Universidad de Valladolid

Facultad de Ciencias Económicas y Empresariales

Trabajo de Fin de Grado

Grado en Economía

Introducción a modelos de datos de panel

Presentado por:

Carlos De La Rosa Pastor

Tutelado por:

Mercedes Prieto Alaiz

Valladolid, 29 de junio de 2016

Índice

Resumen

El presente trabajo tiene como finalidad realizar una introducción a los modelos de datos de panel. Se explicarán los casos más sencillos y la manera de proceder a estimarlos. El trabajo se centra en el modelo de efectos fijos y en el modelo de efectos aleatorios, las técnicas de estimación de ambos modelos y las ventajas e inconvenientes que presentan si existe o no correlación entre la heterogeneidad individual inobservable y los regresores. Se realiza el contraste de Hausman para determinar si existe o no correlación entre los efectos latentes y los regresores y en consecuencia, se escogerá el procedimiento mas adecuado.

This paper aims to perform an introduction to panel data models. The simplest cases and how to proceed will be explained to estimate. The paper focus on the fixed effects model and the random effects model, the estimates technique for both of them and the advantages and disadvantages of wether o not correlation between the unobservable indivdual heterogeneity and regressors. the Hausman is performed to determine whether or not correlation between the latent effects and regressors and accordingly , the most suitable procedure will be chosen.

1. INTRODUCCIÓN

La Econometría se encarga, entre otras cosas, de estudiar y analizar las características de una variable económica utilizando información de otras variables que puedan explicar el comportamiento de la primera. Los principales objetivos del análisis econométrico son la especificación de un modelo que relacione variables económicas, la utilización de información muestral sobre dichas variables para cuantificar la magnitud de la dependencia entre ellas, la validación de las hipótesis propuestas por la teoría económica acerca de esta relación y la realización de un seguimiento y previsión de las variables analizadas.

La naturaleza de los datos en economía puede ser muy diversa. Esto va a condicionar el tipo de análisis que se puede utilizar por eso, debemos diferenciar entre tres tipos de datos:

- **Los datos de series de tiempo** son un conjunto de observaciones sobre los valores de una variable en diferentes momentos temporales. Tal información debe recopilarse en intervalos regulares, es decir, en forma diaria (precios de acciones, informes del tiempo), semanal (cifras de oferta monetaria), mensual (Índice de Precios al Consumidor (IPC)), trimestral (el PIB) o anual.

  Cuando se trabajo con datos de series temporales, la relación que aparece entre una variable dependiente y un conjunto de variables explicativas puede ser:

  - una relación estática, en la que todas las variables del modelo están referidas al mismo periodo de tiempo. Por ejemplo:$y_t = \alpha + \beta_1 x_t + \varepsilon_t$.
  - o una relación dinámica, en la que aparece alguna variable exógena o endógena retardada algún periodo. Por ejemplo: $y_t = \alpha + \beta_1 x_1 + \beta_2 y_{t-1} + \varepsilon_t$.

- **Los datos transversales** proporcionan información en un mismo instante de tiempo de diversos individuos de una naturaleza similar.

- **Los datos panel** es un conjunto de observaciones de series temporales sobre una muestra de unidades individuales, es decir, un conjunto de individuos son observados en distintos momentos en el tiempo, como el Panel de Estudio Dinámico de los Ingresos o la Encuesta de Condiciones de vida.

Entre los ejemplos de datos de panel más utilizados, se encuentra el «Panel Study of Income Dynamics(Panel de Estudio Dinámico de los Ingresos (PSID))», realizado por la Universidad de Michigan, que recoge información anual sobre los distintos miembros que forman las familias del estudio tomando información sobre los cambios profesionales, cambios de ingreso, cambios en el estado civil y otras características socioeconómicas y demográficas.

En España, se realiza la Encuesta de Condiciones de Vida (ECV), elaborada por el Instituto Nacional de Estadística (INE), cuyo objetivo fundamental es disponer de una fuente de referencia sobre estadísticas comparativas de la distribución de ingresos y la exclusión social en el ámbito europeo. La ECV no es un panel puro debido a que la muestra sobre la que se realiza dicha encuesta se renueva cada cuatro años, por lo tanto, no son siempre los mismo individuos. Dicha encuesta sirve a la Comisión Europea como un instrumento estadístico para medir distintos indicadores, como pueden ser: la pobreza, la desigualdad, la cohesión social en el territorio europeo.

Hasta ahora hemos hablado del tipo de información disponible de las variables con las que se puede trabajar, pero no de la forma de la relación funcional entre ellas. Esta forma funcional puede ser lineal o no lineal, cuando se habla de relaciones no lineales en los parámetros. Así, un ejemplo de modelo lineal sería $y_t = \alpha + \beta x_t + \varepsilon_t$ y un ejemplo de modelo no lineal sería $y_t = \alpha x_t^{\beta} \varepsilon_t$.

Este trabajo abordará varios tipos de modelos lineales y estáticos adecuados para datos de panel. En primer lugar, vamos a plantear un modelo de datos de panel, señalando sus principales características y, en función de éstas, clasificar los distintos modelos que pueden aparecer. Asimismo, presentaremos la principales ventajas e inconvenientes de este tipo de modelo. Seguidamente, trabajaremos con el modelo de datos panel más sencillo, aquel en el que aparecen dos periodos temporales, para presentar la principal ventaja de dichos modelos, a saber, controlar la heterogeneidad individual inobservable. A continuación, estudiaremos los dos tipos de modelos lineales y estáticos más importantes: el modelo de efectos fijos y el modelo de efectos aleatorios. Además, aplicaremos toda esta teoría, mediante el programa informático E-views, a los datos utilizados en el trabajo seminal de Baltagi y Griffin (1982) sobre la demanda de gasolina en la OCDE.

Finalmente extraeremos las conclusiones más importantes del trabajo realizado.

2. MODELO DE DATOS PANEL

Como dijimos anteriormente, en un conjunto de datos panel, se tienen observaciones de series temporales sobre una muestra de unidades individuales. Para una variable y_{it}, se supone que se tiene $i = 1, ..., N$ observaciones de corte transversal y $t = 1, ..., T$ observaciones temporales, donde i y t hacen referencia a los individuos y al periodo de tiempo, respectivamente.

Atendiendo al número de observaciones que se disponen de cada individuo, se puede diferenciar entre panel balanceado (mismo número de observaciones para todos los individuos) y no balanceado (cuando hay algún individuo con distinto número de observaciones).

Además, dependiendo del número de observaciones de los cortes transversales y temporales, se podrá diferenciar entre los siguientes tipos de datos de panel:

- Micropanel: si hay un reducido número de observaciones temporales para cada individuo y el número de individuos es muy elevado. Por lo tanto, para analizar las propiedades de los estimadores de los parámetros en los modelos que se planteen con este tipo de datos, se considerará que $N \to \infty$ y T es fijo.

- Macropanel: si hay un gran número de observaciones temporales y pocos individuos. Así, para analizar las propiedades que se planteen con este tipo de datos, se considerará que N es fijo y $T \to \infty$.

- Campo aleatorio: si se dispone de un elevado número de observaciones temporales y de individuos. Por consiguiente, para analizar las propiedades que se planteen con este tipo de datos, se considerará que $N \to \infty$ y $T \to \infty$.

En nuestro caso, trabajaremos, tanto teóricamente como de manera práctica con los modelos que se plantean para los datos de tipo micropanel.

El modelo más sencillo de datos de panel es una extensión del modelo de regresión lineal clásico, formulado de la siguiente manera:

$$y_{it} = x'_{it}\beta + \varepsilon_{it};\ i = 1...N;\ t = 1...T$$

donde $x'_{it} = (x_{1it}, x_{2it}, \ldots, x_{kit})$ es el vector k×1 formado por las observaciones de los k regresores del individuo i en el periodo t, β [1]es el vector de orden k×1 de parámetros de interés y ε_{it} la perturbación aleatoria.

La característica más sobresaliente de este modelo es que la observación de cada variable tienen dos subíndices: i, para la dimensión transversal, y t, para la dimensión temporal. En este caso, si las perturbaciones cumplen las hipótesis clásicas, el

[1]Sin pérdida de generalidad, se puede considerar que β incluye una ordenada en el origen o termino constante.

estimador por Mínimos de Cuadrados Ordinarios (MCO, también llamado estimador POOLED) será ELIO (estimador lineal, insesgado y óptimo).

En numerosas ocasiones, existen características de los individuos que afectan a la variable endógena, que no recogen los regresores y que permanecen constantes a lo largo del tiempo para cada individuo, por ejemplo: la habilidad, la inteligencia o la cultura en el caso de que i esté referida a personas. Por tanto, resulta adecuado incorporar esta heterogeneidad individual que, generalmente, es inobservable en el modelo de la siguiente forma:

$$y_{it} = x_{it}'\beta + \eta_i + \varepsilon_{it} \; i = 1...N; \; t = 1...T$$

donde η_i es la heterogeneidad individual inobservable y constante a lo largo del tiempo para cada uno de los individuos. Esto dificulta la estimación del modelo ya que si η_i está correlacionado con x_{it}' el estimador por MCO, en general, será sesgado.

Un ejemplo de modelo de datos de panel es el trabajo planteado por Mundlak (1961) y recogido también en los trabajos de Chamberlain (1984) y Arellano (1991) donde se analiza la función de producción Cobb-Douglas de un producto agrícola, en el que la heterogeneidad individual inobservable recoge la calidad del suelo. Además, en Arellano (1991) también se muestran otros ejemplos de modelos de datos de panel. Por ejemplo: El análisis de la oferta de trabajo intertemporal, donde la heterogeneidad individual inobservable está relacionada con la función de utilidad marginal de la riqueza; el estudio sobre los rendimientos de la educación, donde la heterogeneidad individual inobservable tiene que ver con la habilidad de cada individuo.

Como veremos posteriormente, trabajar con datos de panel sirve para controlar algunos efectos que influyen en la variable endógena y que no son recogidos por la variables explicativas del modelo. De manera que, los modelos con datos de panel controlan la heterogeneidad individual inobservable, es decir, la existencia de efectos latentes no observables específicos de cada agente encuestado, generalmente, constantes en el tiempo que inciden sobre el modo en que éste toma sus decisiones. En el caso de que esos efectos latentes no se recojan explícitamente en el modelo, se producirá un problema de variables omitidas, particularmente graves, cuando existe correlación entre los efectos latentes y los regresores del modelo.

Otra de las ventajas que presentan los modelos con datos de panel es que proporcionan una mayor cantidad de información, más variabilidad, menos colinealidad entre variables y una mayor precisión. Por último, los datos panel proporcionan 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 en unidades de corte transversal.

En contra, uno de los inconvenientes que aparecen al trabajar con datos de panel es que los individuos pueden abandonar la muestra, por lo que no es posible realizar su seguimiento a lo largo del tiempo. Además, pueden aparecer desequilibrios en la muestra, de manera que se tenga más información de algunos individuos que de otros, dando lugar a un panel no equilibrado o incompleto.

3. MODELO CON DOS PERIODOS TEMPORALES

El motivo principal para utilizar modelos con datos de panel datos de panel es la capacidad de controlar la heterogeneidad individual inobservable invariante en el tiempo. A continuación, vamos a analizar esta ventaja para un sencillo ejemplo donde k=1 y T=2. El modelo de datos panel se representa de la siguiente manera:

Si consideramos el modelo para el caso en el que t=1:

$$y_{i1} = \beta x_{1i1} + \eta_i + \varepsilon_{i1}.\ i = 1, ..., N. \tag{1}$$

donde β es el parámetro de interés que pretendemos estimar η_i recoge las características específicas de cada individuo (heterogeneidad individual inobservable), que suponemos permanecen constantes en el tiempo y ε_{i1} es la perturbación aleatoria del modelo que suponemos cumple las hipótesis clásicas. se pueden plantear dos casos:

- Si η_i es observable, β puede ser estimado por MCO y el estimador es insesgado y consistente.
- Si η_i no es observable, como ocurre en la mayoría de los casos, estaríamos cometiendo un error por omisión de variable relevante. En este caso, nos podemos encontrar con las siguientes dos nuevas situaciones:

  1. Si x_{i1} y η_i están incorrelacionados, el estimador de MCO, al menos será consistente.
  2. Si x_{i1} y η_i están correlacionados, el estimador de MCO sería inconsistente. En este caso, habría que utilizar el método de variables instrumentales que proporciona estimadores consistentes, siempre que exista un instrumento que esté incorrelacionado con ε_{i1} y η_i, pero correlacionado con x_{i1}.

Si ninguna de estas dos opciones es viable, disponer de un panel de datos, supone una alternativa para obtener estimadores al menos consistentes de β . A continuación, se muestra que trabajar con datos de panel permite plantear un modelo que conserva los parámetros de interés y en el que se elimina la heterogeneidad individual inobservable

(η_i), así si tenemos una nueva observación de las variables para los mismos individuos en un segundo periodo de tiempo, T=2, es decir, tenemos y_{i2} y x_{i2} , tal que:

$$y_{i2} = \beta x_{i2} + \eta_i + \varepsilon_{i2} \tag{2}$$

donde ε_{i1} y ε_{i2} cumplen $E(\varepsilon_{it} \mid x_{i1}, x_{i2}, \eta_i) = 0$ con t=1,2. Al tomar primeras diferencias:

$$(y_{i2} - y_{i1}) = \beta\,(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \tag{3}$$

En todo caso podemos observar que, aunque η_i no sea observable, la heterogeneidad individual desparece del modelo y se puede obtener un estimador consistente de β a partir de MCO, es decir,

$$\hat{\beta} = \frac{Cov(\triangle x_{i2}, \triangle y_{i2})}{Var(\triangle x_{i2})}, \tag{4}$$

siendo $\triangle x_{i2} = (x_{i2} - x_{i1})$ y $\triangle y_{i2} = (y_{i2} - y_{i1})$.

Esta ventaja que se ha visto en el caso más sencillo, se puede extender cuando se dispone de más observaciones y regresores. Esto es lo que abordaremos en la siguiente sección.

4. MODELOS DE EFECTOS FIJOS Y MODELO EFECTOS ALEATORIOS

Como hemos mencionado anteriormente, el modelo con el que vamos a trabajar es un modelo de regresión lineal que incluye efectos individuales inobservables, este modelo lo podemos representar de forma escalar o matricial de la siguiente manera:

De manera escalar:

$$y_{it} = x'_{it}\beta + \eta_i + \varepsilon_{it} \quad i = 1, 2, ..., N; \quad t = 1, 2, ..., T. \tag{5}$$

De forma matricial:

$$Y = X\beta + C\eta + \varepsilon \tag{6}$$

donde Y es un vector de (NTx1) que contiene la información de todos los individuos en todos los periodos del tiempo, es decir,

$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Y_N \end{bmatrix} \text{, siendo } Y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix};$$

X es una matriz de orden (NT×k) de las observaciones de las k variables explicativas en todos los individuos y todos los periodos de tiempo, es decir,

$$X = \begin{bmatrix} x_{111} & \dots & x_{k11} \\ \vdots & \ddots & \vdots \\ x_{11T} & \dots & x_{k1T} \\ x_{121} & \dots & x_{k21} \\ \vdots & \dots & \vdots \\ x_{12T} & \dots & x_{k2T} \\ x_{1N1} & \dots & x_{kN1} \\ \vdots & \dots & \vdots \\ x_{1NT} & \dots & x_{kNT} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ X_N \end{bmatrix} \text{siendo } X_i = \begin{bmatrix} x_{1i1} & \dots & x_{ki1} \\ \vdots & \ddots & \vdots \\ x_{1iT} & \dots & x_{kiT} \end{bmatrix};$$

β es un vector de orden (k×1) que contiene los parámetros de interés, es decir,

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix};$$

C es una matriz de orden (NT×N) formada por 0 y 1, de tal forma que cada columna recoge una variable que toma el valor 1 dependiendo del individuo en el que nos encontremos. Así, la primera columna consta de T unos que corresponden con las observaciones del primer individuo y cero en el resto de individuos. La segunda columna representa al segundo individuo y toma el valor uno en sus observaciones y cero en el resto.

$$C = I_N \otimes \iota_T = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & 1 & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \dots & \vdots & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{bmatrix};^2$$

η es un vector de orden (N×1) que contiene la heterogeneidad individual inobservable, es decir,$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \vdots \\ \eta_N \end{bmatrix}$;

ε es un vector de orden (NT×1) que contiene las perturbaciones aleatorias de todos los individuos en todos los periodos de tiempo, es decir,

$$\varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1T} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2T} \\ \vdots \\ \varepsilon_{N1} \\ \vdots \\ \varepsilon_{NT} \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_N \end{bmatrix}, \text{ siendo } \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}.$$

En el modelo planteado anteriormente, si η_i , la heterogeneidad transversal inobservable de los i individuos, fuera omitida se estaría cometiendo un error por omisión. La discusión crucial radica en si los efectos de η_i están correlacionados o no con las variables explicativas observables x'_{it}. Si η_i está correlacionado con x'_{it} nos encontramos en un modelo de efectos fijos. Por el contrario, si los efectos latentes η_i no están correlacionados con x_{it} estamos en el modelo de efectos aleatorios.

[2]El producto de Kronecker, $\otimes$, se utiliza para multiplicar dos matrices, en este caso multiplica la matriz indentidad, I, por el primer uno de ι_T. Depsues otra vez la matriz indentidad por el segundo uno de la matriz ι_T y asi N veces.

Los supuestos bajo los que se construyen estos modelos asumen que ε_{it} cumple la hipótesis clásicas, es decir,

- $E(\varepsilon_{it}) = 0$
- La no correlación entre perturbaciones de cada uno de los grupos y la no correlación temporal, $cov(\varepsilon_{it}, \varepsilon_{jt}) = 0, \forall i \neq j, \forall t \neq s$
- Perturbaciones homocedásticas, $var(\varepsilon_{it}) = \sigma_{\varepsilon}^2$.

En este trabajo se presenta dos estimadores consistentes de los parámetros de interés en presencia de una variable explicativa latente no observable. Hay que señalar que los supuestos anteriormente mencionados son fácilmente violables en cuyo caso se requerirán otras técnicas de estimación que no se abordan en este trabajo.

En el caso de que η_i este correlacionado con los regresores, el modelo de efectos fijos, los η_i son tratados como un conjunto de N coeficientes adicionales que se pueden estimar junto con los parámetros β.

En el caso de que η_i no este correlacionado con los regresores, el modelo de efectos aleatorios, se supone que η_i es una variable aleatoria no observable independiente de x'_{it}, y por tanto, pasa a formar parte de un nuevo término de error, ω_{it} , que tiene dos componentes:

$$\omega_{it} = \eta_i + \varepsilon_{it} \tag{7}$$

Por este motivo, a estos modelos también los podemos llamar como modelos de errores compuestos.

A continuación, procederemos a estudiar en profundidad el modelo de efectos fijos y, posteriormente, el modelo de efectos aleatorios.

4.1. Modelo de efectos fijos

Este modelo supone que existe heterogeneidad transversal inobservable, constante en el tiempo y correlacionada con los regresores. En este caso, η_i es tratado como un parámetro adicional que representa un término independiente para cada individuo del panel[3]. Este modelo se denomina normalmente como el modelo de Mínimos Cuadrados de Variables Ficticias (MCVF), aunque hay que señalar que la parte del nombre «Mínimos Cuadrados» se refiere a la técnica que se utiliza habitualmente para estimarlo, no al modelo como tal. A continuación, presentamos un estimador consistente de β , que también puede obtenerse a partir de transformaciones del modelo, que eliminan los efectos individuales.

A partir de la expresión de la forma matricial de modelo que hemos presentado anteriormente(6),

$$Y = C\eta + X\beta + \varepsilon$$

se puede estimar β, los parámetros de interés, por MCO, utilizando los resultados de la regresión particionada (Véase, por ejemplo, Greene (2007, pp. 535-560)). Así, el estimador de β por MCO se puede escribir de la forma

$$\widehat{\beta} = \left(X'\overline{M}X\right)^{-1} X'\overline{M}Y \tag{8}$$

$$\overline{M} = \begin{bmatrix} M^0 & 0 & \dots & 0 \\ 0 & M^0 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & M^0 \end{bmatrix}$$

donde $\overline{M} = I_{NT} - C\left(C'C\right)C' = I_N \otimes M^0$ con $M^0 = I_T - \frac{\iota_T \iota_T'}{T}$. $\overline{M}$ es una matriz simétrica e idempotente y $\overline{M} \times C = 0$.

Este estimador se puede obtener de forma alternativa a partir de una transformación del modelo (6). Así, si premultiplicamos el modelo (6) por $\overline{M}$, obtenemos un modelo que conserva los parámetros de interés β y elimina los efectos individuales. Este modelo se puede expresar como:

$$\tilde{Y} = \tilde{X}\beta + \widetilde{\varepsilon}$$

donde $\tilde{Y} = \overline{M}Y$, $\tilde{X} = \overline{M}X$ y $\widetilde{\varepsilon} = \bar{M}\varepsilon$. El estimador por MCO de β en este

[3]En el caso de que β contenga un término constante habría que imponer la restricción $\sum \eta_i = 0$. En este caso, η_i representa la diferencia en el termino independiente atribuida al individuo i

modelo sería:

$$\tilde{\beta} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{Y} \tag{9}$$

Dado que la matriz $\overline{M}$ es idempotente y simetrica la expresión (8) y (9) son equivalentes.

Se puede comprobar que los elementos de $\tilde{X}$ e $\tilde{Y}$ son las desviaciones de las observaciones originales con respecto a las medias temporales de cada individuo. Así, por ejemplo los elementos de $\tilde{Y}$, $\tilde{y_{it}}$, serían de la forma

$$\tilde{y_{it}} = y_{it} - \bar{y}_i .\ i = 1, \ldots, N \ y \ t = 1, \ldots, T.$$

con $\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}$

El estimador de $\hat{\beta}$ y $\widetilde{\beta}$ es llamado **estimador intragrupos.** La consistencia de este estimador intragrupos, cuando T es fijo y $N \to \infty$, no depende de la especificación de η_i, porque los efectos siempre se eliminan debido a la transformación del modelo. El requisito que garantiza la consistencia de este estimador es que las variables x'_{it} sean estrictamente exógenas, con respecto a ε_{it}, es decir, $E\left(\varepsilon_{it} \mid x'_{it}\right) = 0$, $\forall t = 1, .., T$.

Sin embargo, este procedimiento tiene una limitación, y es que si las variables x'_{it} son invariantes en el tiempo, al calcular las desviaciones respecto a las medias temporales, desaparecerán del modelo, por lo que no podrán estimarse los parámetros correspondientes.

4.2. Modelo de efectos aleatorios

En el modelo de efectos aleatorios, se supone que η_i recoge los efectos no observables de cada agente que suponemos constantes en el tiempo y que están incorrelacionadas con los regresores.

Partiendo del modelo escrito en su forma escalar (5):

$$y_{it} = x'_{it}\beta + \eta_i + \varepsilon_{it} \quad i = 1, 2, ..., N; \quad t = 1, 2, ..., T.$$

Lo característico del modelo de efectos aleatorios es que η_i pasa a formar parte del término error del modelo (5):

$$y_{it} = x'_{it}\beta + \omega_{it}. \quad i = 1, 2, ..., N. \quad t = 1, 2, ..., T. \tag{10}$$

donde el término error ω_{it} se ajusta a la ecuación (6), es decir,$\omega_{it} = \eta_i + \varepsilon_{it}$

Para la estimación del modelo (10), se trabaja bajo los siguientes supuestos:

- $E\left(\varepsilon_{it}\right) = E\left(\eta_i\right) = E\left(\eta_i \varepsilon_{it}\right) = 0 \ \forall it$

- $E(x_{it} \mid \eta_i) = E(x_{it} \mid \varepsilon_{it}) = 0 \, \forall it$
- $E(\eta_i \eta_j) = \begin{cases} \sigma_\eta^2 \; Si \; t = s & i = j \\ 0 & i \neq j \end{cases}$
- $E(\varepsilon_{it} \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 \; Si \; t = s & i = j \\ 0 \quad t \neq s & i \neq j \end{cases}$

La varianza del nuevo término error, $\omega = \eta_i + \varepsilon_{it}$, es de la forma: $E[\omega_{it}^2] = \sigma_\omega^2 = \sigma_\eta^2 + \sigma_\varepsilon^2; \forall i, t$. La covarianza para el término error para un mismo individuo sería, $Cov(\omega_{it}\omega_{is}) = \sigma_\eta^2$, por lo que la matriz de varianzas y covarianzas de los T valores correspondientes al i-ésimo individuo, $\omega_i' = [\omega_{i1}, \omega_{i2}, \ldots, \omega_{iT}]$, es una matriz (T×T), de la forma:

$$\Omega = E[\omega_i \omega_i'] = \begin{bmatrix} \sigma_\eta^2 + \sigma_\varepsilon^2 & \sigma_\eta^2 & \ldots & \ldots & \sigma_\eta^2 \\ \sigma_\eta^2 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ldots & \ddots & \ddots & \sigma_\eta^2 \\ \sigma_\eta^2 & \ldots & \ldots & \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\varepsilon^2 \end{bmatrix} = \sigma_\varepsilon^2 I + \sigma_\eta^2 i_T i_T'$$

Esta matriz muestra que la correlación entre dos observaciones procedentes del mismo individuo es constante y no desaparece con el tiempo, debido a la presencia del efecto individual η_i, σ_η^2.

De manera que, la matriz de varianzas y covarianzas de todas las perturbaciones del modelo completo, $\omega = [\omega_1, \omega_2, \ldots, \omega_N]$, es diagonal a bloques y tiene la forma:

$$E[\omega\omega'] = V = I_N \otimes \Omega = \begin{bmatrix} \Omega & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Omega \end{bmatrix}.$$

Si el modelo (10) se estima por MCO, los estimadores serían consistentes pero no eficientes al no considerar la correlación entre las observaciones correspondientes al mismo individuo. Por este motivo, el método de estimación eficiente sería Mínimos Cuadrados Generalizados (MCG), es decir

$$\widehat{\beta}_{MCG} = \left[X'V^{-1}X\right]^{-1} \left[X'V^{-1}Y\right] \tag{11}$$

Generalmente los elementos de Ω son desconocidos. Por lo tanto, el estimador por MCG no se puede obtener, así que, previamente necesitamos estimar σ_ε^2 y σ_η^2, y obtener el estimador por MCGF.

$$\widehat{\beta}_{MCGF} = \left[X'V\hat{^{-1}} - 1X\right]^{-1} \left[X'V\hat{^{-1}}Y\right]$$

Balestra y Nervole (1966) porponen un forma de estimar $\widehat{\sigma_\varepsilon^2}$ y $\widehat{\sigma_\eta^2}$ es la siguiente:

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widetilde{y}_{it} - \hat{\beta}' \tilde{x}_{it} \right)^2$$

$$\widehat{\sigma_\eta^2} = \frac{1}{N} \sum_{i=1}^{N} \left(\overline{y}_i - \hat{\beta}' \overline{x}_i \right)^2 - \frac{1}{T} \widehat{\sigma_\varepsilon^2}$$

donde$\widetilde{y}_{it} = y_{it} - \overline{y}_i$; $\tilde{x}_{it} = x_{it} - \overline{x}_i$ y $\hat{\beta}$ estimador intragrupos.

En este caso el estimador por MCGF es consistente y asintóticamente infeciente.

4.3. Elección entre efectos fijos y aleatorios

En el caso de que haya incorrelación entre η_i y x'_{it} será conveniente utilizar el modelo de efectos aleatorios, debido a que el estimador de Balestra y Nerlove (BN o MCGF) es consistente y eficiente, mientras que el estimador intragrupos es consistente, pero no eficiente. Por otro lado, si η_i y x'_{it} están correlacionados será conveniente utilizar el modelo de efectos fijos, ya que el estimador intragrupos es consistente y eficiente, pero el estimador de MCG no es consistente. Si existe mucha diferencia entre los valores de ambas estimaciones será un indicio que hay efectos individuales no observados, si no la hay será un indicio de que ese, modelo es de efectos aleatorios.

Una forma de juzgar si estas diferencias son estadísticamente significativas es mediante el contraste de Hausman (1978). En esta prueba, se contrasta bajo la hipótesis nula la ausencia de correlación entre los efectos latentes, η_i, y las variables explicativas, es decir,

$$H_0 : \; E\left(\eta_i \mid x_{11}, \ldots, x_{it}\right) = 0$$

$$H_1 : \; E\left(\eta_i \mid x_{11}, \ldots, x_{it}\right) \neq 0$$

El estadístico de Hausman, se basa en la comparación directa entre el estimador intragrupos $\widehat{\beta}$ (modelo de efectos fijos) y el estimador de Balestra Nervole $\widehat{\beta}_{MCGF}$ (modelo de efectos aleatorios). Bajo la hipótesis nula el estimador MCGF es consistente y eficiente, siendo inconsistente bajo la hipótesis alternativa, es decir, cuando las variables latentes están correlacionadas con las variables explicativas. Bajo la hipótesis nula y alternativa el estimador intragrupos es consistente. El estadístico de contraste utilizado para el Contraste de Hausman, es:

$$h = \left(\widehat{\beta}_{MCGF} - \widehat{\beta} \right)' \left(\Sigma_{\hat{\beta}} - \Sigma_{\beta_{M\hat{C}GF}} \right)^{-1} \left(\widehat{\beta}_{MCGF} - \widehat{\beta} \right) \rightsquigarrow \chi_k^2 \qquad (12)$$

siendo $\Sigma_{\hat{\beta}}$ la matriz de varianzas y covarianzas correspondientes a un modelo de efectos fijos y $\Sigma_{\hat{\beta_{MCGF}}}$la matriz de varianzas y covarianzas correspondientes a un modelo de efectos aleatorios.

Donde el estadístico h, bajo la hipótesis nula, se distribuye asintóticamente como una χ^2 con k grados de libertad. Si h es pequeño, no se rechaza H_0 de ausencia de correlaciones, rechazándose en caso contrario.

Si no se rechaza H_0, debe utilizarse el estimador de MCGF que es óptimo en tales condiciones. Si se rechaza H_0 es preferible utilizar el estimador intragrupos que, al menos, es consistente.

Para decidir cual de los modelos, el de efectos fijos o el de efectos aleatorios, es más adecuado a la hora de ajustar a una situación real, no solo debe tomarse dicha decisión bansándose en el resultado de un contraste de especificación, sino que debe adaptarse a las características que se plantean a la hora de elegir uno u otro modelo, es la existencia de correlaciones entre efectos individuales η_i y los regresores x'_{it}.

5. UN EJEMPLO

Hasta ahora este trabajo había abordado los datos de panel de forma teórica. Con el objetivo de obtener una visión más amplia de este tipo de datos se va a proceder a analizar un caso práctico sobre la demanda de gasolina en la OCDE desarrollado por Baltagi y Griffin. En este estudio el programa informático que se va utilizar es «E-views8», en el Anexo se recoge de manera detallada la forma de analizar datos de panel en la versión 8 de E-views.

EL modelo que vamos a plantear para estudiar la demanda de gasolina en la OCDE es:

$$ln\left(GAS/CAR\right)_{it} = \eta_i+\beta_1 ln\left(Y/N\right)_{it}+\beta_2 ln\left(Pmg/Pgdp\right)_{it}+\beta_3 ln\left(CAR/N\right)_{it}+\beta_4 ln\left(\varepsilon_{it}\right) \tag{13}$$

$$i = 1, \ldots, 18;\ t = 1960, \ldots, 1978.$$

donde

(GAS/CAR)= consumo de gasolina por coche;

(Y/N)= ingresos per cápita;

(PMG/PGP)= precio de la gasolina;

(CAR/N)= stock de coches per cápita;

En la tabla 5.1, se muestran los resultados de las estimación por MCO de modelo para cada país. En general, la estimación de los parámetros de los distintos países proporciona resultados bastantes coherentes. Las elasticidades de la renta per cápita sobre la demanda de gasolina son positivas sobre la demanda de gasolina en todos los países, menos en Japón, España y Suecia que tiene un efecto negativo. Por el conrario, las elasticidades del precio de la gasolina son negativas sobre la demanda de esta. El incremento de coches per cápita genera una tendencia a reducir el uso del coche. A pesar de que la estimación presenta resultados muy coherentes en signo y en valor, muchos de ellos no son significativos. Por ejemplo, en diez países(Dinamarca, Grecia, Irlanda, Italia, Japón, Holanda, España, Suecia, Turquía, USA), la renta per cápita no tiene un efecto significativo y, en el caso del precio de la gasolina, son siete países en los que no es significativo. No parece logico que muchas de ellas no sean significativas por lo que puede deberse a un problema de omisión de heterogeneidad individual inobservable. Por eso, dadas las características de los modelos de datos de panel, podemos mejorar estos resultados.

Tabla 5.1: resultados de la estimación del modelo planteado para cada uno de los países.

País	ln(Y/N)	ln(PMG/PGP)	ln(CAR/N)	Const.	$\bar{R}^2$	S.E
Austria	0.760	-0.793	-0.519	3.726	0.680	0.039
Bélgica	0.845	-0.041	-0.673	3.041	0.890	0.034
Canadá	0.392	-0.362	-0.438	3.125	0.791	0.011
Dinamarca	0.092	-0.137	-0.517	0.236	0.963	0.030
Francia	1.119	-0.194	-0.844	3.191	0.735	0.025
Alemania	0.401	-0.167	-0.222	4.263	0.492	0.017
Grecia	0.594	-0.343	-0.473	3.693	0.914	0.074
Irlanda	0.353	-0.099	-0.181	4.822	0.190	0.039
Italia	0.117	-0.371	-0.356	1.273	0.977	0.032
Japón	-0.048	-0.144	-0.560	-1.219	0.998	0.025
Holanda	0.362	-0.402	-0.618	0.623	0.980	0.040
Noruega	0.801	-0.230	-0.655	2.913	0.935	0.031
España	-0.830	-0.078	-0.101	-1.561	0.967	0.057
Suecia	-0.710	-0.616	0.039	-2.886	0.478	0.026
Suiza	1.067	-0.404	-0.617	4.925	0.923	0.028
Turquía	0.318	-0.260	-0.602	0.479	0.929	0.087
U.K	0.560	-0.061	-0.332	4.487	0.684	0.026
U.S.A	0.107	-0.276	-0.095	4.328	0.452	0.016

Fuente: Elaboración propia.

A continuación, abordaremos la estimación de la demanda de gasolina en la OCDE de forma conjunta y no individualmente, a partir de la estimación de el modelo de efectos fijos y el modelo de efectos aleatorios.

Finalmente el contraste de Hausman nos servirá para determinar el modelo más correcto.

La tabla 5.2, donde el p-valor aparece entre parentesis, muestra los resultados de la estimación de ambos modelos. Se observa que en este caso los signos de las variables son coherentes y todas son significativas tanto individual como conjuntamente, algo que no pasaba en la estimación individual por países. Aunque todas las variables comparten el mismo signo, la magnitud de las elasticidades no es la misma en los dos modelos.

Tabla 5.2: resultados estimación de modelo usando la estimación intragrupos y Baslestra Nerlove

Técnica estimación	ln(Y/N)	ln(PMG/PGP)	ln(CAR/N)	Const.	S.E	p-valor Fstatistic
Intragrupos	0.662 (0.000)	-0.321 (0.000)	-0.640 (0.000)	-	0.092	0.000
Balestra y Nerlove	0.554 (0.000)	-0.420 (0.000)	-0.606 (0.000)	1.996 (0.000)	0.095	0.000

Fuente: Elaboración propia.

Aunque todas la variables de las dos estimaciones comparten el mismo signo, . Se observa que para la estimación intragrupos de las elasticidades de la renta per cápita y número de coches per cápita son mayores, sin embargo el precio de la gasolina la elasticidad es algo mayor en la estimación Balestra y Nerlove.

Con el fin de evaluar si estas diferencias son estadísticamente significativas, se realiza el contraste de Hausman. En nuestro caso, la $\chi^2 = 26{,}495$ y el p-valor=0.000. Por tanto, podemos decir que se rechaza la hipótesis nula, ausencia de correlaciones entre la heterogeneidad individual inobservable y los regresores. Así que, el estimador de efectos fijos será mas adecuado que el estimador de efectos aleatorios.

6. CONCLUSIONES

Este trabajo tiene por objetivo realizar una introducción a los modelos de datos panel, mostrando la importancia de poder estimar modelos en los que aparece una heterogeneidad individual inobservable. Para ello hemos partido del caso más sencillo, con dos periodos de tiempo, y luego generalizarlo a más periodos. Con esta primera introducción, se muestra la importancia del análisis de datos panel y cómo obtener un estimador consistente en presencia de la heterogeneidad. Posteriormente, dentro de los datos panel, hemos distinguido entre el modelo de efectos fijos y el modelo de efectos aleatorios. Para ello, hemos establecido los supuestos bajo los que se construye cada uno de esos modelos y el método de estimación en cada caso, diferenciando si existe o no correlación entre la heterogeneidad individual inobservables y los regresores. Terminamos haciendo una comparación entre ambos métodos, estudiando las ventajas e inconvenientes de cada uno y cuál utilizar en cada situación. Para poder tomar esta decisión, nos apoyamos en el contraste de Hausman, que nos permite determinar si existe o no correlación entre los efectos individuales y los regresores. En el caso de que exista tal correlación, es conveniente utilizar el modelo de efectos fijos, mientras que si no existe dicha correlación es preferible el modelo de efectos aleatorios. Esta preferencia depende de las propiedades que presentan los métodos de estimación en cada situación.

Referencias

[1] Amparo, S y Guadalupe, S (2004): *Econometría para Económicas*. Madrid. http://www.uv.es/~sancho/panel.pdf. Última visualización: 30/06/2016

[2] Arellano, M y Bover, O. (1990)*: La Econometría de Datos de Panel*. Editorial London School of Economics, Londres.

[3] Arellano.M. (1991): *Introducción al Análisis Econométrico con Datos de Panel*. Editorial Banco de España, Madrid.

[4] Arellano, M.(2003): *Panel Data Econometrics.* Editorial Oxford University Expess, London.

[5] Baltagi, B y Griffrin, J (1982): "Gasoline Demand in the OCDE", *European Economic Review,* 22, pp.117-137.

[6] Chamberlain, G. (1984): *Data Panel*. Editorial Univerisity of Wisconsin-Madison and NBER.

[7] Greene.W.H. *(2007): Análisis econometrico*. cap14, pp. 535-560. Editorial Prentice Hall.

[8] Gujarati.D. y Porter.D.(2010): *Econometría*. Cap16, pp. 591-61. Editorial McGrawHill, Mexico.

[9] Mahía, R. (2000): *Introducción a la especificación y estimación de modelos con datos de panel.* https://www.uam.es/personal_pdi/economicas/rmc/investiga/introdat.pdf. Última visualización: 30/06/2016.

[10] Mauricio, M y Evelyn, S. (2000): *La técnica de datos de panel una guía para su uso e interpreteción*. Editorial Banco central de Costa Rica, Costa Rica.

[11] Johnston, J y Dinardo, J. (1997): *Métodos de econometría*. Cap12, pp. 445-470. Editorial Vinces Vives, Barcelona.

[12] Wooldridge, J.M(2009): *Introducción a la econometría un enfoque moderno*. Cap13, pp. 473-510. Editorial South-Western Cenage Learning, Londres.

[13] Wooldridge, J.M(2009): *Introducción a la econometría un enfoque moderno*. Cap14, pp.510-538. Editorial South-Western Cenage Learning, Londres

[14] Web oficial: http://www.eviews.com/home.html Última visita el 28/06/2016.

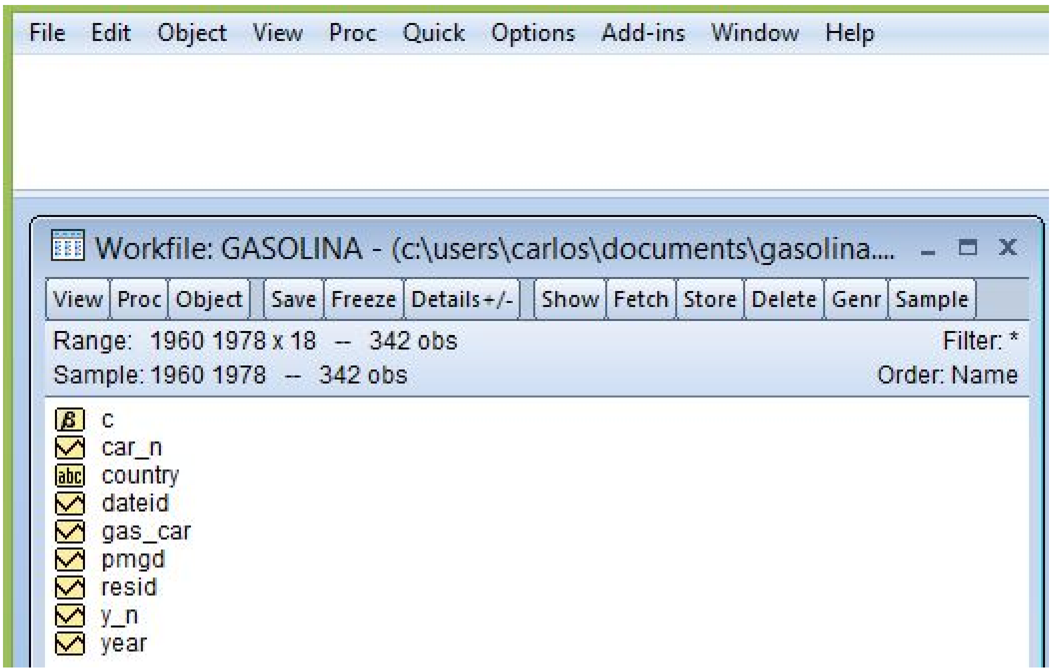
Anexo I

En este anexo se recoge la manera de proceder para analizar datos de panel con Eviews8.

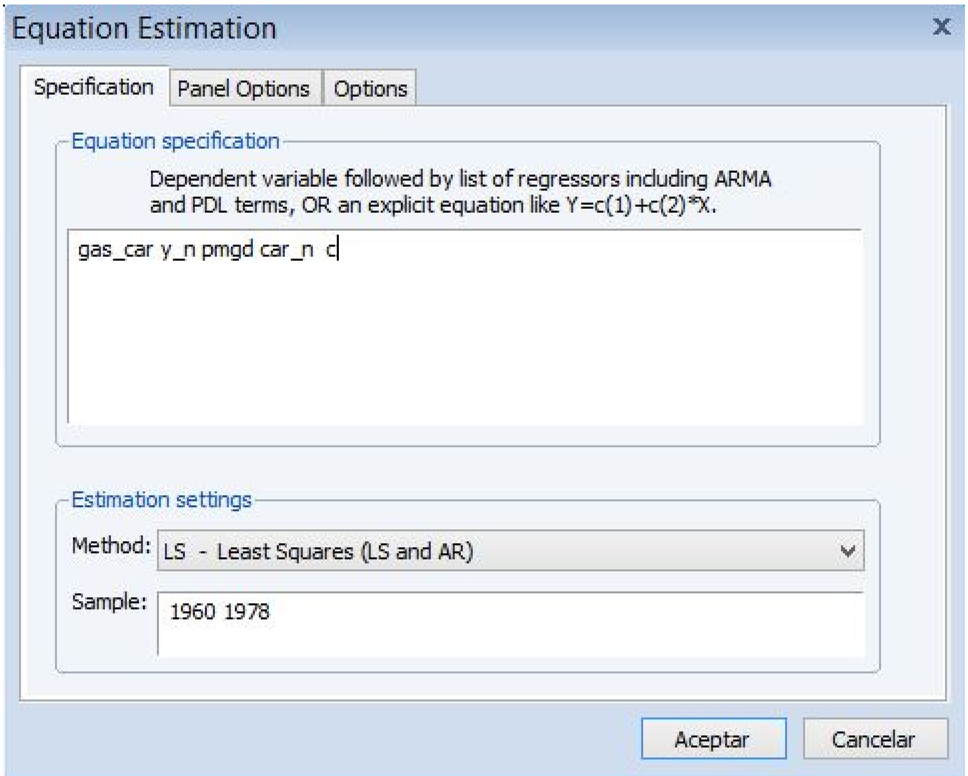
- Disponer de una hoja de calculo excel con toda la información ordenada,

country	year	gas/car	y/n	pmgd	car/n
AUSTRIA	1960	4,173244	-6,47428	-0,33455	-9,76684
AUSTRIA	1961	4,100989	-6,42601	-0,35133	-9,60862
AUSTRIA	1962	4,073176	-6,40731	-0,37952	-9,45726
AUSTRIA	1963	4,059509	-6,37068	-0,41425	-9,34315
AUSTRIA	1964	4,037689	-6,32225	-0,44534	-9,23774
AUSTRIA	1965	4,033983	-6,29467	-0,49706	-9,1239
AUSTRIA	1966	4,047536	-6,25255	-0,46684	-9,01982
AUSTRIA	1967	4,052911	-6,23458	-0,50588	-8,9344
AUSTRIA	1968	4,045507	-6,20689	-0,52241	-8,84797
AUSTRIA	1969	4,046355	-6,15314	-0,55911	-8,78869
AUSTRIA	1970	4,080888	-6,08171	-0,59656	-8,7282
AUSTRIA	1971	4,10672	-6,04363	-0,65446	-8,6359
AUSTRIA	1972	4,128018	-5,98105	-0,59633	-8,53834
AUSTRIA	1973	4,19938	-5,89515	-0,59445	-8,48729
AUSTRIA	1974	4,018496	-5,85238	-0,46603	-8,4304
AUSTRIA	1975	4,029018	-5,86936	-0,45414	-8,38281
AUSTRIA	1976	3,985412	-5,8117	-0,50008	-8,32223
AUSTRIA	1977	3,931676	-5,83329	-0,42192	-8,24956
AUSTRIA	1978	3,92275	-5,76202	-0,4696	-8,21104
BELGIUM	1960	4,164016	-6,21509	-0,16571	-9,40553

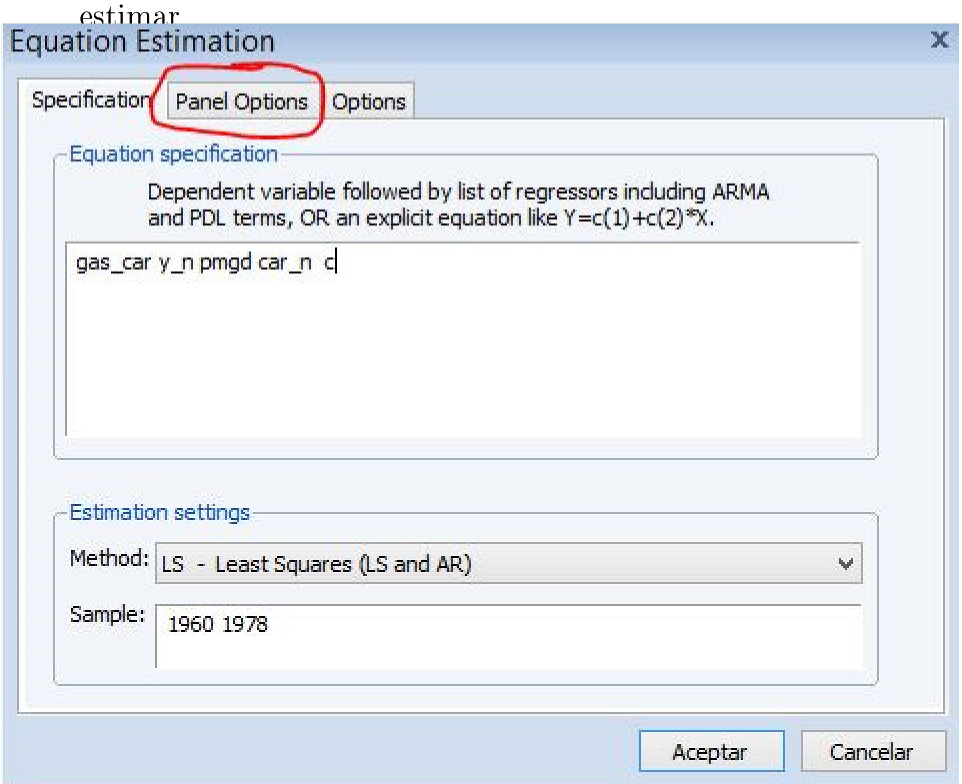
- Para trasladar los datos a la aplicación E-views la forma utilizada ha sido: situar el cursor del ratón sobre la hoja de calculo excel que se desea abrir, pulsar el botón derecho del ratón abrir con→eviews8. E-views automáticamente te reconoce las variable y se crean por defecto, se pulsa el botón "finalizar".



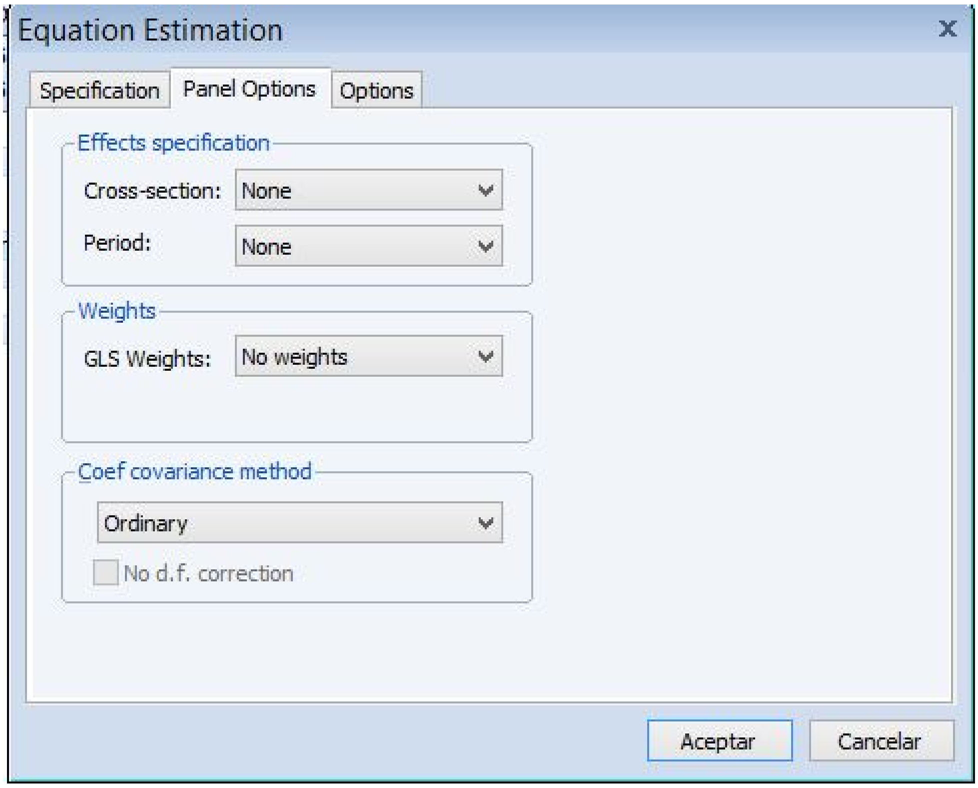
- Seguidamente se procede a estimar la ecuación, Quick→ estimate equation, se abrirá una pestaña donde se introducen la variables del modelo, primero la endógena y después el resto.



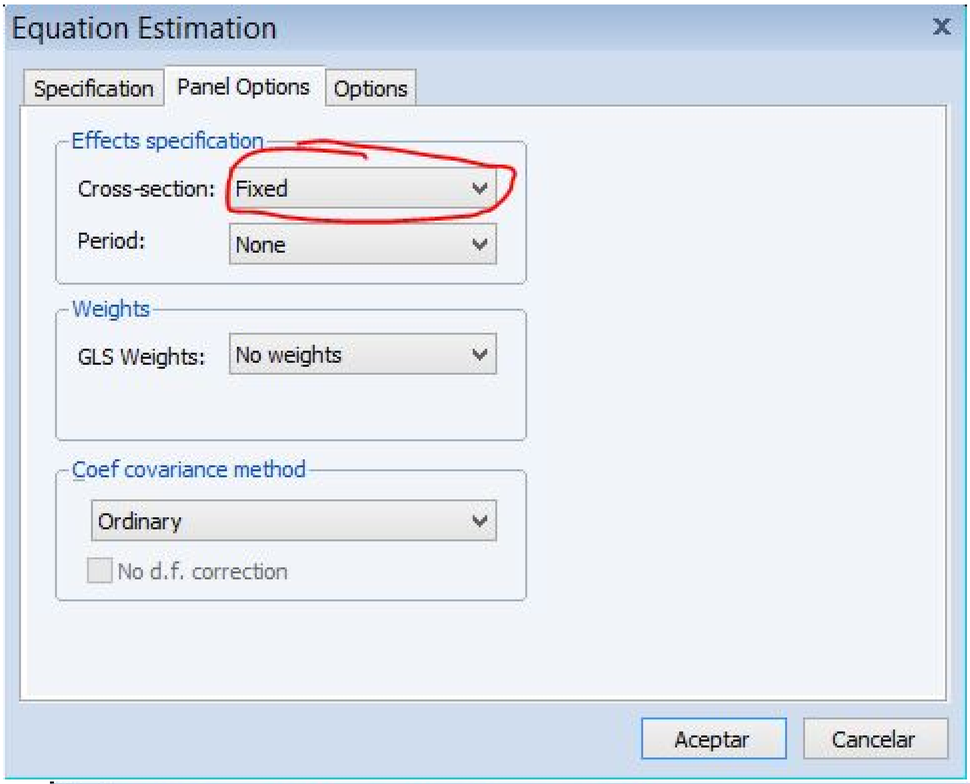
- A continuación, se pincha la pestaña de "panel options" que nos llevará a otra ventana, donde seleccionaremos el modelo (fijos o aleatorios) que queremos estimar

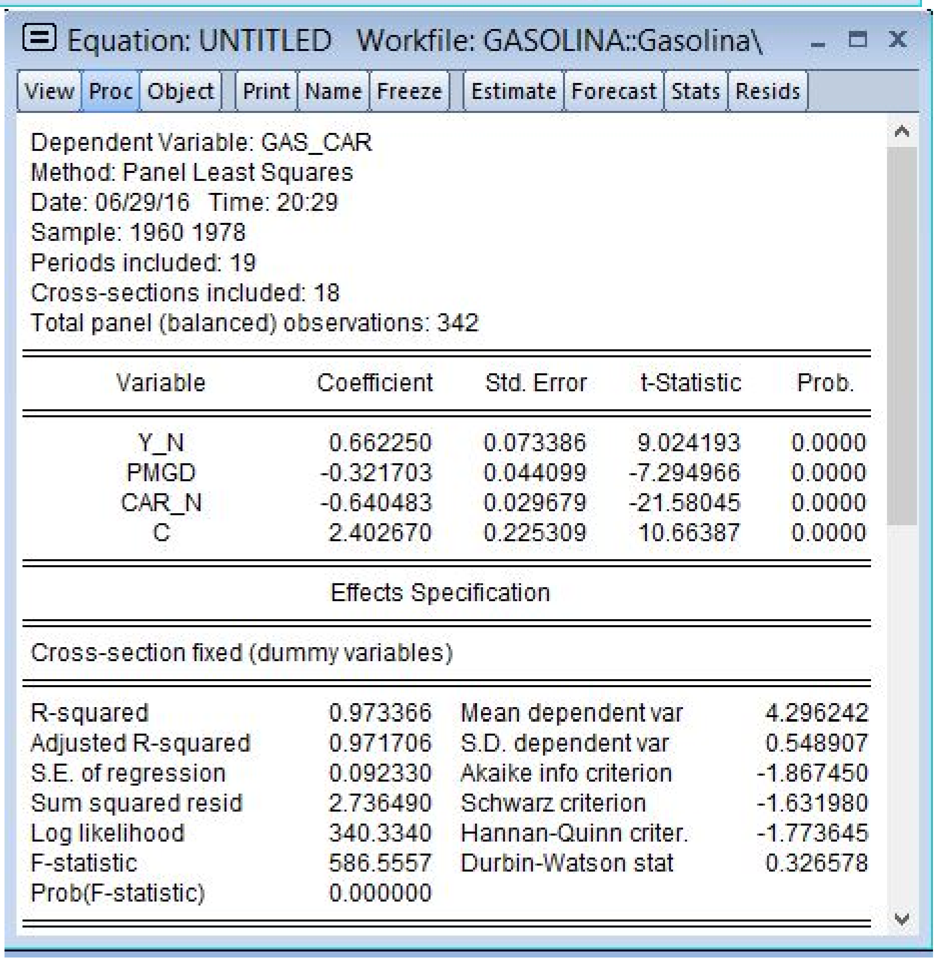


- En "panel options" se encuentra un apartado llamado "effects specification" donde hay dos opciones "cross section" y "period", el que vamos a utilizar es "corss section". Se despliega "cross section" se verá como aparecen otras dos opciones: "fixed" y "random", que hacen referencia a efectos fijos y aleatorios, respectivamente.

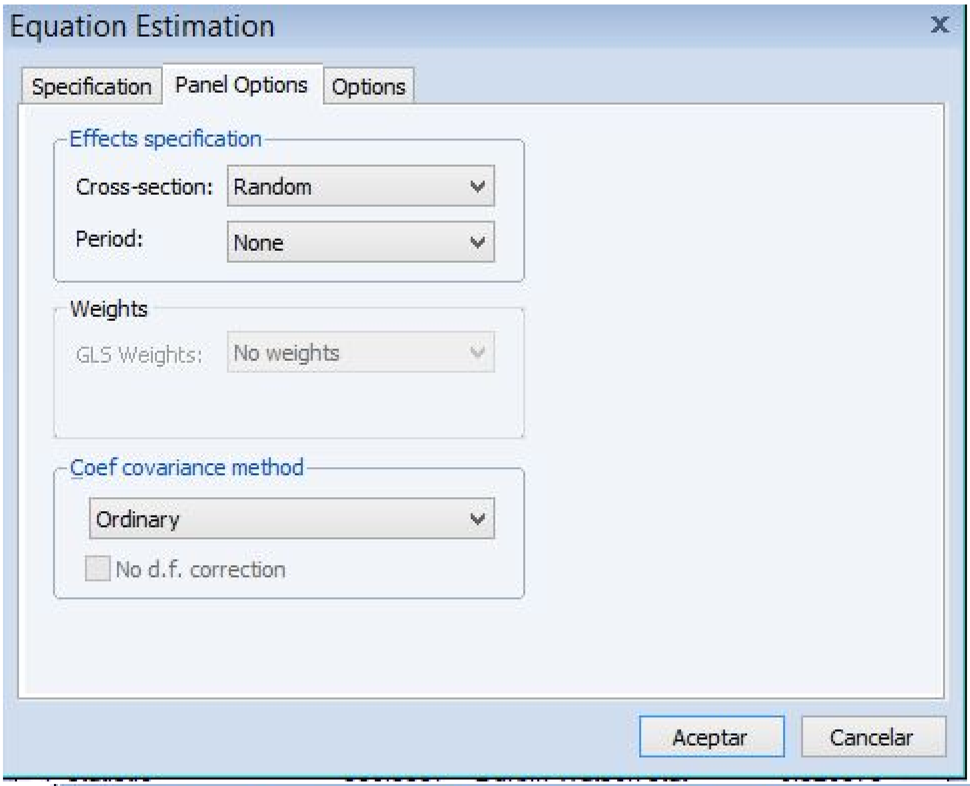


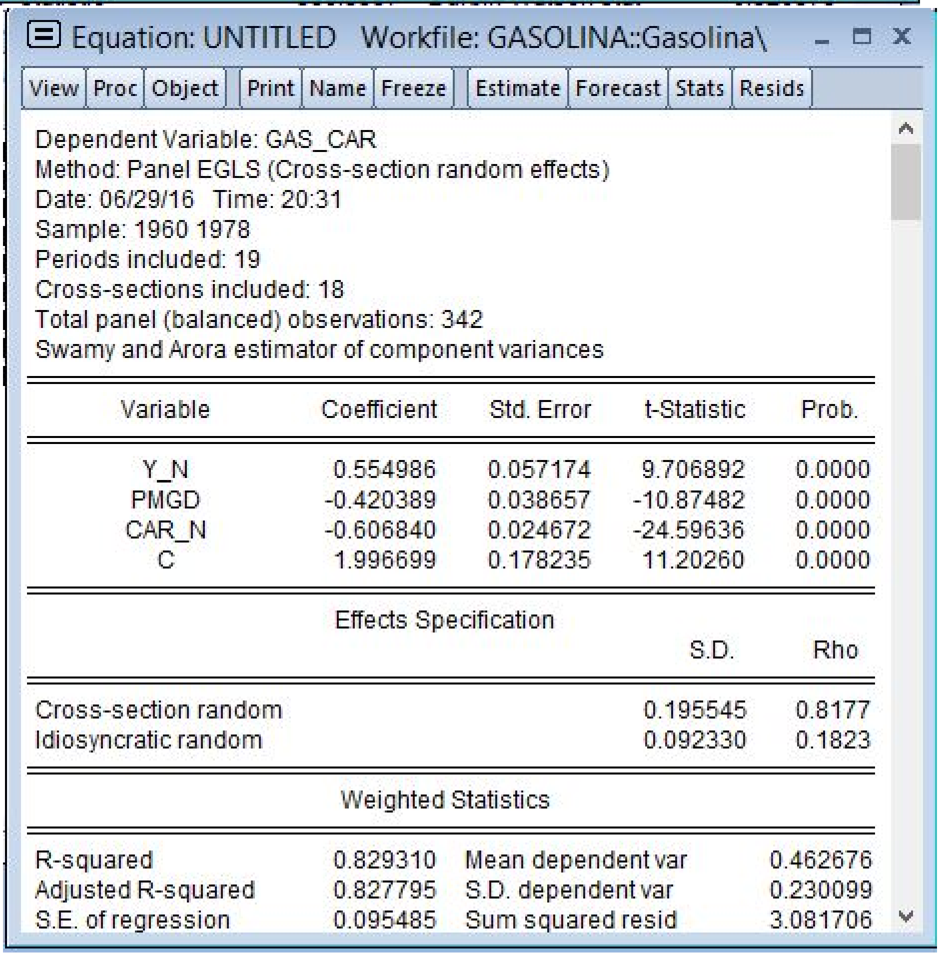
- Para estimar el modelo de efectos fijos se selecciona la opción de "fixed" y damos a "aceptar".





- El mismo procedimiento para efectos aleatorios





Dependent Variable: GAS_CAR
Method: Panel EGLS (Cross-section random effects)
Date: 06/29/16 Time: 20:31
Sample: 1960 1978
Periods included: 19
Cross-sections included: 18
Total panel (balanced) observations: 342
Swamy and Arora estimator of component variances

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y_N	0.554986	0.057174	9.706892	0.0000
PMGD	-0.420389	0.038657	-10.87482	0.0000
CAR_N	-0.606840	0.024672	-24.59636	0.0000
C	1.996699	0.178235	11.20260	0.0000

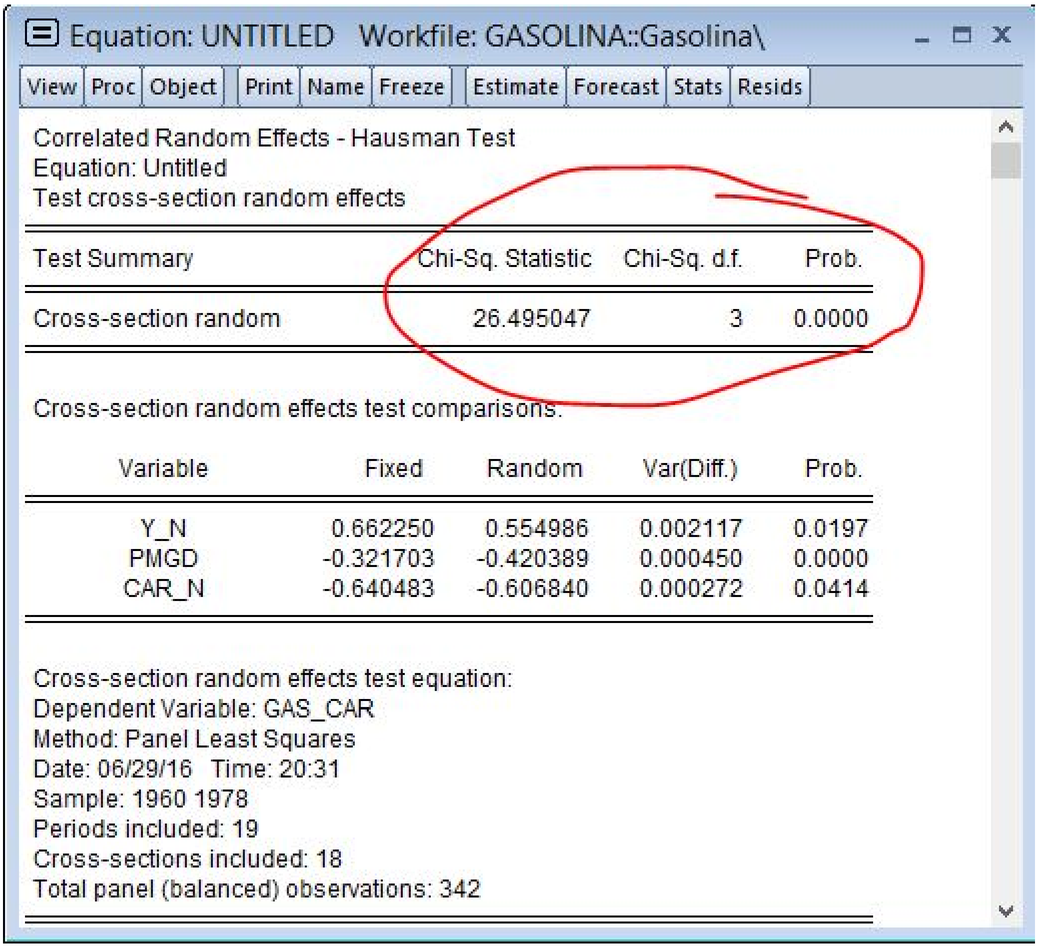
Effects Specification

	S.D.	Rho
Cross-section random	0.195545	0.8177
Idiosyncratic random	0.092330	0.1823

Weighted Statistics

R-squared	0.829310	Mean dependent var	0.462676
Adjusted R-squared	0.827795	S.D. dependent var	0.230099
S.E. of regression	0.095485	Sum squared resid	3.081706

- Por último se realiza el contraste de Hausman. Sobre la pantalla de los resultados de la estimación del modelo de efectos aleatorios, pulsamos "view"→"fixed/random effects testing" → "correlated random effects-Hausman test".



Equation: UNTITLED Workfile: GASOLINA::Gasolina\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Correlated Random Effects - Hausman Test
Equation: Untitled
Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	26.495047	3	0.0000

Cross-section random effects test comparisons:

Variable	Fixed	Random	Var(Diff.)	Prob.
Y_N	0.662250	0.554986	0.002117	0.0197
PMGD	-0.321703	-0.420389	0.000450	0.0000
CAR_N	-0.640483	-0.606840	0.000272	0.0414

Cross-section random effects test equation:
Dependent Variable: GAS_CAR
Method: Panel Least Squares
Date: 06/29/16 Time: 20:31
Sample: 1960 1978
Periods included: 19
Cross-sections included: 18
Total panel (balanced) observations: 342