Econometría de Económicas

Apuntes para el tema 6

Curso 2004-2005

Profesoras
Amparo Sancho
Guadalupe Serrano

Modelos de panel de datos

Datos de Panel son aquellos que surgen de la observación de una misma sección cruzada o corte transversal con N individuos a lo largo del tiempo. En ellos, se obtiene información para cada uno de los individuos, i = 1, 2, 3...N, para cada momento del tiempo, t = 1, 2, 3....T, tratándose de una muestra de NxT observaciones. Generalmente las variables observadas se identifican para cada individuo, i, y momento del tiempo, t: Y_{it}.

	i =1	i=2	i=3	….	i = N
1980	3.5	2.5	4.2		1.8
1981	2.8	2.1	4.6		1.9
1982	2.6	1.9	4.1		1.7
1983	3.2	2.2	3.9		2.0

Se denomina:
Micropanel si la información a analizar corresponde a agentes individuales. En general se dispone de un número muy elevado de individuos y pocas observaciones temporales para cada uno.
Macropanel si la información a analizar corresponde a otras unidades de análisis (países, regiones, etc) para los que se dispone de muchas observaciones temporales correspondientes a pocos individuos.

Ejemplos de Fuentes estadísticas con información de panel:
- datos referidos a empresas de la Central de Balances del Banco de España,
- Encuestas de presupuestos familiares
- Encuesta de Población Activa.
- Encuesta permanente de consumo.
- Encuesta industrial
- Bases de datos multi-país como OCDE o EUROSTAT y bases de datos multi-región como Contabilidad Regional de España o REGIO de Eurostat.

¿Por que Datos de Panel?

- Proporciona una información muy válida de los individuos siguiéndolos a través del tiempo, lo que ofrece una visión más completa del problema, interpretando mejor la dinámica del cambio. Ejemplo: estudio de la movilidad laboral a través de encuesta de población activa o hábitos de consumo (respuestas dinámicas).
- Elimina el sesgo de la agregación al trabajar con datos desagrupados.
- Elimina el sesgo de especificación que tienen los modelos de series temporales que no tienen en cuenta las características inobservables de los individuos que podrían estar condicionando su comportamiento, o bien efectos latentes en cada período de tiempo que pueden alterar el comportamiento de un mismo individuo en distintos momentos del tiempo.
- La unión de la dimensión temporal e individual del problema proporciona mayor número de grados de libertad en el análisis.
- Proporciona información que permite mitigar o reducir los problemas de multicolinealidad respecto a los modelos de serie temporal.
- Explica mejor los fenómenos más complejos como el cambio tecnológico.

Los problemas asociados a estos modelos son:
- Abandono de la muestra por ciertos individuos, por lo que no es posible realizar su seguimiento a lo largo del tiempo. Si esto ocurre de forma sistemática puede proporcionar problemas de censura en la muestra. En general, se puede solucionar mediante la observación temporal de individuos con características muy similares (datos de cohortes).
- No aleatoriedad de la muestra, lo que implicaría decisiones erróneas y no representativas de la muestra
- Desequilibrios en la muestra que haga que se tenga más información de algunos individuos que de otros (Panel no equilibrado o incompleto).

Paneles artificiales o cohortes: Cuando se establece un panel mediante la observación temporal de individuos con características similares y no de un mismo individuo.

PLANTEAMIENTO DEL PROBLEMA

En general, el análisis del comportamiento de varios individuos a lo largo del tiempo se podría realizar mediante diferentes modelos econométricos para:

- contrastar una hipótesis y su robustez para distintos individuos, es decir, contrastar la homogeneidad en el comportamiento de distintos individuos (viendo si los parámetros del modelo para cada uno de ellos son iguales)
- Analizar los factores que influyen en las decisiones o el comportamiento de los agentes.

Puesto que pueden existir diferencias en el comportamiento de los individuos, la fuente de variación de la muestra es importante en la formulación y la estimación de los modelos económicos. Dependiendo si se considera que los parámetros del modelo son distintos entre individuos o entre períodos de tiempo se pueden analizar diferentes especificaciones que permiten recoger dichas diferencias en el comportamiento de los individuos y también entre períodos de tiempo.

Para un caso donde se observan una variable endógena Y y dos variables explicativas X_1 y X_2 para cada individuo y periodo de tiempopodríamos tener las siguientes especificaciones:

1.- Todos los coeficientes son constantes en el tiempo y para todos los individuos. Modelo de coeficientes constantes o Modelo de POOL de datos.

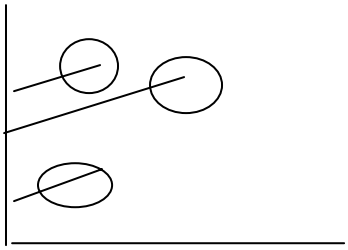
La intersección y los coeficientes son constantes respecto al tiempo y entre individuos. Queda por lo tanto asimilado al término aleatorio las posibles diferencias entre individuos y diferentes momentos del tiempo:

$$Y_{it}= \alpha + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (2)$$

La solución a este modelo sería considerar toda la información sin diferenciar entre individuos o periodos temporales (pool de datos). El modelo se estimaría por MCO.

En estos modelos puede encontrarse problemas de autocorrelación debido a que la varianza de las perturbaciones pueda ser diferente respecto a los individuos o en el tiempo, y/o heterocedasticidad, solucionándose mediante la aplicación de MCG.

2.- Todos las pendiente o coeficientes de las variables son constantes pero no así la intersección.



En este modelo, la heterogeneidad en el comportamiento de las unidades maestrales queda recogida a través de los términos independientes, es decir que en el caso de la igualdad de los valores medios de las variables explicativas entre individuos, el valor medio de la variable dependiente sería diferente. Sin embargo, el impacto de las variables explicativas sobre la variable indígena sería el mismo para todos ellos.

$$Y_{it}= \alpha_i + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (3)$$

Este caso es semejante al resultado de aplicar variables ficticias D para modelizar el parámetro independiente para cada uno de los i individuos.

$Y_{it}= \alpha_1 +\alpha_2 D_2 + \alpha_3 D_3 +...+ \alpha_N D_N + \beta_2 X_{2it} + \beta X_{3it} +u_{it}$ (4)

Con el fin de estimar este modelo, la forma operativa de hacerlo es transformar las variables en desviaciones respecto a su media temporal para cada individuo.
El modelo propuesto quedaría

$$(Y_{it} - \overline{Y}_i) = \sum_{k=2}^{k} \beta_k (X_{kit} - \overline{X}_{ki}) + u_{it} - \frac{\sum u_{it}}{T}$$

donde $\overline{X}_{ki} = \sum_{t=1}^{T} X_{kit} / T$

Con el fin de determinar si la heterogeneidad del modelo proviene de las diferencias en el término independiente, y, por tanto que el modelo (3) supone una mejor especificación que el modelo (2) se puede realizar mediante un contraste basado en el estadístico F, que se formula de la forma siguiente: $F_{N-1,NT-k-N} = \frac{(R_G^2 - R_{rest}^2) / N - 1}{(1 - R_G^2) / NT - k - N}$

Donde: R^2_G es el coeficiente de determinación del modelo general (modelo 3) y R^2_{rest} es el coeficiente de determinación del modelo restringido (modelo 2). Este contrste también se denomina Análisis de la Covarianza.

3.- Los coeficientes son constantes pero la intersección varia conforme a los individuos y el tiempo.

En este caso el modelo se presenta de la forma siguiente:

$Y_{it}= \alpha_1 +\alpha_2 D_2 + \alpha_3 D_3 +...+ \alpha_N D_N + \gamma_1 + \gamma_2$ año2 + ...+ γ_t año t + $\beta_2 X_{2it} + \beta X_{3it} +u_{it}$ (3)

4.- Existe una heterogeneidad en las pendientes entre individuos y la intersección permanece constante.

$Y_{it}= \alpha + \beta_{2i} X_{2it} + \beta_{3i} X_{3it} +u_{it}$ (3)

Sin embargo, esta especificación no tiene interés dado el objetivo del análisis econométrico.

No obstante, se da el caso de que individuos con idénticas características observables se comportan o adoptan decisiones diferentes. Ello implica la existencia de factores no observables, específicos para cada individuo, que hacen que su comportamiento sea diferente respecto al de otros individuos. Asimismo, un mismo individuo puede comportarse de diferente manera en distintos periodos de tiempo debido a factores temporales no observables característicos de cada período temporal. Si dichos factores o efectos no observables no se consideran en la especificación del modelo existirá un problema de variables omitidas: la estimación de sus parámetros está sesgada porque recoge parte de estos efectos individuales o temporales no observables.

Si se supone que dichos efectos individuales son constantes en el tiempo, o bien los temporales son constantes entre individuos, permite realizar inferencia sobre el comportamiento de los individuos aún existiendo dichos factores no observables. Esta es la idea que subyace en los modelos de datos de panel.

En general, el estudio de los modelos de datos de panel hace referencia a los casos 2 y 3, en los que la heterogeneidad de los parámetros implica que considerar la información como un panel de datos y no como un pool mejorará la estimación del modelo evitando el problema de la omisión de variables.

En los modelos de datos de panel, la discusión se centra en el análisis de los efectos individuales, α_i, que se consideran factores no observables y constantes en el tiempo que son específicos para cada uno de los individuos y que pueden estar correlacionados con las variables explicativas (por ejemplo, la habilidad de los comerciales de cada una de las empresas, además de sus horas de trabajo, a la hora de determinar las ventas de las mismas; la habilidad de los trabajadores, además de su cualificación, para determinar su salario). De acuerdo con dicha posibilidad, los efectos individuales se han tratado en la literatura como efectos individuales fijos, cuando los α_i, están correlacionados con las variables explicativas y los efectos individuales aleatorios, cuando no existe tal correlación. Dicha consideración de efectos fijos o aleatorios, condiciona el análisis y la inferencia realizada con el modelo.

Si α_i es fijo en el muestreo, la inferencia estará condicionada a estos efectos individuales muestrales que no tienen porqué representar a los poblacionales. En este caso los efectos individuales estarán correlacionados con las variables explicativas y la inferencia estará condicionada a las realizaciones de los α_i en la muestra. Por el contrario, si α_i es aleatorio, se realizará inferencia no condicionada puesto que los efectos individuales no están correlacionados con las variables explicativas y además se realizan hipótesis sobre la distribución poblacional de los mismos.

Presentación del modelo de Panel

El modelo de Panel se representa de la forma siguiente:

$Y_{it}=\alpha_i+ \beta X_{it} +u_{it}$

con una sola variable explicativa y siendo i=1,2,....N grupos (o individuos); t=1,2,...T periodos de tiempo. Además se supone que $E(u_{it})=0$ y tiene varianza constante σ_u^2

Para el caso multivariante:
$Y_{it}= \alpha_i + X_{Ki} \beta_k +u_{it}$ donde X es una matriz de K variables explicativas para cada uno de los individuos de forma que:

$$Y_{it}= \begin{bmatrix} y_{1t} \\ y_{2t} \\ : \\ y_{Nt} \end{bmatrix}$$ donde y_{it} es el vector que contiene la información del individuo i en todo t

X_{kit} es la matriz de observaciones de la variables explicativa k, para el individuo i, en el tiempo t.

$$X_{Kit} = \begin{bmatrix} X_{11t} & X_{21t} & .. & X_{K1t} \\ X_{12t} & X_{22t} & .. & X_{K2t} \\ : & : & .. & : \\ X_{1Nt} & X_{2Nt} & .. & X_{KNt} \end{bmatrix}$$ a su vez $\beta_k = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \\ \beta_K \end{bmatrix}$ y $u_{it} = \begin{bmatrix} u_{1t} \\ u_{2t} \\ : \\ u_{Nt} \end{bmatrix}$ vector que contiene las

t perturbaciones aleatorias de cada individuo.

El vector de parámetros $\alpha'_i = [\alpha_1, \alpha_2 ,\ldots, \alpha_N]$ recoge los **efectos individuales**.

Los supuestos que se realizan en estos modelos son fundamentalmente la no correlación entre las perturbaciones de cada uno de los grupos, y la no correlación temporal, y que las varianzas de las perturbaciones son homocedásticas y no autocorrelacionadas.

$E[u_{it}] = 0$
$var[u_{it}]=\sigma^2$
$cov[u_{it}\ u_{js}]= o$

Estos supuestos son fácilmente violables pudiéndose dar correlación entre grupos y en periodos de tiempo.

Modelo de efectos fijos

Este modelo supone que las diferencias entre unidades pueden captarse mediante diferencias en el término constante. Por ello cada α_i debe de ser estimado.
Este método sería equivalente a estimar por MCO un modelo con tantas variables ficticias como individuos. Este modelo sería equivalente al modelo de variables ficticias expuesto en el apartado 2. El modelo puede ser estimado por m.c.o. si el tamaño muestral longitudinal es pequeño. Sin embargo, cuando la muestra longitudinal es muy extensa (un elevado número de individuos) el problema radicaría en el elevado número de parámetros a estimar.

En este modelo, dada la correlación entre los efectos individuales y las variables explicativas su estimación por MCO

$$Y_{it}= \alpha_i + \beta_1 X_{1it} + \beta_2 X_{2it} + u_{it}$$

sería no consistente por problemas de especificación, puesto que en dichas estimaciones no se podría separar el efecto de las explicativas y de los efectos individuales (ejemplo si una de las explicativas es una variable ficticia, ésta puede ser colineal con el efecto fijo de esa familia).

La solución propuesta consiste en realizar una transformación del modelo que elimine dichos efectos y permita separar el de las variables explicativas mediante el siguiente procedimiento:

a) se calculan las medias temporales para cada individuo $X_{it} - \overline{X}_i$
b) se transforman las variables en desviaciones a su media para cada individuo de forma que se eliminan dichos efectos individuales

$$(Y_{it} - \overline{Y}_i) = \sum_{k=1}^{2} \beta_k (X_{kit} - \overline{X}_{ki}) + u_{it} - \frac{\sum u_{it}}{T}$$

donde $\overline{X}_{ki} = \sum_{t=1}^{T} X_{kit} / T$

c) se aplica MCO al modelo transformado obteniéndose estimadores consistentes y eficientes de los β. A partir de la obtención de los β se obtiene los α_i como el residuo medio del grupo i es decir: $\hat{\alpha} = \overline{y}_{i.} - \hat{\beta}\overline{X}_{i.}$
d) Éste es el denominado estimador intragrupos (within groups).

Cuando se presenta un modelo de Panel donde N es grande la utilización del proceso de efectos fijos puede ser inviable. Igualmente, modelo de efectos fijos sería exclusivamente aplicable a las unidades de la muestra, y no a unidades fuera de ella.¿¿¿¿Explicar mejor???

Modelo de efectos aleatorios

En este caso, se considera que el parámetro α_i es una variable aleatoria, cuyas realizaciones son los efectos individuales de los agentes que componen el panel (escogidos mediante un muestreo aleatorio) y distribuida independientemente de X. Este valor es diferente por lo tanto para cada individuo y se supone que difiere en cada uno de ellos de un valor medio α

De forma que :

$\alpha_i = \alpha + \varepsilon_i$ A su vez ε_i es una $NI(0,\sigma^2_\varepsilon)$

Con lo que el modelo de efectos aleatorios quedaría expuesto de la forma siguiente:

$$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + \varepsilon_i + u_{it}$$

Con ello el nuevo término de perturbación sería $w_{it} = \varepsilon_i + ui_t$. Bajo el supuesto de que no están correlacionados, la varianza de $w = \sigma^2_{\varepsilon} + \sigma^2_u$.

Si en este caso se estimara por MCO, los estimadores serían consistentes, pero no eficientes al no considerar σ^2_{ε} . Por ello, el método de estimación eficiente en este caso y dado la composición del término de perturbación aleatoria es el de los mínimos cuadrados generalizados (Mínimos cuadrados ponderados).

Modelo de efectos fijos vs modelo de efectos aleatorios

Si el tamaño muestral es grande en cuanto al tiempo (es decir T es grande) y pequeño en cuanto a los individuos (I es pequeño) existe poca diferencia entre los dos métodos. Si por el contrario I es grande y T pequeña estos estimadores pueden cambiar.
Si no se han utilizado extracciones aleatorias es mejor el modelo de efectos fijos, en caso contrario es mejor el de efectos aleatorios

Se puede utilizar el test de Hausman para enfrentar a ambos modelos valorando para ello las matrices de varianzas-covarianzas de los estimadores obtenidos.

La hipotesis nula planteada es que no existe correlación entre las X y los efectos individuales por lo que, bajo la hipótesis nula, el estimador de MCG (efectos aleatorios) sería consistente y eficiente mientras que el estimador intragrupos sería solo consistente (efectos fijos). Bajo la hipótesis alternativa, solo el estimados intragrupos sería consistente.
El test se plantea con la expresión siguiente

$$H = (\beta^{EF} - \beta^{EA})\,(M_1 - M_0)^{-1}\,(\beta^{EF} - \beta^{EA}) \sim \chi^2_k$$

Siendo:

M_1 y M_0 la matriz de varianzas covarianzas del modelo estimado por efectos fijos y aleatorios respectivamente.
β^{EF} es el vector de coeficientes estimados del modelo de efectos fijos
β^{EA} es el vector de coeficientes estimados del modelo de efectos aleatorios

Si se rechaza la hipótesis nula es decir el valor del test de Hausman es mayor que el valor de las tablas χ^2_k se aceptaría el modelo de efectos fijos.

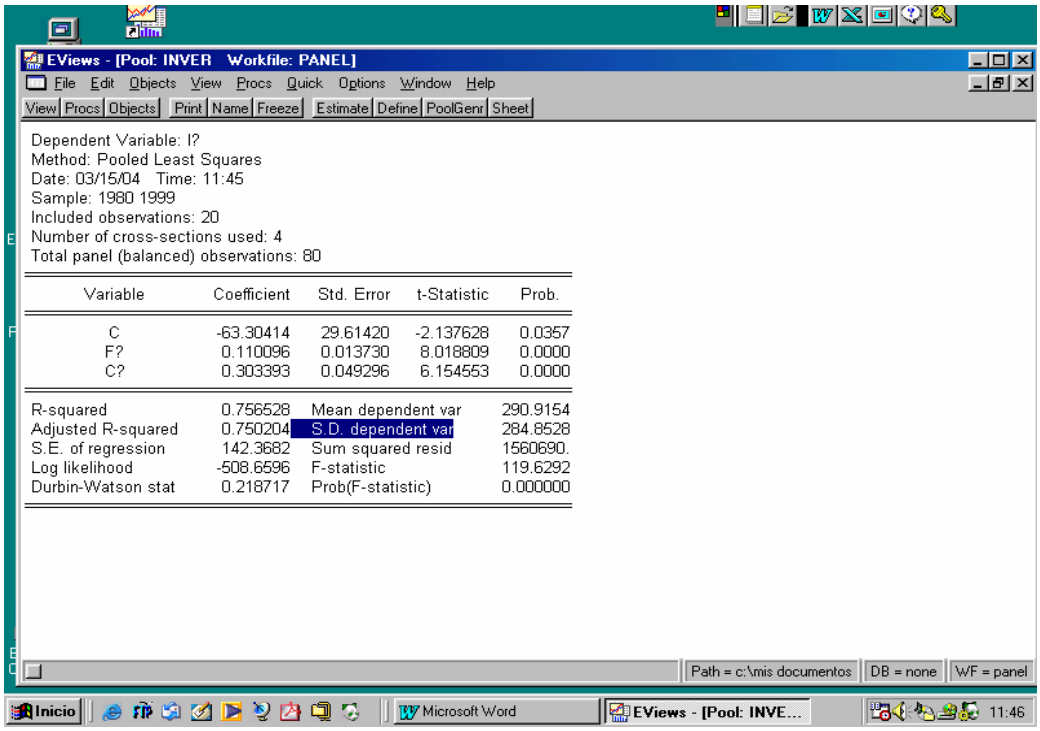
Caso práctico

Se realiza un modelo done se considera que la variable Innovación está en función de las variable F (Gasto en I+D) y C (nivel educativo del factor humano). Se dispone de información para cuatro países diferentes: GE (España), GM (Alemania), US(Estados Unidos), WEST (Francia) . El periodo temporal en el que se recoge la información es entre 1980 y2000. Por lo tanto e dispone de un Panel de información. Los datos están disponibles en el fichero /perfiles de satsuma/profesor/sanchoa/tema6/panel y en la web: www.uv/~sancho/docencia/panel1.

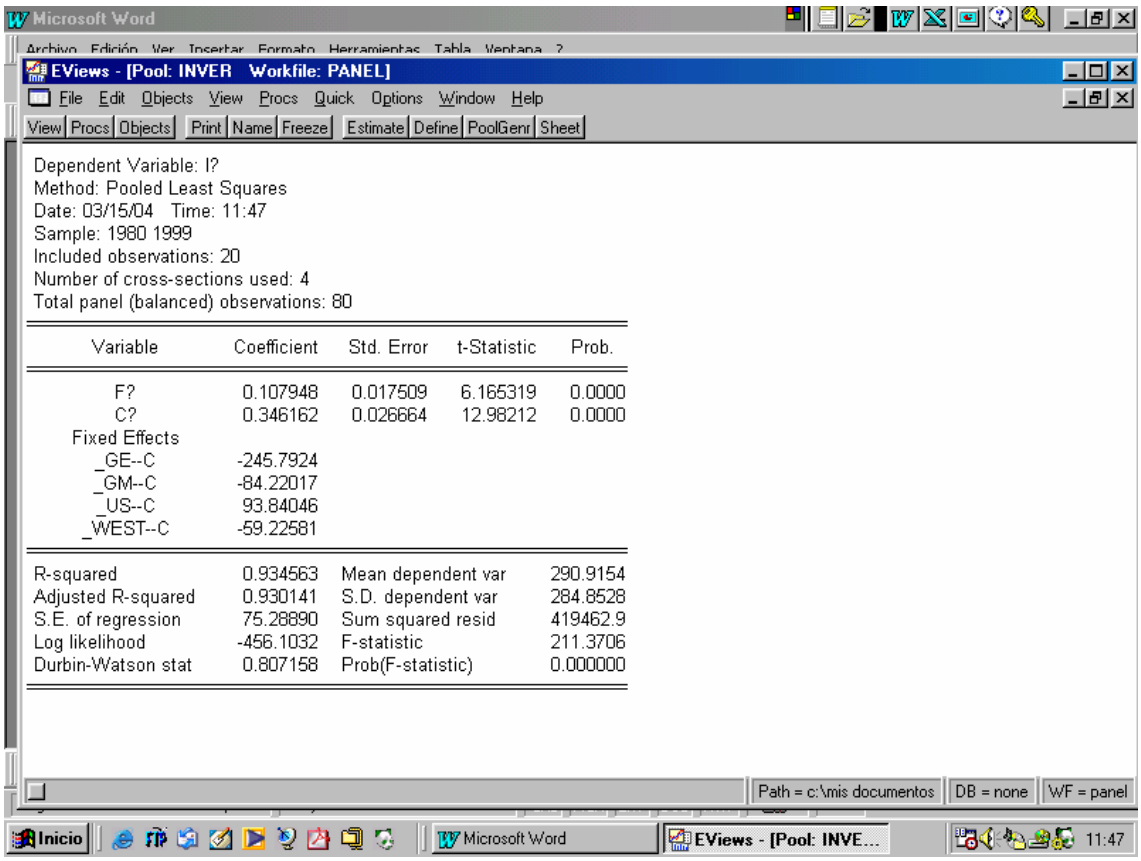
Analizando los casos anteriormente expuestos se obtiene los resultados siguientes

Caso 1.- Todos los coeficientes son constantes

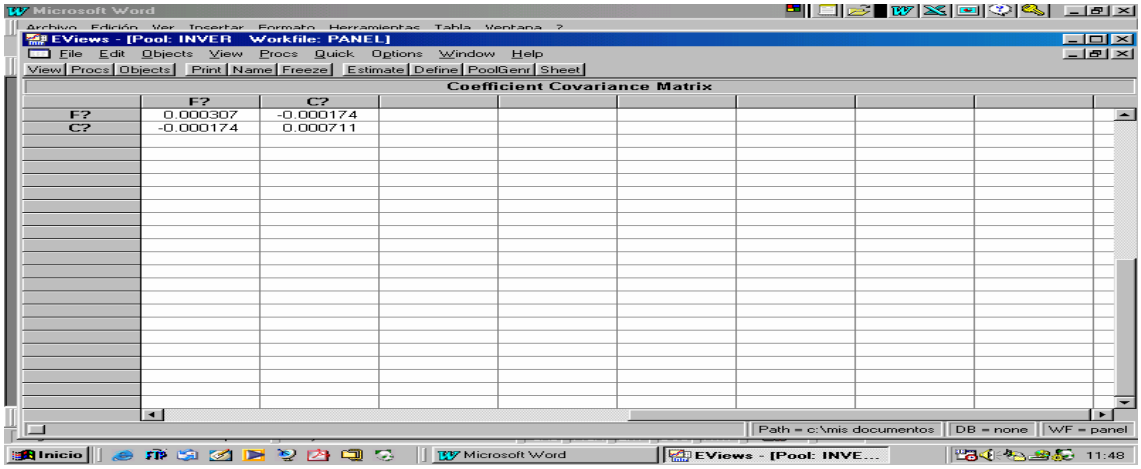
Este ejercicio se comporta como si se estimara el modelo con un pool de información. La estimación se realiza por m.c.o.



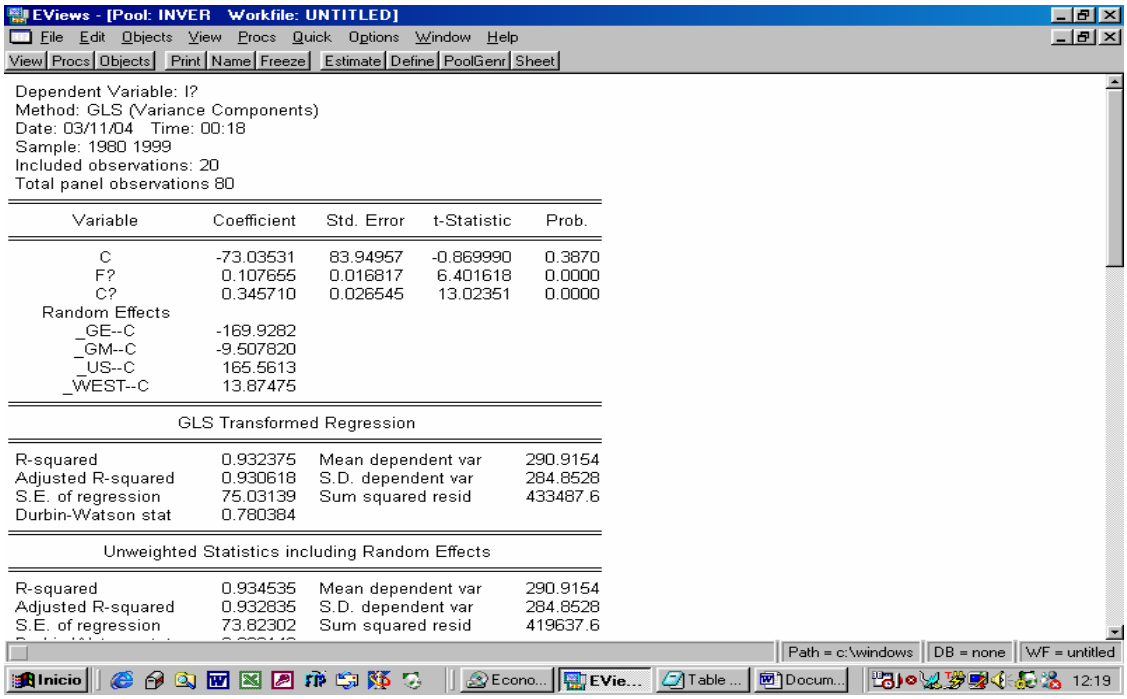
Caso 2.- Modelo de efectos fijos



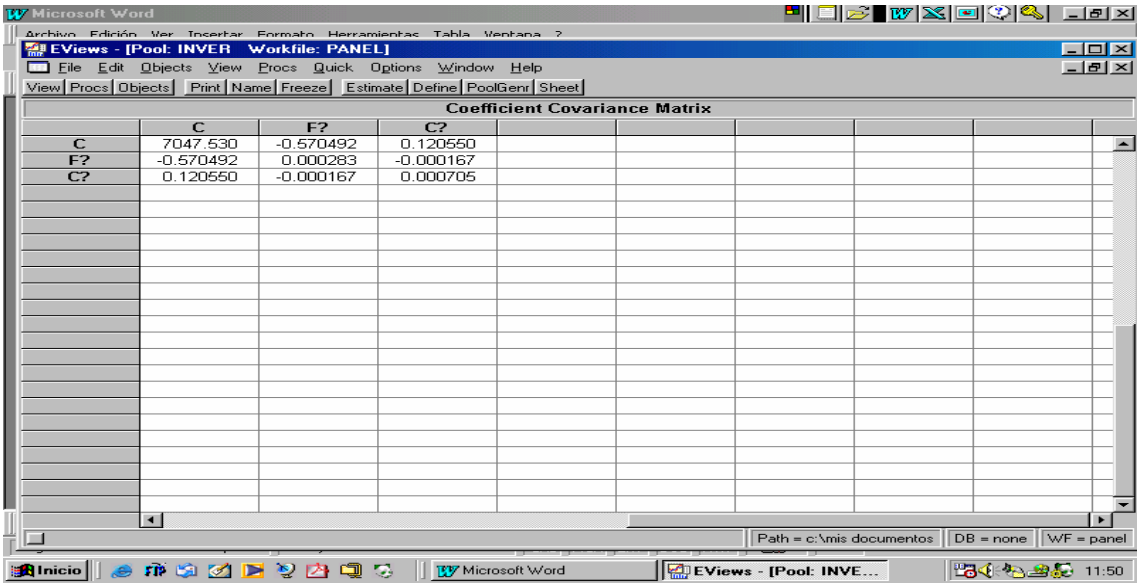
La matriz varianzas covarianzas de los estimadores



Modelo de efectos aleatorios



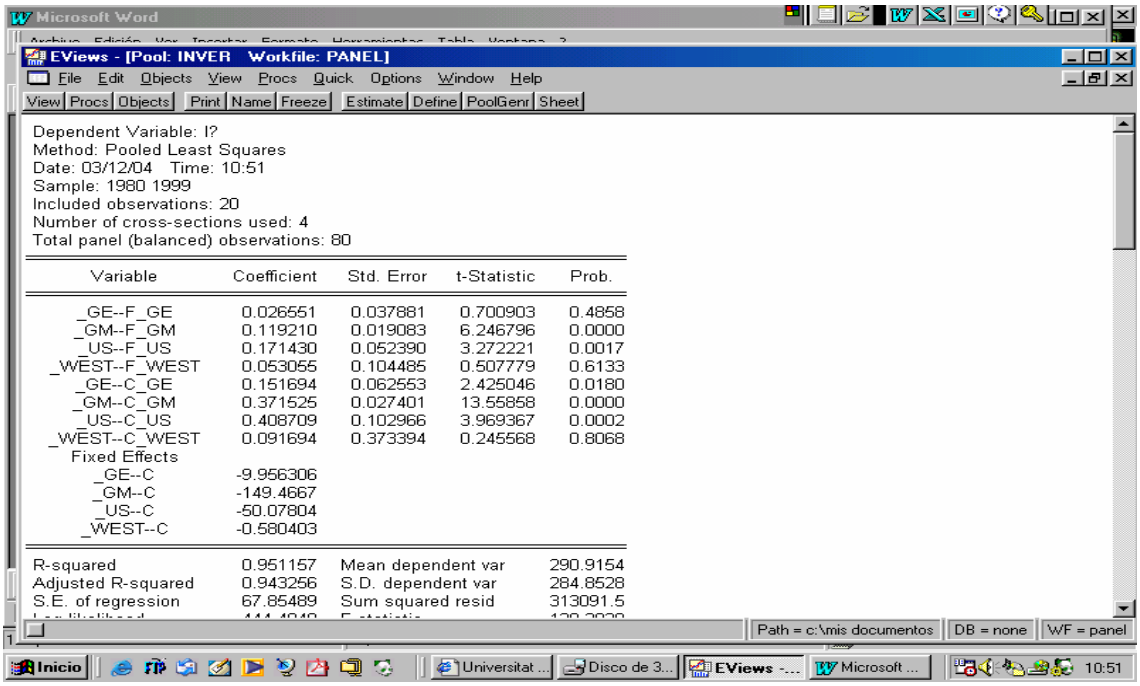
La matriz var covarianzas de los estimadores sería en esta estimación:



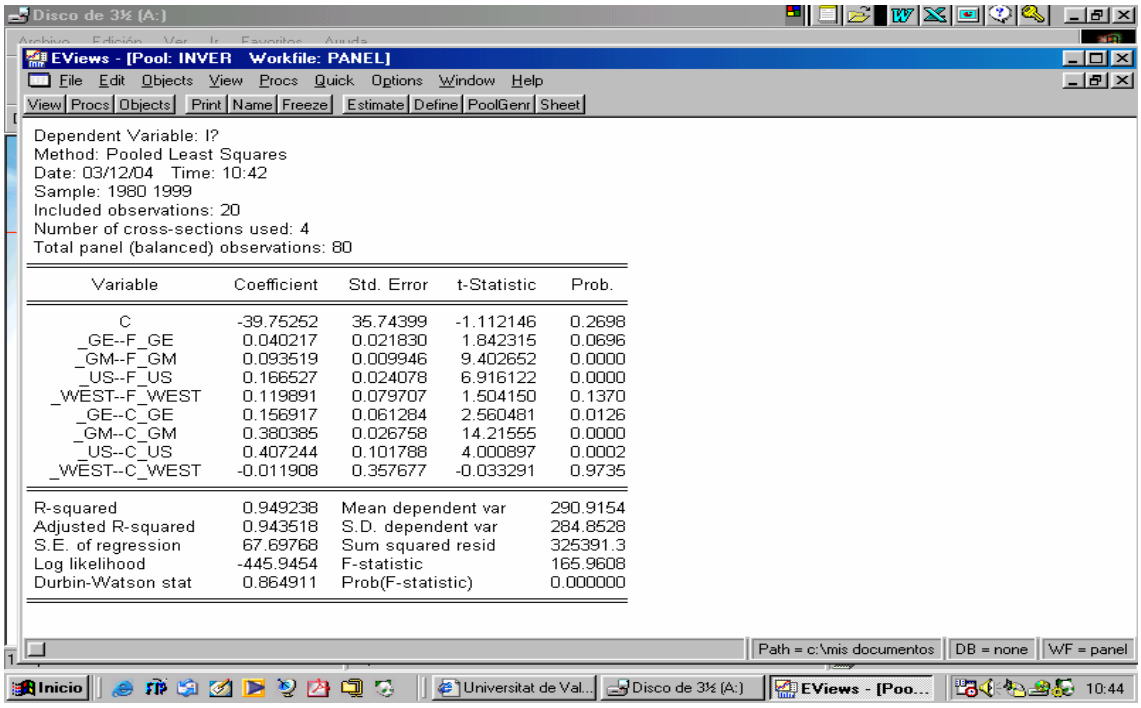
Con la información proporcionada por los vectores de valores estimados de los parámetros y las matrices varianzas y covarianzas puede obtenerse el test de Hausmasn es 0,0764.

Caso 4: Los coeficientes varían entre individuos

Es el resultado de estimar los países de forma independiente. Con lo que se aprecia que hay una ordenada en el origen por país así como las pendientes de las variables son diferentes por países.



Modelo con ordenada en el origen común a todos ellos y pendientes que varían según los países.



EViews - [Pool: INVER Workfile: PANEL]

File Edit Objects View Procs Quick Options Window Help

View | Procs | Objects | Print | Name | Freeze | Estimate | Define | PoolGenr | Sheet

Dependent Variable: I?
Method: Pooled Least Squares
Date: 03/12/04 Time: 10:42
Sample: 1980 1999
Included observations: 20
Number of cross-sections used: 4
Total panel (balanced) observations: 80

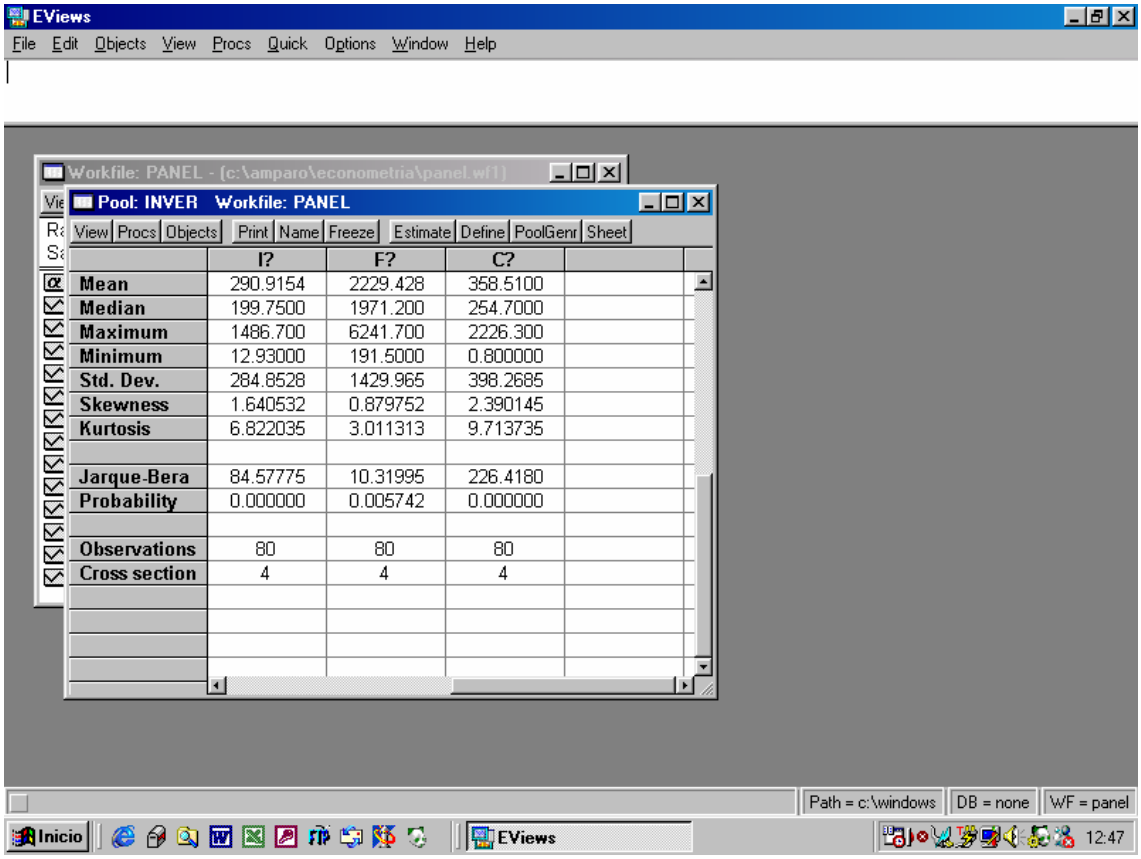
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-39.75252	35.74399	-1.112146	0.2698
_GE--F_GE	0.040217	0.021830	1.842315	0.0696
_GM--F_GM	0.093519	0.009946	9.402652	0.0000
_US--F_US	0.166527	0.024078	6.916122	0.0000
_WEST--F_WEST	0.119891	0.079707	1.504150	0.1370
_GE--C_GE	0.156917	0.061284	2.560481	0.0126
_GM--C_GM	0.380385	0.026758	14.21555	0.0000
_US--C_US	0.407244	0.101788	4.000897	0.0002
_WEST--C_WEST	-0.011908	0.357677	-0.033291	0.9735

R-squared	0.949238	Mean dependent var	290.9154
Adjusted R-squared	0.943518	S.D. dependent var	284.8528
S.E. of regression	67.69768	Sum squared resid	325391.3
Log likelihood	-445.9454	F-statistic	165.9608
Durbin-Watson stat	0.864911	Prob(F-statistic)	0.000000

Path = c:\mis documentos | DB = none | WF = panel

Inicio | Universitat de Val... | Disco de 3½ (A:) | EViews - [Poo... | 10:44

Información sobre las variables :



EViews

File Edit Objects View Procs Quick Options Window Help

Workfile: PANEL - (c:\amparo\econometria\panel.wf1)

Pool: INVER Workfile: PANEL

View | Procs | Objects | Print | Name | Freeze | Estimate | Define | PoolGenr | Sheet

	I?	F?	C?
Mean	290.9154	2229.428	358.5100
Median	199.7500	1971.200	254.7000
Maximum	1486.700	6241.700	2226.300
Minimum	12.93000	191.5000	0.800000
Std. Dev.	284.8528	1429.965	398.2685
Skewness	1.640532	0.879752	2.390145
Kurtosis	6.822035	3.011313	9.713735
Jarque-Bera	84.57775	10.31995	226.4180
Probability	0.000000	0.005742	0.000000
Observations	80	80	80
Cross section	4	4	4

Path = c:\windows | DB = none | WF = panel

Inicio | EViews | 12:47