Modelos de Regresión con datos de panel (Introductorio)

Expositora: Luisa Fernanda Ames Santillán

18 de Febrero de 2023

Regresión con datos de panel

- Modelos de datos de panel estático: Efectos fijos y aleatorios.

- Elección de Modelos alternativos:

Modelo de efectos individuales versus el modelo Pool
Modelo de efectos fijos versus efectos aleatorios

En el campo económico:
Los efectos de los ingresos en el ahorro, con datos de países a través de los años.
El efecto de la educación sobre los ingresos, con datos de individuos a través del tiempo.

- Los datos de panel es un concepto bidimensional, donde los mismos individuos se observan repetidamente durante diferentes períodos de tiempo. tienen dimensiones transversales y series temporales.
  Los datos del panel incluyen individuos observados en " períodos de tiempo regulares.

Los datos del panel se pueden equilibrar cuando se observan todos los individuos en todos los periodos del tiempo $T_i = T$ para todo i . Se pueden desequilibrar cuando los individuos no se observan en todos los periodos del tiempo

$$T_i \neq T$$

- Se asume correlación (agrupamiento) a lo largo del tiempo para un individuo dado, pero independencia sobre individuos. Por ejemplo El ingreso para el mismo individuo se correlacional a lo largo del tiempo pero es independiente entre individuos.

- La regresión de datos de panel es una manera de controlar las dependencias de variables independientes no observadas en una variable dependiente, lo que puede conducir a estimadores sesgados en los modelos de regresión lineal.

- En general, los datos de panel pueden verse como una combinación de datos transversales y de series de tiempo. Los datos transversales se describen como una observación de varios objetos y las variables correspondientes en un punto específico en el tiempo (es decir, se toma una observación una vez). Los datos de series de tiempo solo observan un objeto de forma recurrente a lo largo del tiempo. Los datos del panel comprenden características de ambos en un modelo mediante la recopilación de datos de múltiples objetos iguales a lo largo del tiempo.

Tipos de datos de panel:
Panel corto: muchos individuos y pocos periodos de tiempo
Panel largo: muchos periodos de tiempo y pocos individuos.
Ambos: muchos periodos de tiempo y muchos individuos.

Comparativo de Datos de panel vs otros

Datos de panel

person	year	x	y
A	2018	3,5	85
A	2019	3,2	83
A	2020	3,8	88
B	2018	1,2	79
B	2019	1,5	83
B	2020	2,3	88
C	2018	5,6	75
C	2019	6	72
C	2020	5,8	78

Secciones transversales agrupadas

year	x	y
2018	3,5	85
2019	3,2	83
2020	3,8	88
2018	1,2	79
2019	1,5	83
2020	2,3	88
2018	5,6	75
2019	6	72
2020	5,8	78

Datos transversales agrupados

¿cuál es el significado detrás de este concepto de datos y por qué deberíamos usarlo?

- La respuesta es la **heterogeneidad y endogeneidad** en los modelos de regresión lineal, en los que la heterogeneidad a menudo conduce a resultados sesgados. Los datos del panel pueden ser una alternativa resolver este problema.
- Dado que la **heterogeneidad y la endogeneidad** son cruciales para comprender por qué utilizamos modelos de datos de panel, vamos hacer una breve explicación.

El problema de la endogeneidad causado por la heterogeneidad no observada

- "La dependencia no observada de otras variables independientes se llama **heterogeneidad no observada** y la correlación entre las variables independientes y el término **de error (es decir, las variables independientes no observadas)** se llama **endogeneidad** ".

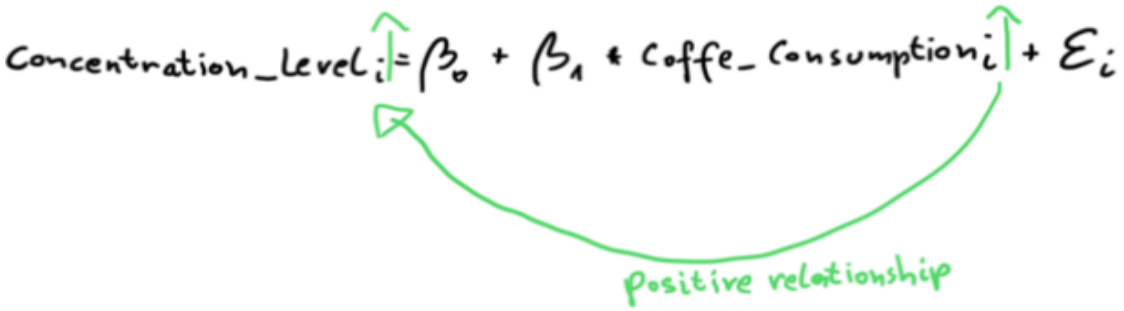
Supongamos se quiere analizar la relación de cómo el consumo de café afecta el nivel de concentración. Un modelo de regresión lineal simple se vería así:

$$Concentration\_Level_i = \beta_0 + \beta_1 * Coffe\_Consumption_i + \varepsilon_i$$

dónde:

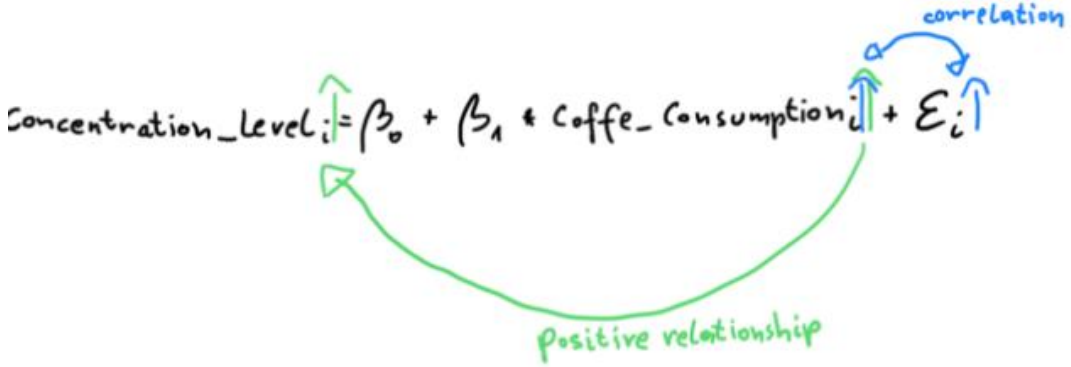
- *Concentration_Level es* la variable dependiente (DV)
- *β0* es la intersección
- *β1* es el coeficiente de regresión
- *Coffe_Consumption* es la variable independiente (IV)
- ε es el término de error

Regresión lineal simple



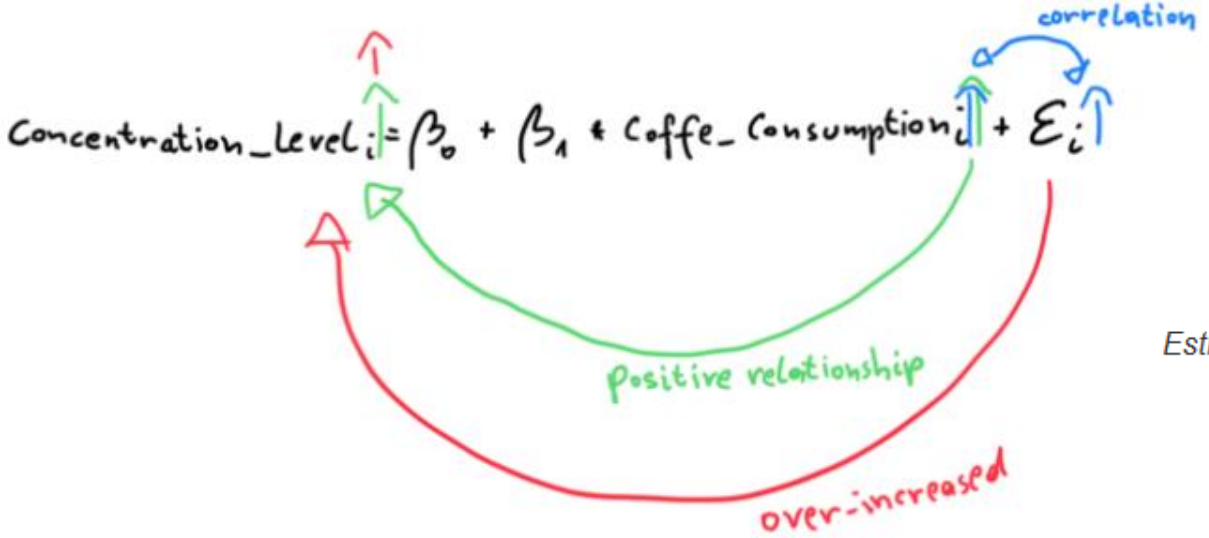
Relación entre IV y DV

Pero, ¿qué sucede si hay otra variable que afectaría a las IV existentes y no está incluida en el modelo? Por ejemplo, el *cansancio* tiene una alta probabilidad de afectar el *consumo de café* (si estás cansado, beberás café, dicha variable se denomina **variable independiente no observada**. Están **"ocultos" osea detrás del término de error** y si, por ejemplo, ***Coffe_Consumption* está relacionado positivamente con dicha variable**, el **término de error aumentaría a medida que aumenta *Coffe_Consumption* :**



Correlación entre IV y término de error

Esto, a su vez, conduciría a un estimador sobre-aumentado de la DV *Concentration_Level* . Por lo tanto, el DV estimado está sesgado y dará lugar a inferencias inexactas. En el ejemplo, el sesgo sería el aumento excesivo en rojo de *Concentration_Level.*



Estimadores sesgados debido a la heterogeneidad

Los datos panel se pueden aplicar para estos casos , la ventaja de los datos de panel es que podemos controlar la **heterogeneidad** en nuestro modelo de regresión reconociendo la heterogeneidad como **fija** o **aleatoria** .

Modelos de regresión de datos panel

Los modelos de datos de panel describen el comportamiento de los individuos tanto a lo largo del tiempo como entre individuos.

$$y_{it} = X_{it}\beta + \alpha_i + u_{it} \quad \text{for } t = 1, \dots, T \text{ and } i = 1, \dots, N$$

Hay tres **tipos de modelos**: el modelo agrupado, el modelo de efectos fijos y el modelo de efectos aleatorios.

Modelo agrupado (Pooled model) o MCO

El modelo agrupado especifica los coeficientes constantes, los supuestos habituales para el análisis de corte transversal.

Este es el modelo de datos de panel más restringido y no se usa mucho en la literatura.

Este MCO simple requiere que no haya correlación entre las variables independientes no observadas y los IV (osea exogeneidad).

$$y_{it} = X_{it}\beta + \alpha_i + u_{it} \quad \text{for } t = 1, \dots, T \text{ and } i = 1, \dots, N$$

$$\text{Cov}(X_{it}, \alpha_i) = 0$$

Supuesto de exogeneidad

El problema con PooledOLS es que incluso el supuesto anterior es cierto, alfa podría tener una correlación en serie a lo largo del tiempo. En consecuencia, PooledOLS es mayormente inapropiado para datos de panel.

$$\text{Cov}(\alpha_i, \alpha_j) = \text{Var}(\alpha_i) = \sigma_\alpha^2 > 0$$

Correlación serial entre alfa

Modelo de efectos fijos (FE)

Este permite que los efectos individuales específicos ai se correlacionen con los regresores x. Incluimos ai como intercepciones. Cada individuo tiene un término de intersección diferente y los mismos parámetros de pendiente

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\beta + u_{it}$$

Podemos recuperar los efectos específicos individuales después de la estimación como: $\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\hat{\beta}$

En otras palabras, los efectos específicos individuales son la variación sobrante en la variable dependiente que no puede ser explicada por los regresores. Se pueden incluir dummies de tiempo en los regresores x.

Dentro de los modelos FE, la relación entre las variables independientes no observadas y los IV (es decir, la endogeneidad) puede existir:

$$y_{it} = X_{it}\beta + \alpha_i + u_{it} \quad \text{for } t = 1, \dots, T \text{ and } i = 1, \dots, N$$

$\text{Cov}(X_{it}, \alpha_i) \neq 0$

Se permite endogeneidad

En este modelo si asumimos *alfa* como constante y restamos los valores medios de cada término de la ecuación, *alfa* (osea la heterogeneidad no observada) obtendrá cero y se puede despreciar

$$y_{it} - \overline{y_i} = \beta(X_{it} - \overline{X_i}) + \underbrace{(\alpha_i - \overline{\alpha_i})}_{=0} + (u_{it} - \overline{u_i})$$

Elimina los efectos individuales en el modelo FE

Únicamente, el error idiosincrásico (representado por *my* = factores no observados que cambian con el tiempo y entre unidades) permanece y tiene que ser exógeno y no colineal.

Modelo de efectos aleatorios (RE)

Determinan los efectos individuales de variables independientes no observadas como variables aleatorias a lo largo del tiempo. Son capaces de "cambiar" entre OLS y FE y, por lo tanto, pueden centrarse en ambas dependencias **entre** y **dentro de los individuos** . La idea detrás de los modelos RE es la siguiente:

$$y_{it} = X_{it}\beta + \alpha_i + u_{it} \quad \text{for } t = 1, \dots, T \text{ and } i = 1, \dots, N$$

En general, si la covarianza entre *alfa* y *IV (s)* es cero (o muy pequeña), no hay correlación entre ellos y se prefiere un modelo MCO. Si esa covarianza no es cero, hay una relación que debe eliminarse mediante el uso de un modelo FE:

$$Cov(\alpha_i, X_{it}) \neq 0 \Rightarrow \text{FE-model}$$
$$Cov(\alpha_i, X_{it}) = 0 \Rightarrow \text{OLS}$$

El problema con el uso de OLS, como se indicó anteriormente, es la correlación serial entre *alfa a lo* largo del tiempo. Por lo tanto, los modelos RE determinan qué modelo tomar de acuerdo con la correlación serial de los términos de error. Para hacerlo, el modelo usa el término *lambda.* En resumen, *lambda* calcula qué tan grande es la varianza de *alfa* . Si es cero, entonces no habrá variación de *alfa* , lo que, a su vez, significa que PooledOLS es la opción preferida. Por otro lado, si la varianza de *alfa* tiende a volverse muy grande, *lambda* tiende a convertirse en uno y, por lo tanto, podría tener sentido eliminar *alfa* e ir con el modelo FE.

$$\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T \cdot \sigma_\alpha^2} \right) \qquad \lambda = 0 \Rightarrow \text{OLS} \qquad \lambda = 1 \Rightarrow \text{FE}$$

¿Cómo decidir qué modelo es apropiado?
Elección entre PooledOLS (MCO) y FE / RE: Básicamente, existen **cinco supuestos para los modelos de regresión lineal simple que deben cumplirse**. Dos de ellos pueden ayudarnos a elegir entre PooledOLS y FE / RE.
Estos supuestos son (1) Linealidad, (2) Endogeneidad, (3a) Homoscedasticidad y (3b) No autocorrelación, (4) Las variables independientes no son estocásticas y (5) No hay multicolinealidad.
Si **se violan los supuestos (2) o (3) (o ambos), entonces FE o RE podrían ser más adecuados.**

Prueba de Multiplicador Breusch Pagan Lagrange:

Esta es una prueba para el modelo de efectos aleatorios basado en el residuo OLS (MCO).
Probar si σ_u^2 o equivalentemente $cor(u_{it}, u_{is})$ es significativamente diferente de cero. Si la prueba LM es significativa, utilice el modelo de **efectos aleatorios** en lugar del modelo OLS (MCO).

La Elección entre FE y RE: Se debe responder si la heterogeneidad individual no observada es un efecto constante o aleatorio. Pero esta pregunta también se puede responder mediante la prueba de Hausman.

Prueba de Hausman:

Es una prueba de **endogeneidad**.
Al ejecutar la prueba de Hausman, **la hipótesis nula** es que la **covarianza entre IV (s) y *alfa* es cero**. Si este es el caso, entonces **se prefiere RE a FE**. Si la hipótesis nula **no es cierta**, debemos seguir el **modelo FE.**

La prueba de Hausman prueba si hay una diferencia significativa entre los estimadores de efectos fijos y aleatorios.
El estadístico de la prueba de Hausman sólo se puede calcular para los regresores variables en el tiempo.

La estadística de la prueba de Hausman es: $$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE}))(\hat{\beta}_{RE} - \hat{\beta}_{FE})$$