iniciales diferentes. Ahora la pendiente promedio es mas chica

Que hace el modelo de efectos fijos exactamente?

- Le pone una constante diferente a cada i.
- Eso va a absorber diferencias entre los individuos que son constantes en el tiempo.
- β_1 solo va a utilizar cambios en x_{it} con respecto a su promedio para identificar el efecto en y_{it}.
- Entonces, si el problema de endogeneidad esta dado porque regiones mas pobladas son mas ineficientes juridicamente, el efecto fijo va a abosrver estas diferencias iniciales y ya no sera un problema.

Estimación de los efectos fijos

- El modelo de efecto fijos les va a permitir estimar de manera consistente y sin sesgo los β (siempre y cuando la variable omitida sea constante en el tiempo).

Resumiendo

- Cuando contamos con datos de panel, el modelo de efectos fijos nos permite eliminar sesgo por variables omitidas que son especficas a *i* y no varian a traves de los *j*.
- Esto es genial porque eliminamos muchos potenciales problemas, aun sin tener las variables en la base de datos.
- Por ejemplo, cuando *j* es tiempo, cualquier variable omitida que no varia en el tiempo no sera un problema!
- Los problemas de endogeneidad que quedaran son los que tienen que ver con variables omitidas que cambian en el tiempo.
- La proxima clase vamos a ver el modelo de diferencias en diferencias (Muy clave. Vengan atentos!!)