Universidad Torcuato Di Tella
Maestría en Economía y Econometría
2022

Datos de Panel
Problem Set 2
Modelos de Datos de Panel Lineales

---

1. Utilice nuevamente la base de datos "cornwell.dta" provista para el Problem Set 1. Considere el siguiente modelo de regresión:

$$\ln crmrte_{it} = \beta_0 + \beta_1 \ln prbarr_{it} + \beta_2 \ln prbconv_{it} + \beta_3 \ln prbpris_{it} + \beta_4 \ln avgsen_{it} + \beta_5 \ln polpc_{it} + \sum_{\tau=82}^{87} \beta_\tau \cdot I\{t = \tau\} + \mu_i + \varepsilon_{it}$$

   a) Utilizando el comando egen de STATA, construya las medias individuales de las variables del modelo.

   b) Aplique la transformación within al modelo. Luego, estime el modelo transformado por POLS.

   c) Comente sobre la validez de los errores estándar del inciso previo.

> **Solution:** La estimación de la varianza de los errores en el inciso previo es
>
> $$\widehat{\sigma}_\varepsilon^2 = RSS/(NT - K).$$
>
> Sin embargo, sabemos que una estimación consistente para σ_ε^2 es
>
> $$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1) - K} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\varepsilon}_{it}^2.$$
>
> Por lo tanto, los errores estándar reportados tienden a ser pequeños comparados a los verdaderos. El problema se encuentra en que los grados de libertad de aplicar OLS al modelo transformado no coinciden con el denominador del estimador consistente para σ_ε^2. Por consiguiente, salvo que T sea lo suficientemente grande, necesitamos corregir este denominador.

   d) Utilice el comando *xtreg* para estimar nuevamente el modelo usando efectos fijos.

> **Solution:** Estimamos utilizando los comandos para datos de panel, en particular, la opción para la estimación por efectos fijos.
>
> ```
> xtset county year
> xtreg lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87, fe
> ```

   e) Estime el modelo usando diferencias finitas de primer orden.

> **Solution:** Al no haber un comando nativo del tipo "fd y x" necesitamos construir las variables. Esto puede hacerse creando las variables

```
gen dlcrmrte = lcrmrte-L1.lcrmrte
for any $xlist : gen dX= X-L1.X
reg dlcrmrte dlprbarr dlprbconv dlprbpris dlavgsen dlpolpc dd82-dd87, nocons
```

Solution: Las estimaciones de este ejercicio se presentan en la siguiente tabla:

	(1) POLS	(2) FE	(3) FD
lprbarr	-0.720***	-0.360***	-0.327***
	(0.110)	(0.0324)	(0.0300)
lprbconv	-0.546***	-0.286***	-0.238***
	(0.0704)	(0.0212)	(0.0182)
lprbpris	0.248**	-0.183***	-0.165***
	(0.109)	(0.0325)	(0.0260)
lavgsen	-0.0868	-0.00449	-0.0218
	(0.113)	(0.0264)	(0.0221)
lpolpc	0.366***	0.424***	0.398***
	(0.121)	(0.0264)	(0.0269)
d82	0.00514	0.0126	0.00771
	(0.0367)	(0.0215)	(0.0171)
d83	-0.0435	-0.0793***	-0.0844***
	(0.0336)	(0.0213)	(0.0235)
d84	-0.109***	-0.118***	-0.125***
	(0.0392)	(0.0216)	(0.0287)
d85	-0.0780**	-0.112***	-0.122***
	(0.0386)	(0.0218)	(0.0331)
d86	-0.0421	-0.0818***	-0.0863**
	(0.0429)	(0.0214)	(0.0367)
d87	-0.0270	-0.0405*	-0.0378
	(0.0381)	(0.0210)	(0.0400)
_cons	-2.082**	-1.604***	
	(0.865)	(0.169)	
N	630	630	540

Standard errors in parentheses

* $p < 0{,}10$, ** $p < 0{,}05$, *** $p < 0{,}01$

2. Utilice la base de datos provista *"murder.dta"*. La base de datos es una muestra longitudinal de estados de EE.UU., para los años 1987, 1990 y 1993.

   a) Estime por OLS el efecto de las ejecuciones (x) sobre la tasa de homicidios (*murder rates*, m) controlando por desempleo (u) y año:

$$m_{i,t} = \alpha + \beta_x x_{i,t} + \beta_u u_{i,t} + \beta_{90} d_{90,t} + \beta_{93} d_{93,t} + \nu_{i,t}$$

Note que se omitió la dummy temporal para el año 1987. Interprete los resultados.

Solution: A continuación, se muestran los resultados de la estimación por OLS de la ecuación anterior

	Tasa de Homicidios
Ejecuciones	0.163
	(0.84)
Desempleo	1.391**
	(3.08)
Dummy año 1990	2.675
	(1.47)
Dummy año 1993	1.607
	(0.91)
Constante	-1.864
	(-0.61)
R^2	0.08
N	153

* $p < 0.05$; ** $p < 0.01$; t-statistics en paréntesis

En la tabla anterior se puede observar que el coeficiente de la cantidad de ejecuciones ($\hat{\beta}_x$) no es estadísticamente significativo para explicar la cantidad de asesinatos. Si la ecuación anterior estuviera bien estimada (es decir, si todos los supuestos necesarios para que OLS sea consistente se cumpliesen), los resultados indicarían evidencia en contra de la pena de muerte como forma de prevenir los homicidios. En otras palabras, una mayor cantidad de ejecuciones no tiene efecto disuasorio sobre posibles homicidas, ya que estadisticamente la variable no resulta significativa en la regresión corrida. Es más, no solo no resulta siginificativa, sino que el signo es el opuesto al que, a priori, los defensores de la pena de muerte afirman que debería ser.

Dentro de las variables incluidas como controles, la única que resulta significativa es el desempleo correspondiente a cada estado ($\hat{\beta}_u$). De acuerdo a los resultados de la estimación, la tasa de desempleo se relaciona de forma positiva con la tasa de homicidios, es decir, $\hat{\beta}_u > 0$, lo cual estaría indicando evidencia a favor de que el malestar socioeconómico (tomando como posible *proxy* la tasa de desempleo) debiera ser un factor a tener en cuenta al momento de analizar temas relacionados con la inseguridad y la criminalidad.

Por otra parte, podemos observar que ninguna de las dummies temporales ($\hat{\beta}_{90}$ y $\hat{\beta}_{93}$) resultan estadísticamente siginificativas, lo cual indica evidencia a favor de la ausencia de efectos fijos temporales.

b) ¿Por qué podría ser importante tener en consideración los efectos temporales agregados en el modelo?

Solution: Podría ser importante si la tasa de homicidios es afectada por factores macroeconómicos externos que afectan a todos los estados de EE.UU. de la misma manera. Por lo tanto, si no incluimos estas variables, debemos suponer que cualquier cambio en la media de la tasa de homicidios en el tiempo se debe a las ejecuciones o a la tasa de desempleo y no a factores externos. Por otra parte, controlar por estas variables hace más factible que se cumpla el supuesto de ausencia de autocorrelación serial.

c) Ahora, considere la siguiente modificación en el modelo:

$$m_{i,t} = \alpha + \beta_x x_{i,t} + \beta_u u_{i,t} + \beta_{90} d_{90,t} + \beta_{93} d_{93,t} + c_i + e_{i,t}$$

donde c_i es un efecto individual por estado. Estime la ecuación usando efectos fijos.

d) Repita la estimación del inciso previo usando diferencias finitas de primer orden.

Solution: Para hallar los estimadores de Efectos Fijos, estimamos la siguiente ecuación:

$$m_{i,t} - \bar{m}_i = \beta_x(x_{i,t} - \bar{x}_i) + \beta_u(u_{i,t} - \bar{u}_i) + \beta_{90}(d_{90} - \bar{d}_{90,i}) + \beta_{93}(d_{93} - \bar{d}_{93,i}) + (e_{i,t} - \bar{e}_i)$$

donde $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$.

Mientras que para obtener los estimadores de Primeras Diferencias, estimamos la siguiente ecuación:

$$\Delta m_{i,t} = \beta_x \Delta x_{i,t} + \beta_u \Delta u_{i,t} + \beta_\Delta d_{90} + \beta_{93}\Delta d_{93} + \Delta e_{i,t}$$

donde Δ indica que se han aplicado primeras diferencias a los datos correspondientes.

A continuación, se muestra una tabla con los resultados de las estimaciones de las ecuaciones anteriores:

	LSDV	FE	FD
Ejecuciones	-0.138	-0.138	
	(-0.78)	(-0.78)	
Desempleo	0.221	0.221	
	(0.75)	(0.75)	
Dummy año 1990	1.556*	1.556*	
	(2.09)	(2.09)	
Dummy año 1993	1.733*	1.733*	
	(2.47)	(2.47)	
Δ Ejecuciones			-0.115
			(-0.78)
Δ Desempleo			0.163
			(0.53)
Δ Dummy año 1990			1.511*
			(2.29)
Δ Dummy año 1993			1.725*
			(2.02)
Constante	5.904	5.822**	
	(1.78)	(3.04)	
R^2	0.91	0.07	0.06
N	153	153	102

* $p < 0.05$; ** $p < 0.01$; t-statistics en paréntesis

En la tabla reportamos tres estimaciones. Las dos primeras columnas son por efectos fijos: una de ellas utilizando el comando especifico que provee Stata y otra utilizando un set de dummies para cada individuo. Observamos que los coeficientes en ambas estimaciones (a excepción de la constante) son iguales. La tercera columna es la estimación de primeras diferencias, donde el coeficiente relevante viene dado por el de la variable Δ Ejecuciones, es decir, la primera diferencia de la variable ejecuciones.

En la estimación del modelo de Efectos Fijos se observa que el coeficiente $\hat{\beta}_x$ no es significativo, es decir que las ejecuciones no tienen impacto sobre la tasa de homicidios, al igual que en el resultado hallando en el inciso anterior. Sin embargo, el signo del coeficiente es opuesto al signo del coeficiente del modelo OLS, siendo ahora negativo, en concordancia con lo que afirman los defensores de la pena capital, aunque sin ser significativo.

Por otra parte, ahora tenemos que el coeficiente que acompaña a la tasa de desempleo no es significativo, por lo que en la estimación de este modelo la tasa de desempleo no tiene impacto sobre la tasa de homicidios. Adicionalmente, las dummies para el año 1990 y el año 1993 en este caso dan significativos, lo cual no ocurría en el inciso anterior.

e) Brinde un ejemplo bajo el cual la variable de ejecuciones no sería estrictamente exógena (condicional en c_i). ***Observación.*** Para obtener estimaciones consistentes, el modelo de efectos fijos asume exogeneidad estricta de las variables explicativas condicionadas en c_i.

Solution: La variable explicativa de cantidad de ejecuciones (x_{it}) podría fallar en cuanto a la exogeneidad estricta si los estados aumentan las ejecuciones futuras en respuesta a los shocks positivos actuales de la tasa de homicidios. Dado el tramo de tiempo relativamente corto de la base de datos, la retroalimentación de las tasas de homicidio a las ejecuciones futuras puede no ser muy preocupante, ya que el proceso judicial en los casos de pena capital tiende a moverse lentamente. (Por supuesto, si se acelerara debido a un aumento de las tasas de homicidio, eso podría violar la exogeneidad estricta). Con una serie temporal más larga podríamos añadir $x_{i,t+1}$ (e incluso valores de un futuro más lejano) y estimar la ecuación por FE, comprobando la significación estadística de la variable $x_{i,t+1}$. En el caso de que se encuentre que esta variable es estadísticamente significativa tendríamos evidencia en favor de que no se cumple el supuesto de exogeneidad estricta.

f) Repita la estimación del inciso c) usando el estimador de GLS para diferencias finitas de primer orden. Compruebe que los coeficientes estimados son iguales a los obtenidos por FE.

Solution: Computamos en Stata:

$$\begin{aligned}\widehat{\beta}^{FDGLS} &= \left(\sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} DX_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} Dy_i\right) \\ &= \left(X'(I_N \otimes D' \left(DD'\right)^{-1} D)X\right)^{-1} \left(X'(I_N \otimes D' \left(DD'\right)^{-1} D)y\right)\end{aligned}$$

g) Reestimar el modelo del inciso c) usando efectos aleatorios. Implementar el test de Hausman. ¿Cuál es el mejor estimador?

Solution: A continuación se presentan los resultados obtenidos de Stata de la estimación del modelo por Efectos Aleatorios:

	RE
Ejecuciones	-0.0543
	(-0.34)
Desempleo	0.395
	(1.39)
Dummy año 1990	1.733**
	(2.32)
Dummy año 1993	1.700**
	(2.41)
Constante	4.635**
	(2.13)
N	153
Number of id	51

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

En la estimación del modelo de Efectos Aleatorios se vuelve a observar que las ejecuciones no tienen efecto sobre la tasa de homicidios dado que el coeficiente que acompaña a esta variable no es significativo. Adicionalmente, el signo de este coeficiente es negativo, alineado a la teoría de quienes defienden la pena de muerte, aunque, nuevamente, repetimos que el efecto sigue sin ser significativo.

Por otra parte, al igual que en el inciso anterior, se vuelve a observar que el desempleo no tiene efecto alguno sobre la tasa de homicidios, mientras que las dummies para el año 1990 y el año 1993 son significativas, lo cual es evidencia a favor de la presencia de efectos fijos temporales.

Test de Hausman

Naturalmente surge la cuestión sobre cuál es el mejor estimador. Para ello se plantea el test de Hausman. De forma resumida, el test se basa en la diferencia de las estimaciones de efectos fijos y efectos aleatorios. Bajo la hipótesis nula de $cov(X, \mu_i) = 0$ para todas las covariables y para todo i, ambos estimadores son consistentes, por lo que $\hat{\beta}_{FE} - \hat{\beta}_{RE}$ converge en probabilidad a cero. Además, como bajo H_0 RE es eficiente, la matriz de varianzas y covarianzas de esa diferencia es la resta de las matrices de varianzas y covarianzas. El estadístico en el que se basa el test es

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\hat{V}(\hat{\beta}_{FE}) - \hat{V}(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

Donde $(\hat{\beta}_{FE} - \hat{\beta}_{RE})$ es un vector columna de dimensión k, donde k es la cantidad de coeficientes estimados (no incluimos la fila correspondiente a la constante porque el test de Hausman por defecto no la usa, lo cual es logico porque en FE no está identificada. Se la puede identificar poniendo la restricción de que $\sum_i^I \mu_i = 0$ como hace Stata, pero de todas maneras no tiene sentido incluir la constante ya que es exógena por definición), $\hat{V}(\hat{\beta}_{FE})$ es la estimación de la matriz de varianzas y covarianzas del estimador de FE y $\hat{V}(\hat{\beta}_{RE})$ es la estimación de la matriz de varianzas y covarianzas del estimador de RE. La distribución asintotica del estimador es χ^2 con k grados de libertad.

Computando el test en Stata notamos en este caso que el programa nos reporta que la matriz de varianzas y covarianzas que no es positiva definida. Uno podría pensar que esta es una manifestación del problema de muestra finita que puede tener el estimador (la diferenecia de las matrices puede no ser definida positiva en muestra finita, aunque asintóticamente bajo H_0 tenga que serlo). Calculando los autovalores de la matriz, vemos que algunos son negativos, confirmando el problema de que la matriz $[\hat{V}(\hat{\beta}_{FE}) - \hat{V}(\hat{\beta}_{RE})]$ no es positiva definida.

Ahora bien, ¿realmente tenemos realmente un problema de muestra finita? La respuesta en este caso es negativa. En el libro de Wooldridge (2010, p. 239) citado en el programa, el autor nos aclara que, en relación al test de Hausman para comparar FE y RE:

"A third caveat concerns the set of parameters that we can compare. Because the FE approach only identifies coeficients on time-varying explanatory variables, we clearly cannot compare FE and RE coeficients on time-constant variables. But there is a more subtle issue: we cannot include in our comparison coeficients on aggregate time efect, that is, variables that change only across time. (*Nota*: como las dummies temporales que incluimos). *As with the case of comparing FE and FD estimates, the problem with comparing coeficients on aggregate time efects is not one of identification; we know RE and FE both allow inclusion of a full set of time period dummies. The problem is one of singularity in the asymptotic variance matrix of the diference between* $\hat{\beta}_{FE}$ *and* $\hat{\beta}_{RE}$*. [...] To summarize, we can estimate models that include aggregate time effects, time constant variables, and regressors that change across both i and t, by RE and FE estimation. But no matter how we compute a test statistic, we can only compare the coefficients on the regressors that change across both i and t."*

Un posible curso de acción en este caso es usar el test para evaluar la diferencia solo de estas dos variables (*exec* u *unem*). Notar que esto tiene cierto sentido teórico: lo que estamos excluyendo es algo que varía solo a través del tiempo, por lo que uno intuiría que $cov(d_t, \mu_i) = 0$, ya que una de las variables solo varía en una dimensión (la temporal) y la otra solo difiere entre individuos. En otras palabras, dado un i, μ_i es constante a traves del tiempo, por lo que la covarianza de dicha variable aleatoria debería ser cero.

Para hacer este test de Hausman solo sobre dos coeficientes, lo que hacemos es tomar la matriz de varianzas original y solo quedarnos con la submatriz de 2×2 que corresponde a las varianzas y covarianzas de la diferencia de los estimadores de *exec* y *unem*. Chequeando los autovalores de esta submatriz, vemos que son ambos positivos, por lo la matriz es definida positiva. Luego computamos el estadístico de Hausman con la formula provista. Obtenemos un valor del estadístico de Hausman de 5,7757, que corresponde a un *p-value* de 0,06 (este valor lo sacamos de la tabla de una distribución χ^2 con dos grados de libertad). En este caso, no rechazamos la hipotesis nula al 5 % de confianza, pero si al 10 %.

En conclusión, el test de Hausman rechaza la hipótesis nula de no correlación entre los regresores y los μ_i a un nivel de significatividad de 10 %. Por lo tanto, por un criterio puramente estadístico, si el nivel adoptado fuese de 10 %, el unico estimador consistente sería efectos fijos, por lo que este sería el mejor.

3. Considere el siguiente modelo:

$$y_{it} = x_{it}\beta + \mu_i + \nu_{it}, \quad i = 1, 2, \ldots, N, \ t = 1, 2, \ldots, T$$

donde $x_{it} \overset{iid}{\sim} \mathcal{N}(0,1)$, $u_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right)$, $\nu_{it} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\nu^2\right)$ y $\mu_i \perp \nu_{it}$ para todo i, t. Suponga que $\beta = \sigma_\mu^2 = \sigma_\nu^2 = 1$ y $T = 10$. La idea es realizar experimentos de Monte Carlo para evaluar la eficiencia de distintos estimadores de β.

a) Caso 1: $N = 5$. Realice un experimento de Monte Carlo con 1000 simulaciones. Reporte media, desvío estándar y RMSE de la estimación de β usando: POLS, RE y FE.

b) Repita el punto anterior con $N = 10, 30, 50, 100$ y 500.

c) Comente los resultados obtenidos y su conclusión de qué estimador debiera utilizarse en la práctica.

Solution:

N	OLS			RE			FE		
	Media	SD	RMSE	Media	SD	RMSE	Media	SD	RMSE
5	0.9948	0.1986	0.1986	0.9959	0.1565	0.1565	0.9963	0.1527	0.1527
10	0.9987	0.1549	0.1549	0.9992	0.1090	0.1090	0.9992	0.1078	0.1078
30	1.0009	0.0828	0.0828	1.0014	0.0617	0.0617	1.0015	0.0619	0.0619
50	1.0022	0.0625	0.0625	1.0019	0.0464	0.0464	1.0019	0.0467	0.0467
100	0.9985	0.0438	0.0438	0.9999	0.0331	0.0331	1.0000	0.0333	0.0333
500	0.9996	0.0446	0.0446	0.9988	0.0332	0.0332	0.9987	0.0334	0.0334

Cuadro 1: Resultados de las simulaciones

En primer lugar, es importante destacar que dados los supuestos del modelo, los tres estimadores en consideración son consistentes. Por lo tanto, deberíamos esperar que a medida que el tamaño muestral aumente, la media de las estimaciones de β con los diferentes estimadores estén cerca del valor 1. Ahora bien, para $N < 10$ el estimador FE es el que mejor funciona en términos de sesgo y de eficiencia. Luego, a partir de un tamaño de muestra de $N = 30$ ya se observa como el estimador RE es el más eficiente de todos, es decir, es el que presenta un menor desvío estándar, lo cual se vincula a que, dados los supuestos del modelo, es el estimador con la menor varianza asintótica. En resumen, si en la práctica trabajaramos con un modelo donde suponemos que se cumplen los supuestos del modelo del inciso, entonces, para N muy pequeños uno optaría por utilizar el estimador FE, mientras que ya a partir de $N = 30$ uno optaría por el estimador de RE por su eficiencia.

4. *Basado en el Ejercicio 10.18 de Wooldridge (2010).* Utilice la base de datos *wagepan.dta* para responder las preguntas a continuación.

   a) Utilizando *lwage* como variable dependiente, estimar un modelo que contenga un intercepto y las variables *dummy* de año *d*81 a *d*87. Estime el modelo por POLS, RE, FE y FD. ¿Qué puede concluir acerca de los coeficientes de las variables *dummy*?

   Solution: Se puede apreciar que las estimaciones de los coeficientes son numéricamente idénticas.

   b) Añada las variables constantes en el tiempo *educ*, *black* e *hisp* al modelo, y estímelo por POLS y RE. ¿Cómo se comparan los coeficientes? ¿Qué ocurre si se estima la ecuación por FE?

   Solution: Las estimaciones de POLS y RE son numéricamente idénticas. Este es un resultado general: si el modelo incluye sólo efectos temporales agregados y covariables específicas del individuo que no tienen variación temporal, entonces, POLS = RE.

   Por otra parte, cuando se utiliza FE, por supuesto, no se pueden estimar los coeficientes asociados a las variables constantes en el tiempo. Las estimaciones de las variables *dummy* de año son las mismas que las de POLS y RE. Sin embargo, cuando POLS y RE incluyen variables constantes en el tiempo, la estimación de la "constante"de FE no es igual a la estimación del intercepto en POLS/RE.

   c) ¿Son iguales los errores estándar de POLS y RE del inciso b)? ¿Cuáles son probablemente más fiables?

   Solution: Los errores estándar reportados para POLS y RE no son los mismos. Los errores estándar de POLS suponen, además de homocedasticidad, que no hay correlación serial en el error compuesto, es decir, que no considera la posible presencia de una heterogeneidad no observada. Al menos, los errores estándar de RE permiten en su estructura estándar la presencia de correlación serial, en particular, la cual es igual para todos los pares de períodos (t, s). Esto puede ser demasiado restrictivo, pero es menos restrictivo que los habituales errores estándar OLS.

   d) Obtenga los errores estándar robustos para POLS. ¿Prefiere estos o los errores estándar habituales de RE?

   Solution: Estos errores estándar robustos permiten cualquier tipo de correlación serial y de heterocedasticidad de los disturbios que varíen en el tiempo. Preferimos estos a los errores estándar habituales de RE ya que estos últimos imponen un tipo especial de correlación serial, y, además, asumen homocedasticidad.

   e) Obtenga los errores estándar robustos de RE. ¿Cómo se comparan con los errores estándar robustos de POLS, y por qué?

   Solution: Son numéricamente idénticos a los errores estándar robustos de POLS porque tenemos un solo estimador (POLS = RE en esta configuración) y, por lo tanto, hay una sóla varianza robusta.