



NOTA TÉCNICA
N.º 005 | 2000

La técnica de datos de panel. Una guía para su uso e interpretación

Mauricio Mayorga M.
Evelyn Muñoz S.

Fotografía de portada: “Presentes”, conjunto escultórico en bronce, año 1983, del artista costarricense Fernando Calvo Sánchez. Colección del Banco Central de Costa Rica.

La técnica de datos de panel. Una guía para su uso e interpretación.

Mauricio Mayorga M^{*}, Evelyn Muñoz S[†]

Las ideas expresadas en este documento son de los autores y no necesariamente representan las del Banco Central de Costa Rica.

Resumen

Este documento pretende ser una guía para la utilización de la técnica de datos de panel proporcionando una descripción e interpretación de los principales estadísticos que pueden ser obtenidos a partir de su aplicación con el uso del paquete estadístico TSP, Versión 4.4. Según Kikut (1999), la salida del TSP brinda una cantidad de resultados y de pruebas estadísticas superiores a los que se logra cuando se utilizan los paquetes econométricos E-VIEWS (micro TSP para windows) y SHAZAM, los cuales también permiten utilizar la técnica de datos de panel.

Palabras clave: Datos panel, Modelación económica.

Clasificación JEL: C22, C23.

* Departamento de Investigación Económica. División Económica, BCCR. mayorgamm@bccr.fi.cr

† Departamento de Investigación Económica. División Económica, BCCR. munozse@bccr.fi.cr

The Panel Data Technique, a Guide for its Use and Interpretation

Mauricio Mayorga M.[‡], Evelyn Muñoz S.[§]

The ideas expressed in this paper are those of the authors and not necessarily represent the view of the Central Bank of Costa Rica.

Key words: Panel data, Economic modeling.

JEL codes: C22, C23.

[‡] Department of Economic Research. Email address mayorgamm@bccr.fi.cr

[§] Department of Economic Research. Email address munozse@bccr.fi.cr

ÍNDICE

INTRODUCCIÓN.....	1
I. DESCRIPCIÓN DE LA TÉCNICA	1
A. CLASIFICACIÓN DE LOS MODELOS ECONÓMICOS.....	1
B. LA TÉCNICA DE DATOS DE PANEL	2
C. ESPECIFICACIÓN GENERAL DE UN MODELO DE DATOS DE PANEL.....	3
D. VENTAJAS Y DESVENTAJAS DE LA TÉCNICA DE DATOS DE PANEL.....	4
II.ALTERNATIVAS DE ESPECIFICACIÓN DE DATOS DE PANEL A PARTIR DEL MODELO GENERAL	5
A. ASPECTOS GENERALES	5
1. Modelo de efectos fijos.....	8
2. Modelo de efectos aleatorios	8
3. Elección del método: efectos fijos o efectos aleatorios?.....	9
B. IDENTIFICACIÓN DE MODELOS ESPECÍFICOS DE DATOS DE PANEL SEGÚN LA SALIDA DEL TSP	10
1. Modelo total	10
2. Modelo BYID.....	11
3. Modelo WITHIN	11
4. Modelo BETWEEN.....	12
5. Prueba de HAUSMAN	12
III.EJEMPLO DE SALIDA DE APLICACIÓN DE PROGRAMA DE DATOS EN PANEL EN TSP, VERSIÓN 4.4	12
IV.BIBLIOGRAFÍA	18

LA TÉCNICA DE DATOS DE PANEL UNA GUÍA PARA SU USO E INTERPRETACIÓN

INTRODUCCIÓN

Este documento pretende ser una guía para la utilización de la técnica de datos de panel proporcionando una descripción e interpretación de los principales estadísticos que pueden ser obtenidos a partir de su aplicación con el uso del paquete estadístico TSP, Versión 4.4. Según Kikut (1999), la salida del TSP brinda una cantidad de resultados y de pruebas estadísticas superiores a los que se logra cuando se utilizan los paquetes econométricos E-VIEWS (micro TSP para windows) y SHAZAM, los cuales también permiten utilizar la técnica de datos de panel¹.

No se pretende abarcar algunos temas teóricos en el campo estadístico relacionados con los supuestos básicos para aplicar dicha técnica o relacionados con los algoritmos internos del software utilizado en el cálculo de los principales parámetros de cada uno de los modelos estimados. Cuando corresponda, se recomendará la bibliografía a utilizar para profundizar en cada uno de dichos temas.

Para complementar lo anterior, se utilizará un ejemplo de aplicación que incorpora información acerca de la tasa de morosidad junto con algunas otras variables financieras de las entidades que conforman el Sistema Bancario Nacional.

Asimismo, este documento se centra en la técnica tradicional de datos de panel y no incorpora las consideraciones a seguir en modelos de naturaleza dinámica, en los cuales se introducen rezagos en las variables tanto independientes como dependiente².

La primera sección del documento se concentrará en efectuar una descripción de la técnica como son los tipos de problemas que se pueden solucionar con la metodología, la especificación de los modelos y sus ventajas y desventajas. Posteriormente, la segunda sección aborda aspectos metodológicos que giran en torno a los diferentes modelos que se pueden estimar y los supuestos que se encuentran detrás de cada uno de ellos. Finalmente, la tercera sección resume algunos comentarios a partir del análisis de un ejemplo utilizando una salida de TSP.

I. DESCRIPCIÓN DE LA TÉCNICA

A. Clasificación de los Modelos Econométricos

Los modelos que se utilizan en el análisis económico o de cualquier otra índole, se pueden clasificar desde dos puntos de vista³: i) según los datos utilizados y ii) según las

¹ Para un mayor detalle acerca del uso de estos otros paquetes estadísticos, puede consultarse a Kikut (1999).

² Para efectuar una consulta sobre los modelos dinámicos, se puede recurrir a Hsiao (1986).

³ Según Martín, G. Et.al. (1997).

relaciones supuestas entre las variables. Por el objetivo de este documento, se analizará el primero de los dos criterios⁴.

En el análisis de la información (económica, social, empresarial, comercial, etc.) pueden existir diferentes dimensiones sobre las cuales interesa obtener conclusiones derivadas de la estimación de modelos que traten de extraer relaciones de causalidad o de comportamiento entre diferentes tipos de variables, a partir de los datos disponibles.

Una de estas dimensiones la constituye el análisis de series de tiempo, la cual incorpora información de variables y/o unidades individuales de estudio durante un período determinado de tiempo (dimensión temporal)⁵. En este caso, cada período de tiempo constituye el elemento poblacional y/o muestral. Por su parte, existe otra dimensión que no incorpora el aspecto temporal sino que más bien representa el análisis de la información para las unidades individuales de estudio, en un momento determinado del tiempo (dimensión estructural). En este tipo de análisis, el cual se denomina de corte transversal, el elemento o unidad muestral no lo constituye el tiempo sino las unidades de análisis⁶.

Ambos tipos de análisis de la información permiten extraer conclusiones relevantes de acuerdo con los intereses del investigador. Un sencillo ejemplo puede ilustrar mejor la diferencia entre cada uno de los enfoques⁷: supóngase que se quiere modelar la rentabilidad de las firmas que pertenezcan a una determinada industria. Un análisis de regresión basado en datos de corte transversal para un año en particular podría incluir una serie de variables explicativas tales como la calidad de la administración, el monto del capital físico, el empleo de mano de obra y el nivel de apalancamiento financiero, por ejemplo. Con el tipo de información incluida en este modelo de corte transversal se podría estar tomando en cuenta cualquier tipo de economía de escala de la que las firmas podrían beneficiarse. Sin embargo, este modelo no podría identificar o tomar en cuenta como variable explicativa de la rentabilidad, cualquier incremento en el rendimiento que pueda ocurrir con el transcurso del tiempo como consecuencia de mejoras tecnológicas que hayan sido incorporadas en la industria.

B. La Técnica de Datos de Panel

A diferencia de lo mencionado en el aparte anterior, un modelo econométrico de datos de panel es uno que incluye una muestra de agentes económicos o de interés (individuos, empresas, bancos, ciudades, países, etc) para un período determinado de tiempo, esto es, combina ambos tipos de datos (dimensión temporal y estructural). A manera de ejemplo, se puede disponer de datos mensuales de los ingresos obtenidos por un grupo de cinco bancos costarricenses durante un período de 48 meses, lo cual sería una base de datos mixta de serie temporal y corte transversal constituyéndose en un panel de datos. En este ejemplo, los elementos muestrales serían el tiempo y los bancos comerciales.

⁴ Con base en el segundo criterio, los modelos se clasificarían en uniecuacionales, multiecuacionales, lineales y no lineales.

⁵ Ejemplos de este tipo de datos los constituyen las series del PIB y de tasas de interés de un país o el número de llamadas telefónicas de una familia a lo largo de un período determinado de tiempo.

⁶ Ejemplos de este tipo de análisis pueden ser la cantidad demandada de alimentos por una muestra de familias durante un período de tiempo o la cantidad producida de televisores por una serie de empresas en el mismo lapso. En ambos casos los elementos muestrales serían la familia y la empresa.

⁷ Tomado de Pindyck y Rubinfeld (1994).

El principal objetivo de aplicar y estudiar los datos en panel, es capturar la heterogeneidad no observable, ya sea entre agentes económicos o de estudio así como también en el tiempo, dado que esta heterogeneidad no se puede detectar ni con estudios de series temporales ni tampoco con los de corte transversal⁸. Esta técnica permite realizar un análisis más dinámico al incorporar la dimensión temporal de los datos, lo que enriquece el estudio, particularmente en períodos de grandes cambios. Esta modalidad de analizar la información en un modelo de panel es muy usual en estudios de naturaleza microeconómica. La aplicación de esta metodología permite analizar dos aspectos de suma importancia cuando se trabaja con este tipo de información y que forman parte de la heterogeneidad no observable: i) los efectos individuales específicos y ii) los efectos temporales.

En lo que se refiere a los efectos individuales específicos, se dice que estos son aquellos que afectan de manera desigual a cada uno de los agentes de estudio contenidos en la muestra (individuos, empresas, bancos) los cuales son invariables en el tiempo y que afectan de manera directa las decisiones que tomen dichas unidades. Usualmente se identifica este tipo de efectos con cuestiones de capacidad empresarial, eficiencia operativa, capitalización de la experiencia, acceso a la tecnología, etc.⁹.

Los efectos temporales serían aquellos que afectan por igual a todas las unidades individuales del estudio pero que no varían en el tiempo. Este tipo de efectos pueden asociarse, por ejemplo, a los choques macroeconómicos que pueden afectar por igual a todas las empresas o unidades de estudio.

C. Especificación General de un Modelo de Datos de Panel

La especificación general de un modelo de regresión con datos de panel es la siguiente¹⁰:

$$Y_{it} = \alpha_{it} + X_{it} \beta + u_{it} \quad (1)$$

$$\text{con } i = 1, \dots, N; \quad t = 1, \dots, T.$$

Donde i se refiere al individuo o a la unidad de estudio (corte transversal), t a la dimensión en el tiempo, α es un vector de interceptos de n parámetros, β es un vector de K parámetros y X_{it} es la i -ésima observación al momento t para las K variables explicativas. En este caso, la muestra total de las observaciones en el modelo vendría dado por $N \times T$.

A partir de este modelo general, y con base en ciertos supuestos y restricciones acerca del valor de algunos de los parámetros, se pueden derivar algunas otras variantes de modelos de datos de panel, las cuales se describirán con más detalle en una sección posterior.

⁸ Véase Burdisso, Tamara (1997).

⁹ Este tipo de efectos tienen bastante lógica cuando uno supone que no todas las unidades económicas toman sus decisiones de una misma forma o tomando en cuenta las mismas consideraciones, aunque estas se vean afectadas por igual por algunos otros factores exógenos (choques macroeconómicos, por ejemplo).

¹⁰ Puede consultarse a Burdisso, Tamara (1997).

Es usual interpretar los modelos de datos de panel a través de sus componentes de errores. El término de error U_{it} incluido en la ecuación (1), puede descomponerse de la siguiente manera:

$$U_{it} = \mu_i + \delta_t + \varepsilon_{it} \quad (2)$$

μ_i representa los efectos no observables que difieren entre las unidades de estudio pero no en el tiempo, que generalmente se los asocia a la capacidad empresarial, por ejemplo.

δ_t se le identifica con efectos no cuantificables que varían en el tiempo pero no entre las unidades de estudio. ε_{it} se refiere al término de error puramente aleatorio.

La mayoría de las aplicaciones con datos de panel utilizan el modelo de componente de error conocido como "one way" para el cual $\delta_t = 0$ ¹¹. Las diferentes variantes para el modelo "one way" de componentes de errores surgen de los distintos supuestos que se hacen acerca del término μ_i . Pueden presentarse tres posibilidades:

- El caso más sencillo es el que considera al $\mu_i = 0$, o sea, no existe heterogeneidad no observable entre los individuos o firmas. Dado lo anterior, los U_{it} satisfacen todos los supuestos del modelo lineal general, por lo cual el método de estimación de mínimos cuadrados clásicos produce los mejores estimadores lineales e insesgados.
- La segunda posibilidad consiste en suponer a μ_i un efecto fijo y distinto para cada firma. En este caso, la heterogeneidad no observable se incorpora a la constante del modelo.
- La tercera alternativa es tratar a μ_i como una variable aleatoria no observable que varía entre individuos pero no en el tiempo.

D. Ventajas y Desventajas de la Técnica de Datos de Panel¹²

La técnica de datos de panel presenta una serie de ventajas y desventajas en comparación con los modelos de series de tiempo y de corte transversal. Las más relevantes son las siguientes:

1. Ventajas:

¹¹ Este tipo de análisis supone que no existen efectos no cuantificables que varíen en el tiempo pero no entre las unidades individuales de estudio. Existe además el modelo "two-way" en el cual el componente de error $\delta_t \neq 0$ a través del cual se pretende capturar efectos temporales específicos (choques) que no están incluidos en la regresión. Para un mayor detalle puede consultarse a Baltagi (1999), capítulo 3.

¹² Para un mayor detalle de estos aspectos se puede consultar a Baltagi (1999) y Hsiao (1986).

- La técnica permite al investigador económico disponer de un mayor número de observaciones incrementando los grados de libertad y reduciendo la colinealidad entre las variables explicativas y, en última instancia, mejorando la eficiencia de las estimaciones econométricas¹³.
- Tal y como se mencionó anteriormente, la técnica permite capturar la heterogeneidad no observable ya sea entre unidades individuales de estudio como en el tiempo. Con base en lo anterior, la técnica permite aplicar una serie de pruebas de hipótesis para confirmar o rechazar dicha heterogeneidad y cómo capturarla.
- Los datos en panel suponen, e incorporan en el análisis, el hecho de que los individuos, firmas, bancos o países son heterogéneos. Los análisis de series de tiempo y de corte transversal no tratan de controlar esta heterogeneidad corriendo el riesgo de obtener resultados sesgados.
- Permite estudiar de una mejor manera la dinámica de los procesos de ajuste. Esto es fundamentalmente cierto en estudios sobre el grado de duración y permanencia de ciertos niveles de condición económica (desempleo, pobreza, riqueza).
- Permite elaborar y probar modelos relativamente complejos de comportamiento en comparación con los análisis de series de tiempo y de corte transversal. Un ejemplo claro de este tipo de modelos, son los que se refieren a los que tratan de medir niveles de eficiencia técnica por parte de unidades económicas individuales (empresas, bancos, etc).

2. Desventajas:

- En términos generales, las desventajas asociadas a la técnica de datos de panel se relacionan con los procesos para la obtención y el procesamiento de la información estadística sobre las unidades individuales de estudio, cuando esta se obtiene por medio de encuestas, entrevistas o utilizando algún otro medio de levantamiento de los datos. Ejemplos de este tipo de limitaciones son: cobertura de la población de interés, porcentajes de respuesta, preguntas confusas, distorsión deliberada de las respuestas, etc.

II. ALTERNATIVAS DE ESPECIFICACIÓN DE DATOS DE PANEL A PARTIR DEL MODELO GENERAL

A. Aspectos Generales

En la tabla N°1 siguiente se aprecia la gama de posibilidades que ofrece la técnica de datos de panel para tratar un problema específico, partiendo de un modelo lineal general.

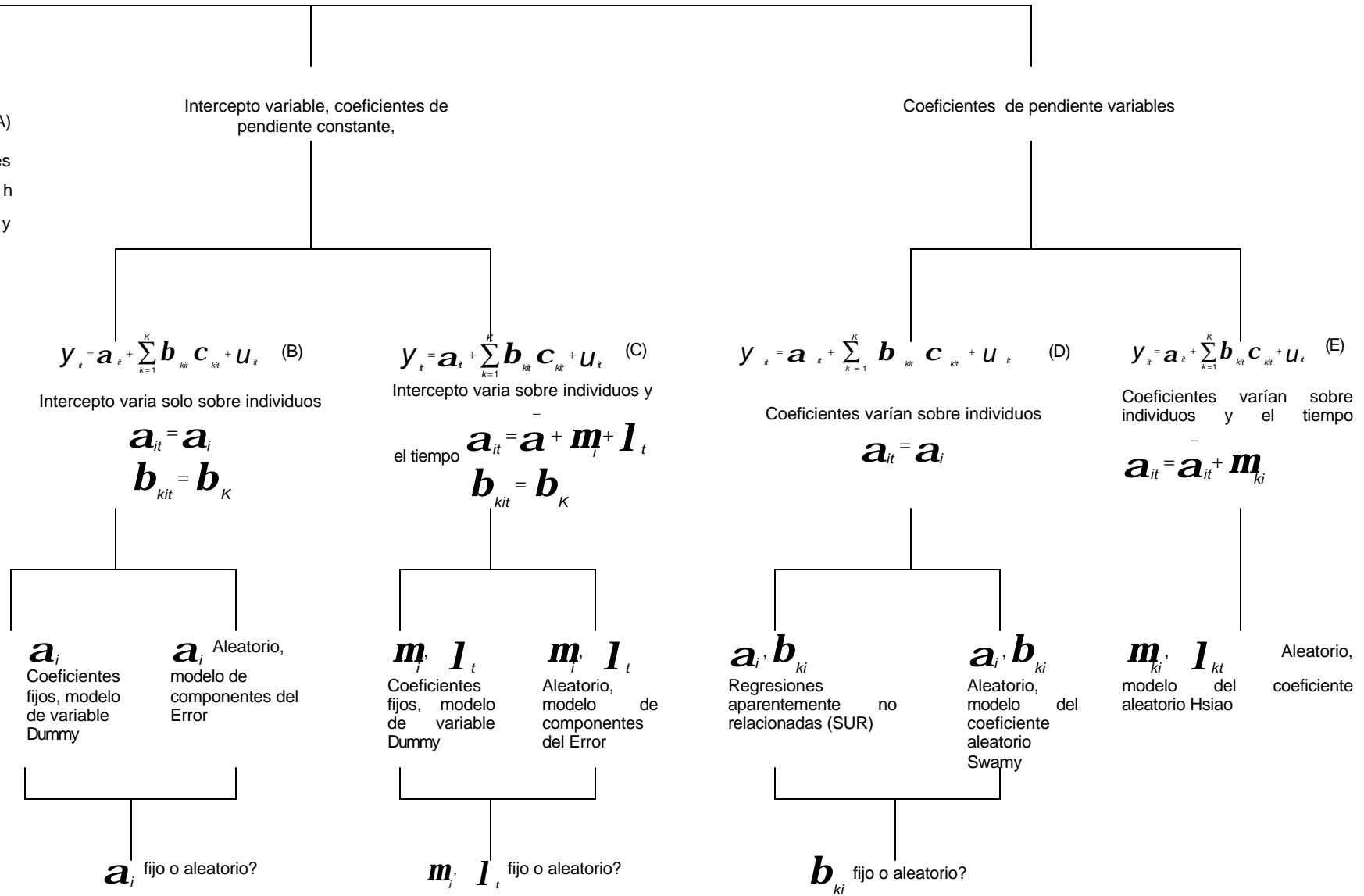
¹³ Hay que recordar que el número de observaciones totales vendría dado por $N \times t$, donde N es el número de unidades de estudio individual y t el período de tiempo.

TABLA N°1
 MODELOS ALTERNATIVOS PARA COMBINAR DATOS DE SERIES DE TIEMPO Y DE CORTE TRANSVERSAL

El Modelo Lineal

$$y_{it} = a_{it} + \sum_{k=1}^K b_{kit} c_{kit} + U_{it}$$

$y_{it} = a_{it} + \sum_{k=1}^K b_{kit} c_{kit} + U_{it}$ (A)
 Todos los coeficientes constantes, $b_{kit} = b_k$ y e_{it} heteroscedástico y autocorrelacionado



Tomado y adaptado de: Judge, et. al. (1980), Capítulo 13..

La primera especificación (A) se refiere al caso en que no existe heterogeneidad no observable en el sistema de datos de panel y por tanto se emplea el método de Mínimos Cuadrados Ordinarios con la ventaja de ganar grados de libertad.

Ahora bien, en los casos en que se rechaza la hipótesis de homogeneidad en un sistema de datos de panel, es decir que existe heterogeneidad no observable ya sea a través del tiempo, entre unidades de estudio (individuos) o en ambos sentidos, debe buscarse una especificación que la capture en forma apropiada, con el fin de evitar el problema de sesgo sobre los estimadores de los parámetros de las variables explicativas¹⁴, que se cometería si se emplea la especificación A.

Existen dos procedimientos adicionales para estimar el modelo en un sistema de datos de panel: uno de ellos implica el reconocimiento de que las variables omitidas pueden generar cambios en los interceptos ya sea a través del tiempo o entre unidades de corte transversal, en este caso el modelo de efectos fijos trata de aproximar estos cambios con variables dummy; el otro modelo es el de efectos aleatorios, que trata de capturar estas diferencias a través del componente aleatorio del modelo.

Como ya se mencionó, la técnica de datos de panel permite contemplar la existencia de efectos individuales específicos a cada firma, invariables en el tiempo que afectan la manera en que cada unidad de corte transversal toma sus decisiones.

Una forma simple, y de hecho la más utilizada, de considerar esta heterogeneidad es empleando los modelos de intercepto variable, identificados en las especificaciones B y C en la tabla N°1. Así, el modelo lineal es el mismo para todas las unidades o individuos bajo estudio, pero la ordenada al origen es específica a cada una de ellas. A partir del modelo general esta situación se representa mediante la siguiente ecuación:

$$Y_{it} = a_i + \sum b_k X_{ki} + m_t \quad (3)$$

El supuesto básico de estos modelos es que dadas las variables explicativas observadas, los efectos de todas las variables omitidas pueden representarse de tres formas posibles:

- a. Una variable por cada individuo, no variable en el tiempo: este es el caso de variables que son las mismas para cada unidad de corte transversal a través del tiempo. Como ejemplos de ellas se tienen: características de la administración de las empresas, sexo, capacidad, religión y otras características sociales.
- b. Una variable por periodo pero no variables entre individuos: son la misma variable para todos los individuos en un momento del tiempo pero varían a lo largo del periodo de estudio. Como ejemplo se tienen los precios, las tasas de interés, el nivel de actividad económica, etc.
- c. Una variable que cambia en el tiempo y por individuo: se trata de variables que cambian entre individuos en un momento del tiempo, y que además cambian a lo largo del tiempo. Como ejemplo de estas variables se pueden mencionar los

¹⁴ Adicionalmente, la eficiencia de los parámetros mejora con el empleo de datos de panel al incrementar el número de grados de libertad.

ingresos totales, el nivel de beneficios, el nivel de capital, razones financieras, entre otras.

Estos modelos de intercepto variable asumen que los efectos de las variables omitidas, ya sean específicas a cada individuo y/o que cambian en el tiempo, no son importantes en forma individual, pero que si son importantes si se consideran en conjunto.

Por otro lado, dado que el efecto de las variables omitidas puede mantenerse constante en el tiempo para cada individuo, o ser el mismo para todos los individuos en un momento en el tiempo, o una combinación de ambos, se pueden capturar en el término constante de un modelo de regresión como un promedio que toma en cuenta explícitamente la heterogeneidad entre individuos y/o en el tiempo contenida en los datos¹⁵.

Seguidamente, se analizarán los principales modelos que arroja el paquete TSP, a partir de la especificación general y de acuerdo con la forma de incorporar la heterogeneidad no observada.

1. Modelo de efectos fijos

Como se indicó brevemente, una posibilidad es explicar los datos con el modelo de efectos fijos considera que existe un término constante diferente para cada individuo, y supone que los efectos individuales son independientes entre sí.

Con este modelo se considera que las variables explicativas afectan por igual a las unidades de corte transversal y que éstas se diferencian por características propias de cada una de ellas, medidas por medio del intercepto. Es por ello que los N interceptos se asocian con variables dummy con coeficientes específicos para cada unidad, los cuales se deben estimar. Para la i-ésima unidad de corte transversal, la relación es la siguiente:

$$Y_i = i\mathbf{a}_i + \mathbf{b}X_i + \mathbf{m}_i \quad (4)$$

Donde el subíndice i representa un vector columna de unos. Debe hacerse notar que en este modelo se presenta una pérdida importante de grados de libertad.

2. Modelo de efectos aleatorios

A diferencia del modelo de efectos fijos, el modelo de efectos aleatorios considera que los efectos individuales no son independientes entre sí, sino que están distribuidos aleatoriamente alrededor de un valor dado. Una práctica común en el análisis de regresión es asumir que el gran número de factores que afecta el valor de las variable dependiente pero que no han sido incluidas explícitamente como variables independientes del modelo, pueden resumirse apropiadamente en la perturbación aleatoria.

Así, con este modelo se considera que tanto el impacto de las variables explicativas como las características propias de cada banco son diferentes. El modelo se expresa algebraicamente de la siguiente forma:

¹⁵ Este es el procedimiento que se emplea por ejemplo al estimar una función Cobb Douglas.

$$Y_{it} = (\mathbf{a} + \mathbf{m}_i) + \mathbf{b}' X_{it} + \mathbf{e}_{it} \quad (5)$$

Donde: “ μ_i ” viene a representar la perturbación aleatoria que permitiría distinguir el efecto de cada individuo en el panel. Para efectos de su estimación se agrupan los componentes estocásticos, y se obtiene la siguiente relación:

$$Y_{it} = \mathbf{a} + \mathbf{b}' X_{it} + U_{it} \quad (6)$$

Donde $U_{it} = \mathbf{d}_i + \mathbf{m}_i + \mathbf{e}_{it}$ se convierte en el nuevo término de la perturbación, U no es homocedástico, donde $\mathbf{d}_i, \mathbf{m}_i, \mathbf{e}_{it}$ corresponden al error asociado con las series de tiempo (δ_i); a la perturbación de corte transversal (\mathbf{m}_i) y el efecto combinado de ambas (ε_{it}).

El método de Mínimos Cuadrados Ordinarios (MCO) no es aplicable dado que no se cumplen los supuestos que permiten que el estimador sea consistente. Por lo que es preferible en este caso utilizar el método de Mínimos cuadrados Generalizados (MCG) cuyas estimaciones son superiores al de MCO en caso de no cumplirse los supuestos tradicionales y son similares en caso contrario.

3. Elección del método: efectos fijos o efectos aleatorios?

La decisión acerca de la estructura apropiada para el análisis, es decir, efectos fijos vrs efectos aleatorios depende en parte de los siguientes aspectos:

a. Los objetivos del estudio

Si se desea hacer inferencias con respecto a la población, es decir que se trabaja con una muestra aleatoria, lo mejor es utilizar una especificación del tipo aleatoria. En caso de que el interés sea limitado a una muestra que se ha seleccionado a conveniencia o bien que se está trabajando con la población, la estimación de efectos fijos será la correcta.

Adicionalmente, si el interés del estudio particular está puesto en los coeficientes de las pendientes de los parámetros, y no tanto en las diferencias individuales, se debería elegir un método que relegue estas diferencias y tratar la heterogeneidad no observable como aleatoria.

El modelo de efectos fijos se ve como un caso en que el investigador hace inferencia condicionada a los efectos que ve en la muestra. El de efectos aleatorios se ve como uno en el cual el investigador hace inferencia condicional o marginal respecto a una población. Se deja al investigador que decida si hace inferencia con respecto a las características de una población o solo respecto a los efectos que están en la muestra.

b. El contexto de los datos, es decir, cómo fueron obtenidos y el entorno de donde provienen

Con el método de efectos fijos la heterogeneidad no observable se incorpora en la ordenada al origen del modelo y con la de efectos aleatorios, como ya se mencionó, se incorporan en el término de error, por lo cual lo que se modifica es la varianza del modelo.

Emplear un modelo de efectos fijos o aleatorios genera diferencias en las estimaciones de los parámetros en los casos en que se cuenta con t pequeño y N grande. En estos casos debe hacerse el uso más eficiente de la información para estimar esa parte de la relación de comportamiento contenida en las variables que difieren sustancialmente de un individuo a otro.

c. Número de datos disponibles

El método de efectos fijos presenta el problema de que el uso de variables “Dummies” no identifica directamente qué causa que la regresión lineal cambie en el tiempo y en los individuos. Además, esto implica la pérdida de grados de libertad.

Asimismo, deberán tomarse consideraciones con respecto a la estructura de los datos con que se cuente, dado que si la N es grande pero si se tiene un T pequeño, podría ser que el número de parámetros de efectos fijos sea muy grande en relación con el número de datos disponibles, con parámetros poco confiables y una estimación ineficiente.

Algunas investigaciones han demostrado que el emplear modelos de efectos fijos produce resultados significativamente diferentes al de efectos aleatorios cuando se estima un ecuación usando una muestra de muchas unidades de corte transversal con pocos periodos de tiempo (629 individuos para 6 periodos, por ejemplo).

B. Identificación de Modelos Específicos de Datos de Panel Según la Salida del TSP

Seguidamente se detallan algunos aspectos metodológicos de la técnica de datos de panel, a los cuales se hará referencia en una sección posterior en la cual se analice un caso práctico aplicando el paquete computacional TSP.

Para efectos prácticos, el procedimiento de identificación de la especificación más apropiada para el panel de datos particular, parte de la comparación entre dos modelos que pueden considerarse como extremos: aquel que considera que no hay heterogeneidad no observable, (ecuación A Tabla N°1) y el que considera que hay diferencias entre unidades de corte transversal tanto en el intercepto como en los coeficientes de las variables explicativas (ecuación D Tabla N°1)

Una vez concluido que no es posible explicar igualmente bien los datos con los modelos extremos, se procede a evaluar otras opciones.

1. Modelo total

Esta es la técnica más sencilla que emplea modelos de datos de panel, combina todas las series de tiempo de las unidades de corte transversal y luego estima el modelo “subyacente” usando Mínimos Cuadrados Ordinarios (MCO). Supone que el conjunto de coeficientes tanto de la pendiente (intercepto α) como del impacto de cada una de las variables explicativas ($\beta=\beta_i$), es el mismo con respecto a la variable dependiente ($y_{it} = \mathbf{a} + X_{it} \mathbf{b}$)

Para determinar si este modelo es el que mejor se ajusta a los datos se utiliza una prueba F, donde la hipótesis nula consiste en que el comportamiento de la variable dependiente se explica “igualmente bien” con el modelo TOTAL que con el modelo donde tanto los interceptos como los coeficientes de las variables explicativas varían a lo largo de las unidades cruzadas (MODELO BYID)

Un problema con el modelo de intercepto y coeficientes constantes es que puede carecer de sentido según el problema que se esté analizando.

2. Modelo BYID

Este modelo asume que tanto los coeficientes de las variables explicativas como de los interceptos varían entre los distintos individuos, tal como lo indica la especificación D de la tabla No.1. Para determinar la conveniencia de este modelo se emplea una prueba F que permite concluir que entre las unidades de corte transversal, tanto los interceptos como el impacto de las distintas variables explicativas es diferente entre los distintos individuos. La prueba se identifica como $H_0: A, B = A_i, B_i$.

3. Modelo WITHIN

Este es un modelo de efectos fijos que asume que cada variable explicativa tiene un solo coeficiente, es decir tiene el mismo impacto sobre la variable dependiente, pero en donde cada individuo tiene distinta constante ($Y_{it} = a_i + X_{it} b$).

Se utiliza una prueba F, con la hipótesis alternativa de que hay suficiente evidencia empírica para decir que aún agregando interceptos diferentes (A_i)¹⁶, se explica mejor el comportamiento de la variable dependiente cuando se estiman diferentes coeficientes para las variables explicativas. En forma esquemática sería: $H_0: A, B = A_i, B_i$.

No rechazar esta primera prueba implica que los datos siguen un modelo de efectos fijos para las variables explicativas, por lo que en segundo lugar debe someterse a prueba que además de que se tienen los mismos coeficientes para las variables explicativas, los parámetros del intercepto varían entre unidades de corte transversal. Esto se realiza con la segunda prueba de hipótesis del modelo WITHIN, que establece como hipótesis nula: $H_0: A, B = A_i, B_i$.

En caso de que se concluya que los parámetros del intercepto son variables, se hace uso de la estimación de cada uno de ellos que muestra el paquete TSP. En caso contrario, se haría uso de la estimación del modelo de efectos aleatorios que calcula este mismo comando.

Este comando está asociado con el análisis de los efectos de corto plazo, ya que operan sobre el componente de los datos asociado con el tiempo, ignorando la variación que existe entre las unidades de corte transversal.

¹⁶ El paquete TSP denota los parámetros con A y B en lugar de a y b.

4. Modelo BETWEEN

Se trata de una estimación de corte transversal que se realiza sobre los promedios de las variables de las unidades de corte transversal a lo largo del periodo observado (OLS on means). Produce estimadores “entre grupos” (promedio para el sistema) y proporciona datos que sugieren la idea de largo plazo ya que se trata de un promedio de los diferentes momentos, ignorando la variación que existe dentro de cada unidad de corte transversal a lo largo del tiempo.

Es un procedimiento exactamente igual que TOTAL con la salvedad de que se realiza una regresión con mínimos cuadrados ordinarios pero utilizando los promedios para cada individuo.

5. Prueba de HAUSMAN

Esta prueba permite determinar qué modelo es el más adecuado para el panel de datos que se está analizando, si el de efectos fijos o de efectos aleatorios.

Utiliza para ello una prueba Chi-cuadrado con la hipótesis nula de que el modelo de efectos aleatorios es el que mejor explica la relación de la variable dependiente con las explicativas, y por tanto se tiene la hipótesis alternativa de que el mejor método que se ajusta es el de efectos fijos.

III.EJEMPLO DE SALIDA DE APLICACIÓN DE PROGRAMA DE DATOS EN PANEL EN TSP, VERSIÓN 4.4¹⁷

En este ejemplo, se utiliza parte de la información recopilada en el Departamento de Investigaciones Económicas del Banco Central de Costa Rica relativa a los bancos públicos y privados que operan en el país.

Originalmente, la información corresponde a 22 periodos (enero de 1997 a octubre de 1998) para 25 bancos. Es importante mencionar que no pueden existir valores perdidos en las series de tiempo para cada una de las unidades individuales, puesto que se obtendrían estimaciones sesgadas de los parámetros.

El objetivo del ejercicio es analizar la relación entre la mora financiera (MORA), la tasa de interés activa promedio (TASA; variable que es común para todos los bancos) y los gastos administrativos en que incurren esas instituciones (GADM, variable que es específica y diferente para cada uno de los bancos).

¹⁷ Para detalles adicionales acerca de los comandos para la elaboración de programas en TSP y de las bases de datos, se puede consultar el manual de referencia de TSP.

EL PROGRAMA PARA TSP Y SU SALIDA

PROGRAMA

```
1 ? EJEMPLO CON DATOS DE LOS BANCOS SUMINISTRADOS POR EL BCCR
1 options crt limwarn=0;
2 freq(panel,t=22,id=@id);
3 set nobs=25*22;
4 smpl 1 nobs;
5 read(file='c:\ok\bccr\poolbccr.xls',format=excel) @id periodo mora
  tasa gadm;
6 panel(id=@id,mean,byid,print) mora c tasa gadm;
  EXECUTION
```

En esta primera parte se despliega el programa utilizado, así como la descripción que el investigador desea que aparezca como encabezado. El programa debe estar en formato texto. La línea 5 permite corroborar la ruta donde está almacenada la base de datos y el orden de lectura de las variables. TSP lee los datos desde un formato de hoja excel (debe verificarse que no sea formato de libro). En la línea 6 se verifica qué información se le solicita al TSP que imprima.

Current sample: 1 to 550

PANEL DATA ESTIMATION

=====

Balanced data: NI= 25, T= 22, NOB= 550

MEANS

	MORA	TASA	GADM
1	3.07342	28.38182	0.92398
2	6.70070	28.38182	0.84730
3	7.13616	28.38182	0.63016
4	2.41288	28.38182	0.63879
5	3.14077	28.38182	0.71722
6	4.77927	28.38182	2.53254
7	1.71817	28.38182	0.71348
8	2.54321	28.38182	0.86220
9	1.00098	28.38182	0.55349
10	1.02436	28.38182	0.55126
11	0.37320	28.38182	0.58144
12	5.79038	28.38182	0.62121
13	0.90772	28.38182	0.36786
14	1.32986	28.38182	0.49799
15	4.92183	28.38182	0.51261
16	1.46918	28.38182	0.79125
17	9.86759	28.38182	0.95033
18	0.62224	28.38182	0.96954
19	1.69292	28.38182	0.63299
20	2.85546	28.38182	0.56887
21	2.80886	28.38182	0.71300
22	11.30215	28.38182	0.45160

23	3.04243	28.38182	0.73554
24	6.94845	28.38182	0.56907
25	9.69629	28.38182	1.22171

Para verificar que los datos han sido leídos correctamente, se despliega la lista anterior con los valores promedio de cada una de las variables en el período bajo estudio. Asimismo se verifica el número total de observaciones que serán utilizadas en la estimación de Panel, en este caso 550.

TOTAL (plain OLS) Estimates:

Dependent variable: MORA

Mean of dep. var. = 3.88634	R-squared = .060968
Std. dev. of dep. var. = 3.55352	Adjusted R-squared = .057535
Sum of squared residuals = 6509.82	LM het. test = 12.6892 [.000]
Variance of residuals = 11.9010	Durbin-Watson = .158498 [.000,.000]
Std. error of regression = 3.44978	

	Estimated	Standard		
Variable	Coefficient	Error	t-statistic	P-value
TASA	.583204	.124192	4.69598	[.000]
GADM	.730922	.229586	3.18365	[.002]
C	-13.2261	3.51474	-3.76303	[.000]

F test of A,B=Ai,Bi: F(72,475) = 46.754, P-value = [.0000]

Critical F value for diffuse prior (Leamer, p.114) = 8.4723

Esta primera prueba es el modelo TOTAL que consiste en una regresión lineal simple para todo el período estimada con el método de mínimos cuadrados ordinarios.

En este punto se hace una prueba de hipótesis, en donde la hipótesis nula es que la morosidad en la banca comercial se explica igualmente bien con un modelo que incorpora igual intercepto e igual sensibilidad para todos los bancos ante cambios en TASA y GADM, en comparación con un modelo en el que esas sensibilidades son diferentes para cada banco, así como el intercepto. En este caso particular se rechaza esta hipótesis y se concluye por lo tanto que existe heterogeneidad en el comportamiento de los bancos.

BYID COEFS

	TASA	GADM	C
1	0.11557	4.47611	-4.34254
2	0.71327	1.06211	-14.44320
3	-0.35235	-2.39846	18.64804
4	0.51092	7.85974	-17.10858

5	-0.49190	-3.43598	19.56601
6	1.17192	0.23780	-29.08407
7	0.23459	0.77310	-5.49146
8	2.11228	3.18979	-60.15750
9	-0.36553	-2.80878	12.93001
10	0.92099	1.47150	-25.92616
11	0.082653	0.43618	-2.22624
12	0.31745	12.13793	-10.75955
13	0.0011299	-2.65242	1.85138
14	0.36149	4.33848	-11.09038
15	1.59274	-7.88303	-36.24199
16	1.12120	1.42465	-31.47975
17	-0.57149	1.16345	24.98195
18	0.32647	-0.012846	-8.63106
19	-0.67903	12.61616	12.97904
20	0.79829	12.56446	-26.94898
21	0.094842	2.27750	-1.50678
22	0.55855	-0.84806	-4.16766
23	0.68651	0.60624	-16.88801
24	1.47809	-0.66911	-34.62171
25	1.96398	-1.45668	-44.26542

Este modelo asume que todas las sensibilidades y los interceptos son diferentes entre los bancos, la estimación de los parámetros se muestra en la lista anterior.

BETWEEN (OLS on means) Estimates:

Dependent variable: MORA

Mean of dep. var. = 3.88634 Std.
 Std. dev. of dep. var. = 3.13768
 Sum of squared residuals = 229.413
 Variance of residuals = 10.4279

error of regression = 3.22922
 R-squared = .029069
 Adjusted R-squared = -.059198
 LM het. test = .111355 [.739]

	Estimated	Standard		
Variable	Coefficient	Error	t-statistic	P-value
TASA	.102090	.048588	2.10111	[.047]
GADM	1.29057	1.59020	.811575	[.425]
C	0.	0.	0.	[1.00]

Este es el modelo que es similar al TOTAL con la única diferencia que el ajuste se realiza con los valores promedios en el tiempo de las variables explicativas de cada uno de los bancos. Esta estimación da idea de la relación de largo plazo.

WITHIN (fixed effects) Estimates:

Dependent variable: MORA

Mean of dep. var. = 3.88634
 Std. dev. of dep. var. = 3.55352
 Sum of squared residuals = 1415.23
 Variance of residuals = 2.70599
 Std. error of regression = 1.64499

R-squared = .795855
 Adjusted R-squared = .785706
 LM het. test = 7.39382 [.007]
 Durbin-Watson = .684319 [.000,.000]

	Estimated	Standard		
Variable	Coefficient	Error	t-statistic	P-value
TASA	.603440	.059416	10.1561	[.000]
GADM	.354998	.141547	2.50799	[.012]

Este modelo asume que cada variable explicativa tiene un solo coeficiente para todos los bancos y un intercepto diferente para cada uno.

F test of $A_i, B=A_i, B_i$: $F(48,475) = 7.5021$, P-value = [.0000]
 Critical F value for diffuse prior (Leamer, p.114) = 7.2679

F test of $A, B=A_i, B$: $F(24,523) = 78.446$, P-value = [.0000]
 Critical F value for diffuse prior (Leamer, p.114) = 6.9075

La primera prueba F permite concluir que, aún agregando interceptos distintos, se explica mejor el comportamiento de la variable dependiente cuando se estiman diferentes coeficientes para las variables explicativas

El valor de la segunda F permite concluir que entre los bancos existen diferentes puntos de "arranque" o interceptos.

En resumen, el análisis conjunto de las tres pruebas F mostradas por TSP, permite rechazar la hipótesis de que TASA y GADM tienen un efecto igual sobre la morosidad, para todos los bancos. Además, cada uno de los bancos poseen diferente intercepto.

Fixed effects

	1	2	3	4	5
1	-14.38131	-10.72680	-10.21426	-14.94060	-14.24057
	6	7	8	9	10
1	-13.24650	-15.66184	-14.88959	-16.32223	-16.29805
	11	12	13	14	15
1	-16.95993	-11.55687	-16.34958	-15.97365	-12.38686
	16	17	18	19	20
1	-15.93843	-7.59649	-16.84866	-15.65851	-14.47321
	21	22	23	24	25
1	-14.57097	-5.98488	-14.34540	-10.38029	-7.86413

Variance Components (random effects) Estimates:

VWITH (variance of U_{it}) = 2.7060
VBET (variance of A_i) = 9.1950
(computed from small sample formula)
THETA (0=WITHIN, 1=TOTAL) = 0.13200E-01

Dependent variable: **MORA**

Mean of dep. var. = 3.88634
Std. dev. of dep. var. = 3.55352
Sum of squared residuals = 6540.35
Variance of residuals = 11.9568
Std. error of regression = 3.45785
R-squared = .056960
Adjusted R-squared = .053512
LM het. test = 5.02084 [.025]
Durbin-Watson = .148208 [.000,.000]

	Estimated	Standard		
Variable	Coefficient	Error	t-statistic	P-value
TASA	.602997	.059412	10.1494	[.000]
GADM	.363220	.140923	2.57743	[.010]
C	-13.5061	1.78360	-7.57239	[.000]

Hausman test of $H_0:RE$ vs. FE : $CHISQ(1) = 0.38396$, P-value = [.5355]

En caso de que la mejor especificación sea la de efectos aleatorios, se toman los coeficientes que se despliegan arriba junto con los coeficientes de los interceptos listados anteriormente para cada uno de los bancos. Asimismo se incluye la prueba de Hausman que contrasta la hipótesis nula de que el modelo de efectos aleatorios es más adecuado que el de efectos fijos; para el caso que nos ocupa no se rechaza dicha hipótesis y se concluye que el modelo de efectos aleatorios es el que mejor ajusta la tasa de morosidad.

Como resultado que vale la pena destacar para el ejemplo que nos ocupa, el modelo estima una relación positiva entre la morosidad y la tasa de interés activa y dicha relación es de 0.602. Lo anterior implica que ante un aumento de un punto porcentual en la tasa de interés activa produce un incremento en la razón de morosidad de 0.602%. De la misma manera es interesante el resultado de una relación positiva entre los gastos administrativos y la morosidad (0.363).

END OF OUTPUT.

IV.BIBLIOGRAFÍA

- Baltagi, Badi. 1995. *Econometric Analysis of Panel Data*. John Wiley & Sons
- Burdisso, Tamara. 1997. *Estimación de una Función de Costos para los Bancos Privados Argentinos Utilizando Datos de Panel*. Banco Central de la República Argentina, Documentos de Trabajo N° 3.
- Hsiao, Cheng. 1986. *Analysis of Panel Data*. Econometric Society Monographs. Cambridge University Press.
- Judge, et.al. 1980. *The Theory and Practice of Econometrics*. John Wiley and Sons, Second Edition.
- Kikut, Otto. 1999. *Uso de Datos de Tiempo y Sección Cruzada*. Monografía. Febrero.
- Martín, Guillermina. 1997. Et.al. *Introducción a la Econometría*. Prentice Hall.
- Pindyck, Robert y Rubinfeld, Daniel. 1993. *Econometric Models & Economic Forecasts*. McGraw-Hill International Editions, Fourth Edition.

mayorgamm@bccr.fi.cr
munozse@bccr.fi.cr