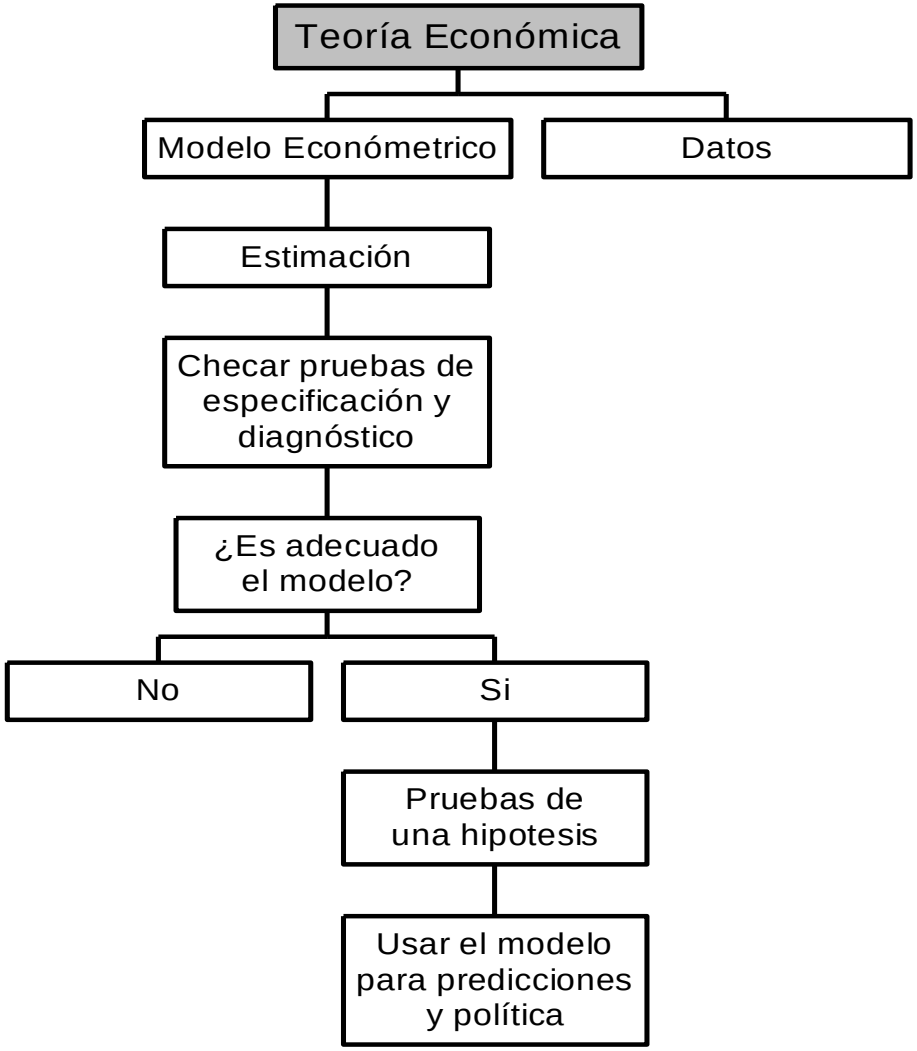
Econometría I
Taller 8. Modelos de datos panel

Esp. Humberto Acevedo

Ciudad Universitaria a 13 mayo de 2021.

PASOS PARA LA FORMULACIÓN DE UN MODELO ECONOMÉTRICO



Tipos de Datos:

Datos de sección cruzada o corte transversal

- En una sección cruzada cada observación representa a un individuo, o familia, o empresa, o estado, o país con diversas características.

- Cada observación es un individuo (empresa consumidores) con información en un punto en el tiempo

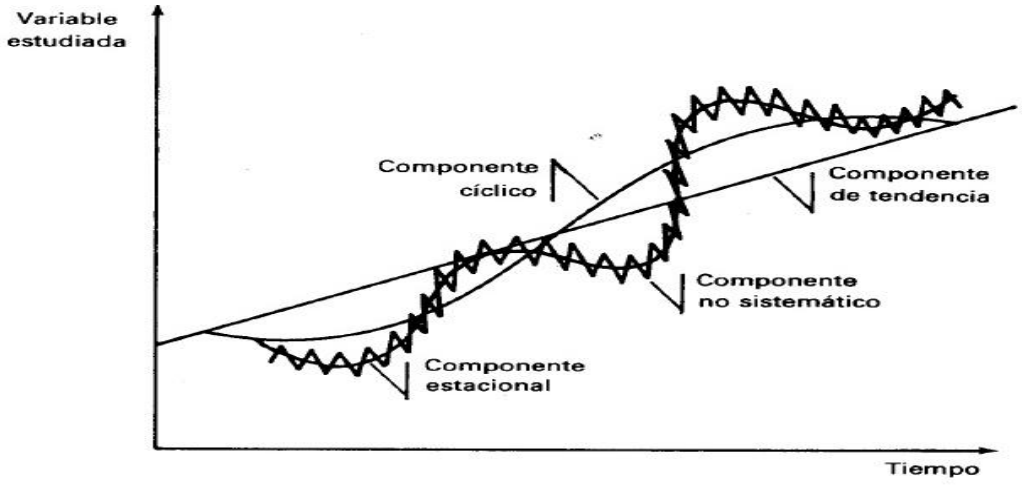
- Si la muestra no se obtuvo de manera aleatoria podemos tener problemas de sesgo

	A	B	C	D	E	F	G	H	I	J	K	L
1	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacemen	gear_ratio	foreign
2	AMC Concord	4099	22	3	2.5	11	2930	186	40	121	3.58	Domestic
3	AMC Pacer	4749	17	3	3	11	3350	173	40	258	2.53	Domestic
4	AMC Spirit	3799	22		3	12	2640	168	35	121	3.08	Domestic
5	Buick Century	4816	20	3	4.5	16	3250	196	40	196	2.93	Domestic
6	Buick Electra	7827	15	4	4	20	4080	222	43	350	2.41	Domestic
7	Buick LeSabre	5788	18	3	4	21	3670	218	43	231	2.73	Domestic
8	Buick Opel	4453	26		3	10	2230	170	34	304	2.87	Domestic
9	Buick Regal	5189	20	3	2	16	3280	200	42	196	2.93	Domestic
10	Buick Riviera	10372	16	3	3.5	17	3880	207	43	231	2.93	Domestic
11	Buick Skylark	4082	19	3	3.5	13	3400	200	42	231	3.08	Domestic
12	Cad. Deville	11385	14	3	4	20	4330	221	44	425	2.28	Domestic
13	Cad. Eldorado	14500	14	2	3.5	16	3900	204	43	350	2.19	Domestic
14	Cad. Seville	15906	21	3	3	13	4290	204	45	350	2.24	Domestic
15	Chev. Chevette	3299	29	3	2.5	9	2110	163	34	231	2.93	Domestic
16	Chev. Impala	5705	16	4	4	20	3690	212	43	250	2.56	Domestic
17	Chev. Malibu	4504	22	3	3.5	17	3180	193	31	200	2.73	Domestic
18	Chev. Monte Carlo	5104	22	2	2	16	3220	200	41	200	2.73	Domestic
19	Chev. Monza	3667	24	2	2	7	2750	179	40	151	2.73	Domestic
20	Chev. Nova	3955	19	3	3.5	13	3430	197	43	250	2.56	Domestic
21	Dodge Colt	3984	30	5	2	8	2120	163	35	98	3.54	Domestic
22	Dodge Diplomat	4010	18	2	4	17	3600	206	46	318	2.47	Domestic
23	Dodge Magnum	5886	16	2	4	17	3600	206	46	318	2.47	Domestic
24	Dodge St. Regis	6342	17	2	4.5	21	3740	220	46	225	2.94	Domestic
25	Ford Fiesta	4389	28	4	1.5	9	1800	147	33	98	3.15	Domestic
26	Ford Mustang	4187	21	3	2	10	2650	179	43	140	3.08	Domestic

Series de tiempo

- Las series de tiempo tienen una observación para cada periodo de tiempo
- A diferencia de los datos de sección cruzada es necesario tomar en consideración problemas distintos.

Por ejemplo: la tendencia y la estacionalidad deben considerarse



Logaritmo natural							
Tasa de interés		Inflación			Producto		
Mes Año	TIIE 28 (tiie_28)	Tasa de inflación observada (p)	Tasa objetivo (P)	Brecha de inflación (gap_p)	IGAE (igae)	Producto potencial (pib_pot)	Brecha del producto (gap_igae)
Ene 2008	2.07	1.31	1.1	0.21	4.60	4.58	0.02
Feb 2008	2.07	1.31	1.1	0.21	4.58	4.58	0.00
Mar 2008	2.07	1.46	1.1	0.36	4.58	4.58	0.00
Abr 2008	2.08	1.53	1.1	0.43	4.62	4.58	0.03
May 2008	2.11	1.61	1.1	0.51	4.62	4.58	0.04
Jun 2008	2.15	1.67	1.1	0.57	4.62	4.58	0.04
Jul 2008	2.16	1.69	1.1	0.59	4.62	4.58	0.04
Ago 2008	2.16	1.72	1.1	0.62	4.60	4.58	0.02
Sep 2008	2.17	1.70	1.1	0.61	4.59	4.58	0.00
Oct 2008	2.17	1.76	1.1	0.66	4.63	4.58	0.05
Nov 2008	2.13	1.82	1.1	0.73	4.60	4.58	0.02
Dic 2008	2.07	1.87	1.1	0.77	4.60	4.58	0.02
Ene 2009	2.03	1.84	1.1	0.74	4.54	4.58	-0.04
Feb 2009	1.90	1.82	1.1	0.73	4.51	4.58	-0.07
Mar 2009	1.75	1.79	1.1	0.69	4.56	4.58	-0.03
Abr 2009	1.66	1.82	1.1	0.73	4.52	4.58	-0.07
May 2009	1.59	1.79	1.1	0.69	4.53	4.58	-0.05
Jun 2009	1.59	1.74	1.1	0.64	4.56	4.58	-0.03
Jul 2009	1.59	1.69	1.1	0.59	4.57	4.59	-0.02

Datos tipo panel

- Corte transversales en periodos de tiempo.

- Con datos de panel o también denominados datos longitudinales se requiere observar al mismo conjunto de unidades en al menos dos momentos del tiempo diferentes.

- Es un conjunto de datos, donde el comportamiento de las observaciones individuales se observan en el tiempo.

Macro-panel: Estados, países, etc. Bases de datos que usualmente se pueden encontrar en páginas del banco mundial, OCDE, CEPAL

Micro-panel: Individuos, hogares, empresas, etc.

Censos del INEGI , ENOE, ENIH, ENDUTIH

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Ventajas de la estimación panel

- Hsiao (2003) & Klevmarken (1989) señalan que la estimación con datos panel permite obtener los siguientes beneficios para nuestra estimación:

  - 1- **Permiten controlar la heterogeneidad individual**: permite conocer hasta cierto punto las características de la heterogeneidad

  - Los modelos de series de tiempo y sección cruzada que no controlan la heterogeneidad, corren el riesgo de obtener resultados sesgados.

- 2- Los modelos con datos panel brindan **más información de los datos, más variabilidad, menos co-linealidad entre las variables , más grados de libertad y eficiencia.**

- 3.- Los modelos panel tienen una **mayor capacidad de estudiar las dinámicas de ajuste:** Las distribuciones de sección cruzada que aparentemente son muy estables encubren una multitud de variaciones

- 4- Los modelos panel **son mejores al identificar y medir efectos que no se identifican a simple vista en los modelos de sección cruzada o series de tiempo**:

  - Si tenemos datos de sección cruzada de mujeres con una tasa de participación laboral del 50% anual, lo cual puede deberse a:

    Cada mujer tiene un 50% de probabilidades de pertenecer a la fuerza laboral en un año determinado (a)

    - 50% de las mujeres trabajan todo el tiempo y el otro 50% no lo hace (b)

    En el caso (a) existe una gran rotación laboral entre las mujeres, en la opción (b) no existe tal rotación laboral; solo un modelo de panel permitirá discriminar entre los dos casos.

- 5- Los modelos con datos panel **permiten construir y probar modelos que presentan comportamientos más complicados que los modelos simples de sección cruzada o series de tiempo**. Por ejemplo los modelos sobre la eficiencia productiva se estudian y modelan de mejor manera con modelos panel.

- 6- Los datos panel a nivel micro como los individuos, las empresas y los hogares se miden en forma más precisa que sus variables similares a nivel macro. **El sesgo resultante de agregar datos a nivel individual se reduce o elimina**.

Limitaciones de la estimación en panel.

- **1- Problemas en el diseño y recopilación de los datos**. Incluyendo problemas de cobertura; la falta de respuesta, información confiable, las frecuencias de las entrevistas, entre otros.

- **2- Distorsiones en las medidas de error**. Errores de medición debido a preguntas mal planteadas, respuestas distorsionadas (sesgo de prestigio), el efecto del entrevistador, etc.

- 3- **Problemas de selectividad**. Incluyendo **auto-selección** o sesgo de selección (calculo de los niveles salariales, sin considerar los salarios de reserva) y la atrición (desgaste), cuando los encuestados mueren o se desplazan a otra región.

- 4- **Dimensiones temporales demasiado cortas**. Las propiedades asintóticas recaen en el componente de los individuos, es un problema común en los micro-paneles.

- 5- **Dependencia a nivel de la sección cruzada**. Los macro-paneles de países o regiones con periodos temporales largos que no tienen en cuenta la dependencia de sección cruzada, puede llevar a inferencias erróneas

- La recolección de los datos de panel puede ser muy costosa

Algunas consideraciones

- Los paneles pueden ser:
  - Balanceados: individuos cada periodo
  - No balanceados: individuos faltantes en algún periodo
- Los paneles puede ser también:
  - Cortos: pocas observaciones temporales mientras $N \to \infty$
  - Largos: Pocas observaciones individuales mientras $T \to \infty$
- Es muy probable que los errores estén correlacionados debido a que la observación de un año no son del todo independientes de la de otros años
- Algunos de los coeficientes pueden variar entre individuos o el tiempo

Tipos de variables del Modelo

- **Se pueden tener 3 tipos de variables:**

- Invariantes en el tiempo Ejemplo: el género

$$x_{it} = x_i$$

- No varían entre los individuos Ejemplo: índice de precios

$$x_{it} = x_t$$

- Varían tanto con el tiempo como con los individuos

X_{it} Ejemplo: productividad laboral

Análisis de las bases de datos en panel

- **Use mus08psidextract.dta**

Base de datos panel de un estudio sobre las dinámicas salariales en Estados Unidos, desde 1976-1982

- describe

variable name	storage type	display format	value label	variable label
exp	float	%9.0g		years of full-time work experience
wks	float	%9.0g		weeks worked
occ	float	%9.0g		occupation; occ==1 if in a blue-collar occupation
ind	float	%9.0g		industry; ind==1 if working in a manufacturing industry
south	float	%9.0g		residence; south==1 if in the South area
smsa	float	%9.0g		smsa==1 if in the Standard metropolitan statistical area
ms	float	%9.0g		marital status
fem	float	%9.0g		female or male
union	float	%9.0g		if wage set be a union contract
ed	float	%9.0g		years of education
blk	float	%9.0g		black
lwage	float	%9.0g		log wage
id	float	%9.0g		
t	float	%9.0g		
tdum1	byte	%8.0g		t== 1.0000
tdum2	byte	%8.0g		t== 2.0000
tdum3	byte	%8.0g		t== 3.0000
tdum4	byte	%8.0g		t== 4.0000
tdum5	byte	%8.0g		t== 5.0000
tdum6	byte	%8.0g		t== 6.0000
tdum7	byte	%8.0g		t== 7.0000

COMANDO
sum

exp	4165	19.85378	10.96637	1	51
wks	4165	46.81152	5.129098	5	52
occ	4165	.5111645	.4999354	0	1
ind	4165	.3954382	.4890033	0	1
south	4165	.2902761	.4539442	0	1
smsa	4165	.6537815	.475821	0	1
ms	4165	.8144058	.3888256	0	1
fem	4165	.112605	.3161473	0	1
union	4165	.3639856	.4812023	0	1
ed	4165	12.84538	2.787995	4	17
blk	4165	.0722689	.2589637	0	1
lwage	4165	6.676346	.4615122	4.60517	8.537
id	4165	298	171.7821	1	595
t	4165	4	2.00024	1	7
tdum1	4165	.1428571	.3499691	0	1
tdum2	4165	.1428571	.3499691	0	1
tdum3	4165	.1428571	.3499691	0	1
tdum4	4165	.1428571	.3499691	0	1
tdum5	4165	.1428571	.3499691	0	1
tdum6	4165	.1428571	.3499691	0	1

Organización en forma de panel

- La información se puede organizar en la forma:

1. Amplia (wide) es decir, con las observaciones en el tiempo para cada individuo

2. Larga (long) Combinando la información de todos los individuos en cada año

- Para cambiar de una forma a otra se usa el comando reshape

- Para comenzar a trabajar con los datos panel en STATA es necesario identificar el número de observaciones y de años

- Declarar el identificador individual y del tiempo

- xtset id t

- panel variable: id (strongly balanced)
- time variable: t, 1 to 7
- delta: 1 unit (se incrementan las observaciones uniformemente en una unidad)

Descripción del panel

- Comando para obtener datos descriptivos del panel
- xtdescribe

```
     id:  1, 2, ..., 595                                 n =         595
      t:  1, 2, ..., 7                                   T =           7
          Delta(t) = 1 unit
          Span(t)  = 7 periods
          (id*t uniquely identifies each observation)

istribution of T_i:   min      5%      25%       50%       75%      95%      max
                        7       7        7         7         7        7        7

    Freq.   Percent     Cum.  |  Pattern
-----------------------------+---------
      595    100.00   100.00  |  1111111
-----------------------------+---------
      595    100.00           |  XXXXXXX
```

Ejemplo panel no balanceado

Distribution of T_i:	min	5%	25%	50%	75%	95%	max
	4	7	7	7	7	7	7

Freq.	Percent	Cum.	Pattern
592	99.50	99.50	1111111
1	0.17	99.66	.1..111
1	0.17	99.83	111.1.1
1	0.17	100.00	11111.1
595	100.00		XXXXXXX

- El comando xtset nos indicaba que la base de datos está fuertemente balanceada y el comando xtdescribe nos confirma tal situación.

- Indica que tenemos considerados 595 individuos que tienen cada uno de ellos 7 observaciones temporales.

- Muestra el patrón de participación de los datos temporales y de sección cruzada

Cambios en las variables

Tanto las variables dependientes como los regresores pueden variar en el tiempo y entre individuos por tanto la variación puede ser:

- **Total (overall)**

$$(\text{en torno a la media total } \overline{x} = {}^{1}/_{NT} \sum_i \sum_t x_{it})$$

$$s_O^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{x})^2$$

- La **variación en el tiempo para un solo individuo** es decir en torno a la **media individual** se llama variación **within (dentro de)**

$$\overline{x}_i = {}^{1}/_{T} \sum_t x_{it})$$

$$s^2_w = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \overline{x}_i + \overline{x})^2$$

- Las **variaciones entre individuos para una observación temporal** se conoce como variación **between (entre)**

$$(\text{variación de } \overline{x}_i \text{ en torno a } \overline{x})$$

$$s_B^2 = \frac{1}{N-1} \sum_i (\overline{x}_i - \overline{x})^2$$

- El comando xtsum calcula las tres variaciones total, within y between

- xtsum id t lwage ed exp exp2 wks south

```
. * Panel summary statistics: within and between variation
. xtsum id t lwage ed exp exp2 wks south tdum1
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
id	overall	298	171.7821	1	595	N =	4165
	between		171.906	1	595	n =	595
	within		0	298	298	T =	7
t	overall	4	2.00024	1	7	N =	4165
	between		0	4	4	n =	595
	within		2.00024	1	7	T =	7
lwage	overall	6.676346	.4615122	4.60517	8.537	N =	4165
	between		.3942387	5.3364	7.813596	n =	595
	within		.2404023	4.781808	8.621092	T =	7
ed	overall	12.84538	2.787995	4	17	N =	4165
	between		2.790006	4	17	n =	595
	within		0	12.84538	12.84538	T =	7
exp	overall	19.85378	10.96637	1	51	N =	4165
	between		10.79018	4	48	n =	595
	within		2.00024	16.85378	22.85378	T =	7
exp2	overall	514.405	496.9962	1	2601	N =	4165
	between		489.0495	20	2308	n =	595
	within		90.44581	231.405	807.405	T =	7
wks	overall	46.81152	5.129098	5	52	N =	4165
	between		3.284016	31.57143	51.57143	n =	595
	within		3.941881	12.2401	63.66867	T =	7
south	overall	.2902761	.4539442	0	1	N =	4165
	between		.4489462	0	1	n =	595
	within		.0693042	-.5668667	1.147419	T =	7
tdum1	overall	.1428571	.3499691	0	1	N =	4165
	between		0	.1428571	.1428571	n =	595
	within		.3499691	0	1	T =	7

Lectura de las variaciones

- Los regresores que no varían en el tiempo tienen variación within cero.
- Por ejemplo el identificador del individuo y la variable ed tienen variación within cero.
- Los regresores invariantes entre los individuos tienen una variación between cero.
- Por ejemplo el identificador temporal t y la variable dummie temporal tdum1 son invariantes entre los individuos
- Por ejemplo la variable wks tiene más variación within que between, una estimación within provocaría mayor eficiencia.

- El comando xttab da información adicional sobre las variaciones between, within y total.

```
. * Panel tabulation for a variable
. xttab south
```

south	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	2956	70.97	428	71.93	98.66
1	1209	29.03	182	30.59	94.90
Total	4165	100.00	610	102.52	97.54
			(n = 595)		

- Indica que el 71% de los 4,165 observaciones tienen valor south=0, y 29% tenia south=1

- El resumen between señala que el 72% de las personas tienen valor south=0 al menos en una ocasión, y el 31% tiene valor south=1 al menos en una ocasión. El porcentaje between total es 102.52%, debido a que un 2.52% de los individuos de la muestra (15 personas) vivieron parte de su tiempo en el sur y fuera del sur por lo tanto se tiene un doble registro.

- El resumen within indica que el 95% de la gente que siempre vivió en el sur siempre se mantuvieron en el sur, y el 99% que residía fuera del sur siempre se mantuvo fuera.

- Las frecuencias se pueden calcular con el comando **xttrans south, freq**

```
. * Transition probabilities for a variable
. xttrans south, freq
```

residence; south==1 if in the South area	residence; south==1 if in the South area 0	1	Total
0	2,527	8	2,535
	99.68	0.32	100.00
1	8	1,027	1,035
	0.77	99.23	100.00
Total	2,535	1,035	3,570
	71.01	28.99	100.00

- Se ha perdido un periodo temporal para calcular las transiciones, por lo que solo se emplean 3570 observaciones.

- Para los datos invariantes en el tiempo, la entradas diagonales deben sumar un 100% y los elementos fuera de la diagonal deben sumar un 0%.

- Para el sur, el 99.2% de las observaciones siempre estuvieron en el sur, si en un periodo vivían en el sur en el siguiente se mantuvieron en el sur.

- Para aquellos que no vivían en el sur por un periodo, el 99.7% se mantuvieron fuera del sur para el siguiente periodo.

DIAGRAMA DE DISPERSION PANEL

- Si tenemos un regresor clave, podemos realizar un diagrama de dispersión de la variable dependiente sobre éste.
- graph twoway (scatter lwage exp) (qfit lwage exp) (lowess lwage exp)
- Cada punto en la gráfica representa la observación de un individuo para el año específico.
- La línea roja es un OLS de lwage con término cuadrático (qfit), la línea verde esta ajustada por una regresión no-paramétrica (lowess).
- log wage se incrementa hasta alcanzar los 30 años de experiencia, declina para años posteriores.

El modelo general

- Un modelo general para datos en panel permite que el intercepto y las pendientes de los coeficiente varíen tanto para cada individuo como en el tiempo. Es decir:

$$y_{it} = \alpha_{it} + \mathbf{x}'_{it}\beta_{it} + u_{it},$$

Donde $i = 1, \ldots, N, \quad t = 1, \ldots, T,$

- Una regresión con datos panel difiere a las regresiones de series de tiempo y sección cruzada por que se considera un subíndice doble

- Donde i, hace referencia a los individuos, hogares, empresas, países, etc. y t denota al tiempo.

- Por tanto se tiene en cuenta la dimensión temporal t y de sección cruzada i

- La mayoría de las estimaciones con datos panel consideran un componente de error como:

$$u_{it} = \mu_i + v_{it}$$

- Donde μ_i denota el efecto individual específico y v_{it} es lo que resta del disturbio.

- Por ejemplo en un modelo de salarios, y_{it} mediría los ingresos del jefe de hogar, mientras que X_{it} puede contener un conjunto de variables como la experiencia, la educación, pertenencia a sindicatos, sexo, raza, etc.

- Si consideramos que μi es invariante en el tiempo y considera cualquier efecto individual específico que no están incluidos en la regresión. Pueden considerarse como las habilidades específicas individuales no observables.

- El remanente v_{it} varia con los individuos y el tiempo, puede considerarse como el residuo usual en la regresión.

- Pero este modelo es muy general y no es estimable debido a que hay más parámetros a estimar que observaciones.

- Por tanto es necesario establecer restricciones a las variaciones de $\boldsymbol{\alpha}_{it}$ y $\boldsymbol{\beta}_{it}$ tanto para los individuos como en el tiempo y con respecto al comportamiento del error u_{it} .

El modelo agrupado (pooled)

- Es el modelo más restrictivo debido a que especifica coeficientes constantes

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

- Si este modelo esta correctamente especificado y los regresores no están correlacionados con el término de error, entonces el modelo se puede estimar usando MCO agrupados

- Realicemos una estimación con mínimos cuadrados ordinarios agrupados (pooled) para log wage usando datos para todos los individuos en todos los años.

- Incorporamos como regresores a la educación, las semanas trabajadas y la experiencia al cuadrado.

- La regresión de y_{it} sobre x_{it} brindan una estimación consistente de β si el error compuesto u_{it} de este modelo no se correlaciona con x_{it}, sin embargo es posible que u_{it} se correlacione en el tiempo para un individuo determinado

Regresión agrupada

- regress lwage exp exp2 wks ed

```
. * Pooled OLS with incorrect default standard errors
. regress lwage exp exp2 wks ed
```

Source	SS	df	MS
Model	251.491445	4	62.8728613
Residual	635.413457	4160	.152743619
Total	886.904902	4164	.212993492

Number of obs	= 4165
F(4, 4160)	= 411.62
Prob > F	= 0.0000
R-squared	= 0.2836
Adj R-squared	= 0.2829
Root MSE	= .39082

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	.044675	.0023929	18.67	0.000	.0399838	.0493663
exp2	-.0007156	.0000528	-13.56	0.000	-.0008191	-.0006121
wks	.005827	.0011827	4.93	0.000	.0035084	.0081456
ed	.0760407	.0022266	34.15	0.000	.0716754	.080406
_cons	4.907961	.0673297	72.89	0.000	4.775959	5.039963

- Los resultados muestran una R2=0.28, y las estimaciones implican que los salarios se incrementan un 0.6% con cada semana adicional trabajada.

- Los salarios se incrementan en 7.6% con cada año de educación adicional.

- Los salarios se incrementan a medida que la experiencia crece hasta alcanzar el pico de los 31 años (=0.0447/(2x0.00072)) y a partir de allí decrecen.

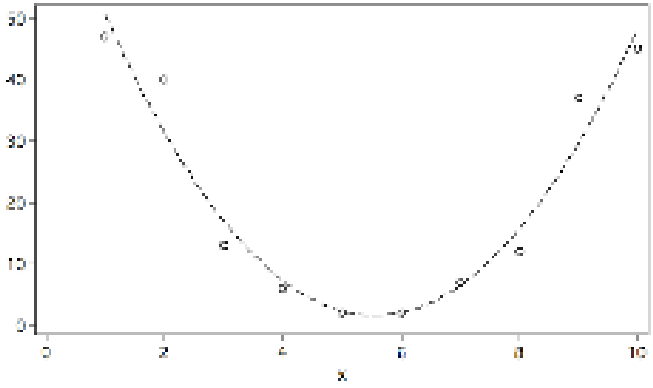
¿Cómo interpretar los términos cuadráticos?

- Si consideramos un modelo: $y=\alpha+\beta_1 X_1+ \beta_2 X_2^2$

- Si conseguimos de dicha estimación los siguientes coeficientes

Model A	Coef.
x	-.1839751
x2	1.016747
constant	.2076584

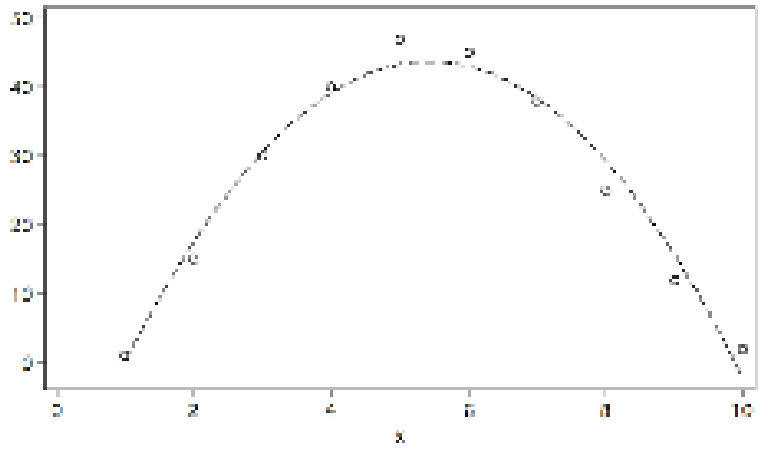
- Posiblemente veríamos una relación no lineal como esta:



- Si conseguimos unos coeficientes como los siguientes:

Model D	Coef.
x	.1839751
x2	-1.016747
constant	99.79234

- Veríamos una relación no lineal como la siguiente



¿Cómo hallar el punto de inflexión?

- El término cuadrático x_2^2, indica en qué dirección se va a inclinar la curva, sin embargo ¿qué implica el término lineal?

- Simplemente tendríamos que hallar la solución del modelo lineal.
- $y=b_0+b_1*X+b_2X_2^2$
- Si derivamos con respecto a X, tenemos:
- $Y'=b_1+2*b_2X_2$

- Esto implica que b1 da la tasa de cambio cuando x es igual cero y b2 señala cual es la dirección y la forma de la curvatura.

- En nuestro ejemplo para poder obtener el punto de inflexión solo es necesario hallar la solución para X, si tenemos la siguiente ecuación:
- $\hat{y}=4.907+0.44675exp-0.00072exp^2+0.0058wks+0.7604ed+\varepsilon$

- Si derivamos respecto a exp tenemos:
- $\hat{y}'=0.4467-(2*(0.00072)exp)$
- Si resolvemos la ecuación para exp, igualando a cero tenemos:

- $exp= 0.4467/2*0.00072=31$

- Sin embargo es muy probable que v_{it} en el modelo pool, tenga algún grado de correlación en el tiempo para un individuo, en el modelo anterior los errores estándar asumen errores i.i.d, **no es correcto asumir esto**; por tanto es necesario emplear errores estándar robustos por grupos de individuos.

regress lwage exp exp2 wks ed, vce(cluster id)

Linear regression

Number of obs	= 4165
F(4, 594)	= 72.58
Prob > F	= 0.0000
R-squared	= 0.2836
Root MSE	= .39082

(Std. Err. adjusted for 595 clusters in id)

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	.044675	.0054385	8.21	0.000	.0339941	.055356
exp2	-.0007156	.0001285	-5.57	0.000	-.0009679	-.0004633
wks	.005827	.0019284	3.02	0.003	.0020396	.0096144
ed	.0760407	.0052122	14.59	0.000	.0658042	.0862772
_cons	4.907961	.1399887	35.06	0.000	4.633028	5.182894

- Como se puede ver los errores estándar son demasiado pequeños las t muy grandes
- Cluster-robust standard errors son casi el doble.
- Cluster-robust t-statistics son la mitad.

- Si no se controla correctamente esta correlación en los errores puede ocasionar una subestimación de los errores estándar, debido a que las observaciones adicionales para una persona determinada brindan menos información nueva que sea independiente.

Population-Averaged estimator (PA)

- Los modelos agrupados (pooled) simplemente realizan la regresión de y_{it} sobre un intercepto y x_{it}, usando tanto variaciones between como within de los datos.

- Es necesario ajustar los errores estándar para cualquier correlación entre los errores temporales de un individuo determinado; permitiendo estimar modelos más eficientes.

- Los modelos agrupados de población agrupada (population average PA) serian más adecuados.

- Consideremos el modelo

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + (\alpha_i - \alpha + \varepsilon_{it})$$

- El modelo tiene el mismo intercepto pero variaciones individuales que se reflejan en los residuos

- El modelo agrupado estimado con MCO puede ser inconsistente si los efectos individuales $\boldsymbol{\alpha_i}$ están correlacionados con los regressors $\mathbf{x_{it}}$

- Dicha correlación implicaría que el error combinado ($\boldsymbol{\alpha_i} - \boldsymbol{\alpha} + \boldsymbol{\varepsilon_{it}}$) esta correlacionado de alguna manera con los regresores

Qué tipo de correlación puede existir y cómo se puede estimar?

Opciones dependen de los supuestos con la correlación entre los errores

1. Pooled estimator xtreg, pa (estimación agrupada con variaciones en los errores) sus opciones son:
2. corr(independent) Cor[u_{it},u_{is}]=0 para s distinto de t. En este caso el estimador es el mismo del pooled model
3. Corr(exchangeable) Cor[u_{it},u_{is}]= ρ para todo s distinto a t es decir los errores están equicorrelacioneados.
4. Corr(ar k) se asume un comportamiento autorregresivo de orden k
5. Corr(stationary g) supone un comportamiento de media móviles MA de orden g

- **xtreg lwage exp exp2 wks ed, pa corr(ar 2) vce(robust)**

```
. * Population-averaged or pooled FGLS estimator with AR(2) error
. xtreg lwage exp exp2 wks ed, pa corr(ar 2) vce(robust) nolog

GEE population-averaged model                 Number of obs      =      4165
Group and time vars:                  id t    Number of groups   =       595
Link:                            identity     Obs per group: min =         7
Family:                          Gaussian                    avg =       7.0
Correlation:                       AR(2)                     max =         7
                                             Wald chi2(4)       =    873.28
Scale parameter:                 .1966639    Prob > chi2        =    0.0000

                                             (Std. Err. adjusted for clustering on id)
```

lwage	Coef.	Semi-robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
exp	.0718915	.003999	17.98	0.000	.0640535	.0797294
exp2	-.0008966	.0000933	-9.61	0.000	-.0010794	-.0007137
wks	.0002964	.0010553	0.28	0.779	-.001772	.0023647
ed	.0905069	.0060161	15.04	0.000	.0787156	.1022982
_cons	4.526381	.1056897	42.83	0.000	4.319233	4.733529

- Los coeficientes cambiaron considerablemente en comparación con el modelo agrupado simple.

- Se puede observar que los errores estándar robustos son menores en comparación con el modelo agrupado para todos los regresores excepto ed.

- Esto implica que se obtuvo una mejora significativa sobre la eficiencia de la estimación, debido a que se está modelando de mejor manera la correlación del error

El estimador between

- El estimador between utiliza solo las variaciones entre los individuos o el contenido de sección cruzada en los datos.

- Promediando la información en todos los años. El modelo between por tanto es

$$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i'\beta + \bar{\varepsilon}_i$$

- Es decir las variaciones de los cortes transversales resultan en el estimador de la regresión de

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\beta + (\alpha_i - \alpha + \bar{\varepsilon}_i), \quad i = 1, \ldots, N.$$

- Donde $\bar{y}_i = T^{-1}\sum_t y_{it}$, $\bar{\varepsilon}_i = T^{-1}\sum_t \varepsilon_{it}$, y $\bar{\mathbf{x}}_i = T^{-1}\sum_t \mathbf{x}_{it}$.

- Este estimador between es consistente sólo si los regresores son independientes del error combinado

- Este sería el caso de un modelo con coeficientes constantes o como más adelante se verá para el de efectos aleatorios

- Pero no para el de efectos fijos en el que $\boldsymbol{\alpha}_i$ está correlacionado con $\mathbf{x}_{it}$, en el cual se permite cierto grado de endogeneidad.

```
. * Between estimator with default standard errors
. xtreg lwage exp exp2 wks ed, be
```

Between regression (regression on group means) — Number of obs = 4165
Group variable: id — Number of groups = 595

R-sq: within = 0.1357 — Obs per group: min = 7
between = 0.3264 — avg = 7.0
overall = 0.2723 — max = 7

F(4,590) = 71.48
sd(u_i + avg(e_i.))= .324656 — Prob > F = 0.0000

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	.038153	.0056967	6.70	0.000	.0269647	.0493412
exp2	-.0006313	.0001257	-5.02	0.000	-.0008781	-.0003844
wks	.0130903	.0040659	3.22	0.001	.0051048	.0210757
ed	.0737838	.0048985	15.06	0.000	.0641632	.0834044
_cons	4.683039	.2100989	22.29	0.000	4.270407	5.095672

Para estimar con robustos incluir la opción vce(bootstrap)

xtreg lwage exp exp2 wks ed, be

xtreg lwage exp exp2 wks ed, be vce(bootstrap)

Modelos: Efectos fijos o aleatorios

- Existen dos modelos sustancialmente diferentes según el tratamiento de α_i.

- Si se asume que el efecto de α_i es un parámetro fijo se tendrá el modelo de efectos fijos, donde el componente α_i del error compuesto se correlaciona con los regresores X_{it}, permitiendo cierto grado de endogeneidad.

- Si se asume que el efecto de α_i es una variable aleatoria entonces tendremos un modelo de efectos aleatorios; en este caso se asume que el componente α_i no se correlaciona con los regresores X_{it}, es decir son estrictamente exógenos.

Efectos fijos

- El modelo a estimar es

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

- En este modelo se permite que α_i esté correlacionada con x_{it}
- Es decir algo de los individuos (efecto within) tiene un efecto en los estimadores y se requiere controlar esto.
- Al hacer que $u_{it} = \alpha_i + \varepsilon_{it}$ se está suponiendo que x_{it} no está correlacionado con ε_{it}
- Por tanto el método elimina el efecto de las características de los individuos invariantes en el tiempo de los estimadores

- Un posible método de estimación seria llevado a cabo en forma conjunta $\alpha 1,\ldots, \alpha N$ y β; pero para un panel corto la teoría asintótica recae sobre $N \rightarrow \infty$, a medida que N crece también lo hace el número de efectos fijos a estimar (problema de los parámetros incidentales)

- El interés radica en estimar β, pero primero es necesario controlar el problema de los parámetros incidentales α_i

- Es posible estimar β en forma consistente, para los regresores que varían en el tiempo con una correcta transformación aplicando diferencias para eliminar α_i

Estimador within: efectos fijos

- La transformación within consigue eliminar los efectos fijos por medio de diferenciar las medias temporales de los individuos sobre la estimación del modelo

- Con ello se elimina el efecto fijo de α_i

- Una limitación es que los coeficientes de las variables que no cambian en el tiempo no se pueden identificar ya que
- $x_{it} = x$ entonces $(x_{it} - media_x)=0$

- Por ejemplo si queremos calcular en una regresión de los salarios el efecto del género o la raza no es posible con este método

Supuestos del modelo de efectos fijos

- Los regresores $\mathbf{x1}_{it}$; …… ; $\mathbf{xk}_{it}$ están correlacionados con α_i.. Por lo que todo el análisis será condicional a α_i.

- Los regresores deben estar no correlacionados con ε_{it}. Es decir:

$$E[\epsilon_{it} \mid \alpha_i, x_{1it}, \ldots, x_{kit}] = 0$$

- Esto implica que

$$E[y_{it} \mid \alpha_i, x_{1it}, \ldots, x_{kit}] = \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + \alpha_i$$

- A pesar de que el regresor es endógeno respecto al término del error compuesto $\mathbf{u_{it}}$, es posible identificar el efecto marginal de β_j como:

$$\frac{\delta E[y_{it}|\alpha_i, x_{1\,it}, \ldots, x_{k\,it}]}{\delta xj, it} = \beta_j$$

- La estimación de β que elimina el efecto fijo de α_i se logra restando la media de cada observación
- La transformación within estima MCO con estas observaciones
- Por lo tanto al modelo de efectos específicos individuales (Modelo panel general)

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + (\alpha_i - \alpha + \varepsilon_{it}).$$

- Le restamos los promedios para cada individuo

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}'_i\beta + (\alpha_i - \alpha + \bar{\varepsilon}_i)$$

- Y obtenemos:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- donde $\overline{X}_i = \frac{1}{T_i}\sum_t X_{it} \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T,$

xtreg lwage exp exp2 wks ed, fe vce(cluster id)

```
. * Within or FE estimator with cluster-robust standard errors
. xtreg lwage exp exp2 wks ed, fe vce(cluster id)

Fixed-effects (within) regression               Number of obs      =        41
Group variable: id                              Number of groups   =         5

R-sq:  within  = 0.6566                         Obs per group: min =
       between = 0.0276                                        avg =         7
       overall = 0.0476                                        max =

                                                F(3,594)           =   1059.
corr(u_i, Xb)  = -0.9107                        Prob > F           =    0.0(

                                              (Std. Err. adjusted for 595 clusters in i
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interva
exp	.1137879	.0040289	28.24	0.000	.1058753	.1217(
exp2	-.0004244	.0000822	-5.16	0.000	-.0005858	-.0002(
wks	.0008359	.0008697	0.96	0.337	-.0008721	.00254
ed	(dropped)					
_cons	4.596396	.0600887	76.49	0.000	4.478384	4.7144
sigma_u	1.0362039					
sigma_e	.15220316					
rho	.97888036	(fraction of variance due to u_i)				

- Los errores estándar son tres veces más grandes porque sólo se está utilizando la variación within
- El coeficiente de la educación no está identificado debido a que el dato de la educación no cambia en el tiempo
- Lo cual es desafortunado por que nos interesa saber como los salarios dependen del nivel educativo.
- Los coeficientes en niveles se interpretan como: para un individuo determinado, tanto como X varia en el tiempo en una unidad, Y se incrementa (decrece) en β%

Fixed effects: *n* entity-specific intercepts (using `xtreg`)

$$Y_{it} = \beta_1 X_{it} + \ldots + \beta_k X_{kt} + \alpha_i + e_{it} \quad \text{[see eq.1]}$$

NOTE: Add the option 'robust' to control for heteroskedasticity

Outcome variable | Predictor variable(s)

Fixed effects option

Total number of cases (rows)

Total number of groups (entities)

```
. xtreg  y  x1, fe

Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289
```

The errors u_i are correlated with the regressors in the fixed effects model

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_cons	2.41e+08	7.91e+08	0.30	0.762	-1.34e+09	1.82e+09
sigma_u	1.818e+09					
sigma_e	2.796e+09					
rho	.29726926	(fraction of variance due to u_i)				

Coefficients of the regressors. Indicate how much *Y* changes when *X* increases by one unit.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

29.7% of the variance is due to differences across panels.

'rho' is known as the intraclass correlation

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

t-values test the hypothesis that each coefficient is different from 0. To reject this, the t-value has to be higher than 1.96 (for a 95% confidence). If this is the case then you can say that the variable has a significant influence on your dependent variable (y). The higher the t-value the higher the relevance of the variable.

For more info see Hamilton, Lawrence, *Statistics with STATA*.

sigma_u = sd of residuals within groups u_i

sigma_e = sd of residuals (overall error term) e_i

Resumiendo

- El modelo de efectos fijos permite que exista cierto grado de endogeneidad, pues los regresores se relacionan con α_i.
- Una limitación es que el modelo de efectos fijos, es que la transformación within no permite identificar los coeficientes de los regresores invariantes en el tiempo
- El modelo LSDV técnicamente permite asociar las características invariantes en el tiempo con las personas o las empresas (dummys) (efectos específicos no observables del individuo)
- Por tanto el modelo de efectos fijos esta diseñado para estudiar las causas de los cambios entre las personas o las entidades

El modelo de efectos aleatorios

- La racionalidad del modelo de efectos aleatorios radica en que a diferencia del de efectos fijos se asume que la variación especifica de los individuos es aleatoria y no está correlacionada con las variables independientes incluidas en el modelo.

- Una ventaja de este modelo es que se pueden incluir las variables que no muestran cambios en el tiempo

- El modelo supone que tanto α_i como ε_{it} son i.i.d

- El modelo de efectos aleatorios es completamente eficiente si los efectos específicos individuales son estrictamente exógenos.

- Es posible estimarlo a partir del modelo transformado por MCG factibles:

$$(y_{it} - \widehat{\theta}_i \overline{y}_i) = (1 - \widehat{\theta}_i)\alpha + (\mathbf{x}_{it} - \widehat{\theta}_i \overline{\mathbf{x}}_i)'\beta + \{(1 - \widehat{\theta}_i)\alpha_i + (\varepsilon_{it} - \widehat{\theta}_i \overline{\varepsilon}_i)\}$$

- Donde el error compuesto U_{it} ahora es

$$\{(1 - \widehat{\theta}_i)\alpha_i + (\varepsilon_{it} - \widehat{\theta}_i \overline{\varepsilon}_i)\}$$

- U_{it} es asintóticamente i.i.d.

- Donde

$$\theta_i = 1 - \sqrt{\sigma_\varepsilon^2 / (T_i \sigma_\alpha^2 + \sigma_\varepsilon^2)}$$

- Este estimador puede transformarse tanto en el estimador pooled como within ya que
  - Será MCO agrupado cuando ($\widehat{\theta_i} = 0$)
  - Será within cuando ($\widehat{\theta_i} = 1$) .

- Además el estimador RE se acerca al within cuando T es grande y σ^2_α es relativamente mas grande comparada con σ^2_e ya que en estos casos $\widehat{\theta_i} \rightarrow 1$.

- Los supuestos del modelo son

$$E[\alpha_i \mid X_{it}] = 0$$

$$Var[\alpha_i \mid X_{it}] = \sigma_\alpha^2$$

$$E[\epsilon_{it} \mid X_{it}] = 0$$

$$Var[\epsilon_{it} \mid X_{it}] = \sigma_\epsilon^2$$

$$E[u_{it} \mid X_{it}] = 0$$

$$Corr(u_{it}, u_{is}) = \frac{\sigma_\epsilon^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}, \text{ para } s \neq t$$

- También es conveniente utilizar las opciones para estimar errores estándar robustos
- Esto debido a que gran parte de las aplicaciones micro-econométricas se considera que los errores son i.i.d.
- Si embargo, es muy común que los errores estén serialmente correlacionados (correlacionados sobre t para un individuo determinado)
- Pueden ser heterocedásticos.

xtset id t

xtreg lwage exp exp2 wks ed, re vce(cluster id)

```
. * Random-effects estimator with cluster-robust standard errors
. xtreg lwage exp exp2 wks ed, re vce(cluster id) theta
```

Random-effects GLS regression — Number of obs = 4165
Group variable: id — Number of groups = 595

R-sq: within = 0.6340 — Obs per group: min = 7
between = 0.1716 — avg = 7.0
overall = 0.1830 — max = 7

Random effects u_i ~ Gaussian — Wald chi2(5) = 87967.78
corr(u_i, X) = 0 (assumed) — Prob > chi2 = 0.0000
theta = .82280511

(Std. Err. adjusted for 595 clusters in id)

lwage	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
exp	.0888609	.0039992	22.22	0.000	.0810227	.0966992
exp2	-.0007726	.0000896	-8.62	0.000	-.0009481	-.000597
wks	.0009658	.0009259	1.04	0.297	-.000849	.0027806
ed	.1117099	.0083954	13.31	0.000	.0952552	.1281647
_cons	3.829366	.1333931	28.71	0.000	3.567921	4.090812
sigma_u	.31951859					
sigma_e	.15220316					
rho	.81505521	(fraction of variance due to u_i)				

- En la salida de STATA, el coeficiente sigma_u nos muestra la desviación estándar del efecto individual α_i , mientras que sigma_e es la desviación estándar del error idiosincrático ε_{it}.

- Rho es igual a la correlación intraclase del error ρ_u

- corr(u_i, Xb) = Mide la correlación de los errors u_i con los regresores

- La salida theta indica la estimación de $\widehat{\theta}_i$ como 0.823 por lo que la estimación de RE es más cercana a la estimación por transformación within que por MCO

NOTE: Add the option 'robust' to control for heteroskedasticity

Outcome variable | Predictor variable(s) | Random effects option

```
. xtreg y x1, re

Random-effects GLS regression                   Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:  within  = 0.0747                         Obs per group: min =        10
       between = 0.0763                                        avg =      10.0
       overall = 0.0059                                        max =        10

Random effects u_i ~ Gaussian                   Wald chi2(1)       =      1.91
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.1669
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	1.25e+09	9.02e+08	1.38	0.167	-5.21e+08	3.02e+09
_cons	1.04e+09	7.91e+08	1.31	0.190	-5.13e+08	2.59e+09
sigma_u	1.065e+09					
sigma_e	2.796e+09					
rho	.12664193	(fraction of variance due to u_i)				

Differences across units are uncorrelated with the regressors

If this number is < 0.05 then your model is ok. This is a test (F) to see whether all the coefficients in the model are different than zero.

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (95%, you could choose also an alpha of 0.10), if this is the case then you can say that the variable has a significant influence on your dependent variable (y)

Interpretation of the coefficients is tricky since they include both the within-entity and between-entity effects. In the case of TSCS data represents the average effect of *X* over *Y* when *X* changes across time and between countries by one unit.

Algunas consideraciones

- La interpretación de los coeficientes
  - La interpretación de los coeficientes es más compleja que en el modelo FE
  - Por tanto los coeficientes β representan el efecto medio de X en Y cuando X cambia en el tiempo y entre los individuos en una unidad.

Within and between R^2

- R^2 se interpreta como la correlación entre los valores observados y los estimados de la variable dependiente, sin embargo los valores ajustados ignoran la contribución de $\tilde{\alpha}_i$

- Si $\widehat{\beta}$ es el coeficiente estimado, si $\rho^2(x, y)$ indica la correlación al cuadrado entre x e y, entonces;

$$\begin{aligned}
&\text{Within } R^2\text{:} && \rho^2\{(y_{it} - \overline{y}_i), (\mathbf{x}'_{it}\widehat{\beta} - \overline{\mathbf{x}}'_i\widehat{\beta})\} \\
&\text{Between } R^2\text{:} && \rho^2(\overline{y}_i, \overline{\mathbf{x}}'_i\widehat{\beta}) \\
&\text{Overall } R^2\text{:} && \rho^2(y_{it}, \mathbf{x}'_{it}\widehat{\beta})
\end{aligned}$$

En los modelos

	Within	Between	RE
Within R2			
Between R2			
Overall R2			

En los modelos

	Within	Between	RE
Within R2	0.66	0.14	0.63
Between R2	0.03	0.33	0.17
Overall R2	0.05	0.27	0.18

Comparación de modelos

- Realicemos una comparación de los estimadores de panel, errores estándar, componentes de la varianza y R^2

- global xlist exp exp2 wks ed
- **POOLED ROBUSTOS**
- regress lwage $xlist, vce(cluster id)
- estimates store OLS_rob

BETWEEN

- xtreg lwage $xlist, be
- estimates store BE

EFECTOS FIJOS

- xtreg lwage $xlist, fe
- estimates store FE

Efectos fijos robustos

EFECTOS FIJOS ROBUSTOS

- xtreg lwage $xlist, fe vce (robust)
- estimates store FE_rob

EFECTOS ALEATORIOS

- xtreg lwage $xlist, re
- estimates store RE

Efectos aleatorios robustos

xtreg lwage $xlist, re vce (robust)

estimates store RE_rob

Vamos a colocar la información en una Tabla

- estimates table OLS_rob BE FE FE_rob RE RE_rob , b se stats (N r2 r2_o r2_b r2_w sigma_u sigma_e rho)

Variable	OLS_rob	BE	FE	FE_rob	RE	RE_rc
exp	0.0447	0.0382	0.1138	0.1138	0.0889	0.08
	0.0054	0.0057	0.0025	0.0040	0.0028	0.00
exp2	-0.0007	-0.0006	-0.0004	-0.0004	-0.0008	-0.00
	0.0001	0.0001	0.0001	0.0001	0.0001	0.00
wks	0.0058	0.0131	0.0008	0.0008	0.0010	0.00
	0.0019	0.0041	0.0006	0.0009	0.0007	0.00
ed	0.0760	0.0738	0.0000	0.0000	0.1117	0.11
	0.0052	0.0049	0.0000	0.0000	0.0061	0.00
_cons	4.9080	4.6830	4.5964	4.5964	3.8294	3.82
	0.1400	0.2101	0.0389	0.0601	0.0936	0.10
N	4.2e+03	4.2e+03	4.2e+03	4.2e+03	4.2e+03	4.2e+
r2	0.2836	0.3264	0.6566	0.6566		
r2_o		0.2723	0.0476	0.0476	0.1830	0.18
r2_b		0.3264	0.0276	0.0276	0.1716	0.17
r2_w		0.1357	0.6566	0.6566	0.6340	0.63
sigma_u			1.0362	1.0362	0.3195	0.31
sigma_e			0.1522	0.1522	0.1522	0.15
rho			0.9789	0.9789	0.8151	0.81

legend:

Qué modelo usar?

- Desde el punto de vista conceptual, el modelo de efectos aleatorios, es apropiado cuando las N unidades transversales son una muestra (aleatoria) de una población mayor (individuos, familias, empresas, etc.); en este caso cabe esperar que el efecto individual se caracterice mejor por una variable aleatoria.

- El estimador de Efectos Aleatorios es más eficiente, si se cumplen supuestos adicionales respecto al de Efectos Fijos.

- El estimador de Efectos Fijos permite estimar el modelo bajo supuestos menos restrictivos:
  - Permite correlación entre los regresores y los efectos individuales.
  - Permite estimar el modelo incluso si los regresores son endógenos.
- Pero es menos deseable
  - Al ser menos eficiente (al explotar solo la variación within).
  - No identifica los coeficientes de regresores que no varían en el tiempo.

Qué modelo usar?

- La elección más importante es entre FE o RE
- Si los efectos son fijos entonces el modelo pooled y RE producen estimadores inconsistentes y por tanto debe utilizarse el estimador within
- Recordar que RE asume un supuesto muy fuerte que α_i tiene una distribución independiente de x_i; por lo tanto, si los efectos son aleatorios, la estimación por FE seria menos eficiente pues solo explota la variación within de los datos

La prueba de Hausman

- Se parte del supuesto que el verdadero modelo es el de efectos aleatorios con α_i , iid [0, σ^2_α] y con el error ε_{it}, iid[0,σ^2_ε], donde ambas no se correlacionan con los regresores y que por tanto β_{RE} eficiente

- La prueba compara los coeficientes estimables de los regresores que varían en el tiempo o se puede aplicar a un subconjunto clave de estos.

- El test de Hausman compara los estimadores de la siguiente manera

$$H = \left(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right)' \left[\hat{V}[\hat{\beta}_{1,W}] - \hat{V}[\tilde{\beta}_{1,RE}]\right]^{-1} \left(\tilde{\beta}_{1,RE} - \hat{\beta}_{1,W}\right),$$

Prueba de Hausman

- La hipótesis nula es que el α_i se comporta como efectos aleatorios, donde el supuesto de exogeneidad estricta garantiza que el estimador es eficiente.

- Pasos

1. Correr el modelo FE y guardar la estimación.
2. Correr el modelo RE y guardar la estimación
3. Hacer la prueba con el comando Hausman

- Como ya tenemos guardados los resultados de la estimación por FE y RE podemos aplicar la prueba directamente

- **hausman FE RE, sigmamore**

- Debido a que la construcción de las varianzas de los estimadores para FE y RE emplean diversas formas en la matriz de varianzas-covarianza, aplicamos la opción sigmamore que específica una estructura similar para hacerlas comparables.

```
* Hausman test assuming RE estimator is fully efficient under null hypothesis
hausman FE RE, sigmamore
```

	Coefficients (b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
exp	.1137879	.0888609	.0249269	.0012778
exp2	-.0004244	-.0007726	.0003482	.0000285
wks	.0008359	.0009658	-.0001299	.0001108

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 1513.02
Prob>chi2 = 0.0000