TELOS. Revista de Estudios Interdisciplinarios en Ciencias Sociales
UNIVERSIDAD Rafael Belloso Chacín
ISSN 1317-0570 ~ Depósito legal pp: 199702ZU31
Vol. 16 (1): 157 - 164, 2014

Modelos de regresión de datos panel y su aplicación en la evaluación de impactos de programas sociales*

Regression Models for Panel Data and their Application to the Evaluation of Social Program Impacts

*Josefa Ramoni Perazzi*** y *Giampaolo Orlandoni Merli****

Resumen

La continuidad de los programas o proyectos sociales depende en gran medida de si éstos logran alcanzar los objetivos que se plantean y de la magnitud del efecto de los mismos. De allí la importancia de la evaluación de impactos, la cual puede basarse en diferentes tipos de datos. Particularmente, los datos panel proporcionan información de tiempo y espacio que si bien puede resultar relativamente más costosa y no siempre viable, permite analizar a fondo el impacto del proyecto sobre la comunidad controlando por factores no observables inherentes tanto al individuo como a la región en que habita. Este trabajo introduce el uso de modelos de datos panel de efectos fijos y efectos aleatorios, aplicados a la evaluación de impactos de programas sociales.

Palabras clave: Evaluación de impactos, programas sociales, contra factual, datos panel, efectos fijos, efectos aleatorios.

* Esta investigación contó con el apoyo del CDCHTA (E-287-08-09-A), ULA. Mérida. Venezuela.

** Economista, Master en Estadística y en Economía; Ph.D. en Economía. Profesora Titular de la Facultad de Ciencias Económicas y Sociales (Universidad de Los Andes). Mérida, Venezuela. Correo electrónico: jramoni@ula.ve

*** Economista, Master Science en Economía; Doctor (HC) en Estadística. Profesor Titular de la Facultad de Ciencias Económicas y Sociales (Universidad de Los Andes). Mérida, Venezuela. Correo electrónico: orlandon@ula.ve

Abstract

The continuity of social projects and programs depends on whether their objectives are fulfilled and the magnitude of their impact is significant. That explains the importance of impact evaluation, which may rely on different types of data. Particularly, panel data provides information about time and space that, even though it may be relatively more expensive and not always feasible, allows for a deeper analysis of the impact of the project on the community, controlling for unobservable factors inherent in the individual as well as the region. This paper introduces the use of fixed effects and random effects panel data models as tools applied to assess social programs.

Keywords: impact evaluation, social programs, counterfactual, panel data, fixed effects, random effects.

Introducción

Los programas sociales tienen por finalidad alcanzar determinados objetivos y beneficiar grupos claramente identificados. Antes de ser implementados, esos programas pueden parecer potencialmente beneficiosos, pero muchas veces no producen los impactos y beneficios esperados.

La evaluación de impactos (EI) tiene como propósito ayudar a los responsables de las políticas públicas decidir si los programas o proyectos están generando los efectos deseados y planificados, promover la transparencia en la distribución de recursos entre diferentes programas, controlando los proyectos que funcionan, y determinar si los cambios en el bienestar de los grupos objetivo pueden atribuirse a dichos proyectos y políticas públicas implementadas (Klugman 2002, p. 117).

La evaluación de impactos forma parte de un programa amplio de formulación de políticas socioeconómicas fundamentadas en evidencias. Consiste en un análisis sistemático de la relevancia, operación y resultados de programas, comparado con un conjunto de estándares prestablecidos, tratando de determinar la potencial diferencia que un programa específico puede causar. Se trata de un análisis "con versus sin": qué cambios ha causado el programa (hecho factual) comparado con qué hubiera pasado sin implementar dicho programa (hecho contrafactual). La EI mide la efectividad de un programa comparando los resultados para un grupo beneficiario versus un grupo control, ambos antes y después de su implementación (Klugman 2002, p. 118; White 2011, p. 4).

La base del análisis está en la definición del grupo de comparación, analizando qué hubiera pasado al grupo beneficiado de no haberse hecho la intervención y haberlo expuesto al programa. La identificación de dicho grupo de comparación o grupo control es la esencia de las evaluaciones de impacto. El análisis estadístico permite construir dicho contrafactual. Para ello se identifica el grupo control, similar en todos los aspectos al grupo intervención. Así las diferencias en los indicadores de interés se comparan en el grupo intervención y grupo control,

luego de la intervención (diferencia ex post). Una evaluación de impacto tiene validez interna si utiliza un grupo de comparación que produce una estimación válida del contrafactual (White 2011, p. 5).

Para poder llevar a cabo esos análisis se requiere tener datos estadísticos referentes al problema. Los tipos de datos para hacer la EI son los datos atemporales, series temporales y datos panel. En los datos atemporales (corte transversal) se toma información de un conjunto de unidades muestrales (UM) en un punto dado del tiempo. Las evaluaciones usando este tipo de datos son poco costosas aunque resultan también poco confiables, pues es difícil distinguir si los cambios observados se deben a la intervención de los programas o a otros factores. En los datos de series temporales se toma información sobre determinadas variables respuesta a intervalos periódicos de tiempo, antes y después del programa. Usualmente se usan en la evaluación de programas de cobertura nacional.

En los datos panel (o micro paneles), la misma UM (familia, empresa, país, individuo) es estudiada a lo largo del tiempo, generalmente antes de la intervención del programa (baseline) y luego de la intervención (al menos una observación en cada caso), por lo que se tienen dos dimensiones: espacial o estructural y temporal. La primera, generalmente predominante, permite modelar diferencias individuales observables, y controlar por las no observables, lo cual viene a solventar una de las debilidades del modelo de regresión clásico; de allí también la potencial presencia de heterocedasticidad. La dimensión temporal da cuenta de la evolución o cambios en el tiempo; siendo por lo general relativamente más corta, los problemas de autocorrelación son menos frecuentes. Este tipo de datos facilita el análisis de situaciones tales como flujos migratorios, efecto de programas de entrenamiento, movimiento de trabajadores entre sectores y son ideales para la EI, pero son costosos y requieren capacitación institucional.

En la EI, el objetivo es determinar la magnitud, si alguna, del impacto del programa, el proyecto o la política sobre la variable de interés, a sabiendas de que todo programa tendrá un efecto distinto sobre los individuos, precisamente debido a las diferencias inherentes a cada uno de ellos. La EI busca en definitiva estimar el efecto promedio del programa para toda la población o Efecto Tratamiento Promedio (ETA), que es la diferencia en la variable respuesta bajo el programa o tratamiento (Y_{it} -Y_{it^*}) con respecto a la situación de control (Y_{it^*}):

$$E[Y_{it} - Y_{it^*}] \qquad (1)$$

También es posible estimar la ganancia promedio que el programa aporta a los participantes (P=1), o Efecto Tratamiento Promedio para los Tratados (ETT)

$$E[Y_{it} - Y_{it^*} | P=1] \qquad (2)$$

Varias son las razones para el auge de los modelos de datos panel (MDP): reconocimiento de las dificultades para estimar modelos de comportamiento individual con datos agregados de series de tiempo; preocupación por las distorsio-

nes que las diferencias individuales introducen en las estimaciones obtenidas a partir de encuestas de corte transversal; avances en las técnicas econométricas, software y disponibilidad de este tipo de datos. Entre las ventajas de los MDP destaca el poder tomar en cuenta de manera explícita la **heterogeneidad no observable**, reduciendo el posible sesgo que ella genera, sin tener que recurrir a variables dicotómicas; el mejor aprovechamiento de la información; menor riesgo de colinealidad; permite estudiar **dinámicas de ajuste**, modelos con retardos, relaciones intertemporales, modelos de ciclo de vida e intergeneracionales, permanencia en el tiempo de fenómenos como desempleo, pobreza; identifica y cuantifica efectos no posibles de detectar con datos de corte transversal o con series de tiempo (comparación de situaciones sin-con o antes-después de una intervención); permite construir y probar modelos de **comportamiento** relativamente más **complejos** sin recurrir a muchas restricciones (eficiencia técnica, cambio tecnológico, economías de escala); **reducen sesgo** de agregación, al recoger información de micro unidades (individuos, firmas, hogares), aumentando la precisión de las estimaciones. Los problemas que se pueden presentar son los inherentes a toda muestra: cobertura, datos faltantes, espaciamiento, sesgos temporales, errores de medida, auto-selección, atrición (Baltagi, 1995; Wooldridge 2002).

Estructura básica de los MDP

Se busca determinar si las variaciones observadas en y se deben a cambios en las variables explicativas, tomando en cuenta las diferencias individuales

$$y_{it} = x'_{it}\,\beta + \varepsilon_{it}, \qquad (3)$$

donde β es el vector de parámetros, X_{it} es un vector de k variables explicativas; i: 1, …, n denota las unidades muestrales, t: 1, …, T indica los periodos, k: 1, …, K representa las covariables, y åit los errores aleatorios. Particularmente, en el caso de la evaluación de un programa o proyecto, cuando menos una de las variables explicativas es una dicotómica que toma valores 1 y 0 para distinguir los individuos sometidos al programa (tratamiento), de aquellos en el grupo de control. Existen diversas fuentes de variabilidad:

a. **efecto individuo:** generalmente invariante en el tiempo. Representa el impacto directo de todas las características individuales **no observables** e invariantes en el tiempo sobre y_{it}.
b. **efecto tiempo:** que puede asumirse invariante entre individuos; cada periodo tiene efectos específicos no observables.
c. **efecto individuo-tiempo**: efectos cambiantes que pueden ser tanto determinísticos como estocásticos.

Suponer que los coeficientes son iguales para los n individuos y/o los t instantes de tiempo puede resultar muy restrictivo. Por su parte, el caso extremo, donde se asume β_{kit} variantes entre individuos y en el tiempo puede ser imposible

de manejar. Se requiere por tanto introducir algunos supuestos (Greene, 2007; Wooldridge, 2002):

Supuesto 1: Exogeneidad contemporánea: x_t y ε_τ ortogonales en el sentido de la media condicional (este supuesto restringe la relación para el mismo periodo de tiempo). $E(\varepsilon_\tau \mid x_t) = 0$, para t= 1, 2,... t.

Supuesto 2: Exogeneidad estricta: extiende la restricción de ortogonalidad en la relación en cualquier periodo. $E(\varepsilon_t \mid x_1, x_2, \ldots, x_t) = 0$, para t= 1, 2,... t.

El supuesto 1 puede resultar difícil de mantener cuando existen variables omitidas. Ello constituye la principal motivación para utilizar datos panel. Así (3) se transforma en

$$y_{it} = X'_{it}\,\beta + Z'_{i}\,\alpha + \varepsilon_{it} \qquad (4)$$

$$y_{it} = X'_{it}\,\beta + \alpha_i + \varepsilon_{it} \qquad (5)$$

donde Zi contiene un término constante y un conjunto de características individuales (observadas o no), todas invariantes en el tiempo. Tanto y como x, z, son variables aleatorias. Interesa conocer los efectos parciales de las variables explicativas x_j, incluido el programa, manteniendo z constante, para lo cual existen tres opciones bien definidas:

a. Si z_i contiene únicamente variables observadas para todos los Y_{it}, o si sólo contiene un término constante, se trata como un tradicional modelo de regresión (estimación por Mínimos Cuadrados Ordinarios, MCO).

b. Si z_i contiene variables no observadas no correlacionadas con Y_{it}, éste será otro factor no observable afectando aY, pero no sistemáticamente relacionado con x y absorbido por el error. Se plantea un modelo de efectos aleatorios en el que el método de MCO genera estimadores consistentes.

c. Si z_i contiene variables no observadas, correlacionadas con x_{it}, MCO genera estimadores sesgados e inconsistentes de , en cuyo caso se plantea un modelo de efectos fijos.

Modelos de efectos fijos y efectos aleatorios

No debe olvidarse que el principal objetivo de los modelos de datos panel es precisamente capturar la heterogeneidad no observable y que es ignorada en los tradicionales modelos de regresión y que puede de alguna manera afectar la estimación de los efectos de las variables *x* sobre *y*.

Modelos de efectos fijos (MEF)

El MEF asume que las diferencias entre los individuos pueden ser capturadas a través de diferencias en el término constante, lo que equivale a asumir estas variaciones como determinísticas. Así, efecto fijo significa $cov(x_{it},z_i)\neq 0$. Siendo que se trata de variables no observadas, la heterogeneidad individual se recoge a

través de un conjunto de n-1 variables dicotómicas (d_i), cuyos coeficientes asociados α_i indican las diferencias individuales con respecto al individuo de referencia y se estiman conjuntamente con las pendientes β_k.

Sin embargo, la inconveniencia de este método para el caso de grandes tamaños de muestra hace que por lo general se tienda a anular el efecto individuo trabajando las variables en desviaciones con respecto a la media temporal de cada individuo, lo cual a su vez impide analizar el efecto de variables invariantes en el tiempo (Baltagi, 1995):

$$(\tilde{y}_{it} \; \bar{y}_{i.}) = (\tilde{X}'_{it} \; \bar{X}_i)\beta + \alpha_i + (\tilde{\varepsilon}_{it} \; / \bar{\varepsilon}_{i.}) \qquad (6)$$

para el cual, el estimador por Mínimos Cuadrados Generales de los parámetros de pendiente viene dado por $\hat{\beta} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y)$, con $\Omega^{-1/2} = [I_n \otimes \Sigma]^{-1/2}$.

Modelos de efectos aleatorios (MEA)

Hasta ahora hemos intentado controlar o eliminar $Z'_i \; \alpha$, pues se consideraba que z_i estaba correlacionada con x_{it}. Si la heterogeneidad individual no observada se asume no correlacionada con las variables incluidas en la matriz x (**$cov(x_{it},z_i)=0$**), el efecto individual constante puede considerarse se distribuye aleatoriamente entre los individuos del corte transversal. En ese caso se tiene un MEA, que puede expresarse como:

$$y_{it} = X'_{it}\,\beta + \eta_{it} \qquad (7)$$

$$\eta_{it} = \alpha_i + \varepsilon_{it} \qquad (8)$$

donde α_i es el componente individual aleatorio similar a ε_{it} pero invariante en el tiempo (within), mientras it es ruido blanco.

Así, en el MEA se supone que i es una variable aleatoria inobservable independiente de x_{it} y que por tanto forma parte del término de perturbación compuesto, por lo que a este tipo de modelos también se les conoce como modelos de error compuesto. El término de error puede incluir también un componente temporal aleatorio, invariable entre individuos (between) $\eta_{it} = \alpha_i + \phi_t + \varepsilon_{it}$; sin embargo, por lo general se asume $\phi_t = 0$. Este tipo de modelos es adecuado cuando se trabaja con muestras muy grandes, extraídas de una población suficientemente grande, donde asumir interceptos diferentes puede resultar muy complejo. Al igual que en caso anterior, la estimación se hace a través de MCG (Baltagi, 1995).

Para la decisión acerca de cuál modelo utilizar se cuenta con test específicos, siendo típico el Test de especificación de Hausman, el cual prueba la ortogonalidad de los efectos aleatorios. En efecto, bajo la hipótesis nula, la no correlación entre z_i y x_{it}, los estimadores MCG del MEA son consistentes y eficientes, en contraste con la hipótesis alternativa de correlación diferente de cero entre ambas matrices de variables, en cuyo caso es conveniente aplicar MEF.

Variables instrumentales en modelos de efectos aleatorios

Recordemos que en MEF se asume que x_{it} no contiene elementos que no varíen en el tiempo (condición de rango completo). Las características de este tipo son absorbidas por el efecto fijo o individual. El MEA, por su parte, asume corr[z_i, x_{it}]= 0, pero permite que el modelo contenga características invariantes en el tiempo.

Hausman y Taylor proponen un método para sobreponerse al primer inconveniente, tomando ventaja del segundo. Sea

$$y_{it} = X1'_{it}\,\beta_1 + X2'_{it}\,\beta_2 + Z1'_i\,\alpha_1 + Z2'_i\,\alpha_2 + U_i + \varepsilon_{it} \qquad (9)$$

$X1_{it}$: k_1 variables que varían en t no correlacionadas con u_i

$X2_{it}$: k_2 variables que varían en t correlacionadas con u_i

$Z1_i$: l_1 variables que no varían en t no correlacionadas con u_i

$Z2_i$: l_2 variables que no varían en t correlacionadas con u_i

Este método basado en variables instrumentales, permite estimar MEA donde algunas de las covariables están correlacionadas con efectos aleatorio individuales no observados.

Ejemplos de aplicación:

Si bien por lo general la EI se basa en datos de corte transversal dado su relativo menor costo de recolección, existen en la literatura múltiples ejemplos de EI con base en datos panel (Klugman, 2002; White, 2011):

a. Programa de Salud en Filipinas: una muestra aleatoria de 274 hogares con niños seleccionados de manera aleatoria observada en dos periodos de tiempo diferentes (1975 y 1979) para analizar el impacto sobre la salud infantil de un programa gubernamental de asistencia médica, incluyendo planificación familiar, en 20 villas de la provincial rural de Laguna en dicho país. El estudio compara, entre otras cosas, los cambios en la salud de niños sometidos al tratamiento con los de otros no atendidos por estos centros asistenciales, utilizando un modelo de efectos fijos para corregir por las diferencias propias de cada una de las comunidades consideradas. Sus resultados indican un crecimiento significativo en la estatura de los niños de familias sometidas al programa de salud y planificación familiar.

b. Construcción de escuelas en Indonesia: Utiliza un modelo de diferencia en diferencia para estimar el efecto de un programa de construcción de 61.000 escuelas primarias en los periodos 1973-74 y 1978-79 sobre la educación y los salaries en dicho país. Básicamente el modelo compara los resultados de regiones no beneficiadas por el programa con aquéllas en las que se llevó a cabo la construcción de escuelas, usualmente las más habitadas. Los resultados evidencian un incremento de entre 0,12 y 0,19 en los años de estudio y un incremento en los sueldos de entre 1,5% y 2,7% entre los beneficiados por el programa.

c. Programa JOVEN de Argentina: Utiliza datos panel para determinar el efecto de un programa de capacitación y formación para el empleo desarrollado por el Ministerio del Trabajo y la Seguridad Social de Argentina entre los años 1993 y 1999, el cual benefició a más de 11.000 participantes provenientes, por lo general, de familias de bajos ingresos. Los resultados sugieren un incremento de hasta el 25% en la probabilidad de conseguir empleo entre los participantes, pero una desmejora en la remuneración.

Conclusiones

Los datos panel proporcionan información de individuos en distintos momentos de tiempo, lo que los hace ideales para evaluar los efectos de programas sociales y proyectos al permitir datos del grupo de tratamiento y del grupo de control antes y después de la intervención. Los modelos de datos panel tienen la gran ventaja de permitir controlar por variables no observadas que pueden de alguna manera afectar el comportamiento de la variable respuesta, y a la vez permiten modelar dinámicas de ajuste y diferencias de comportamientos vitales en la evaluación de impactos.

Esta endogeneidad individual puede o bien controlarse a través de modelos de efectos fijos, o puede asumirse aleatoria en los modelos de efectos aleatorios. La desventaja de los primeros sobre los segundos es que no permiten conocer el efecto sobre la variable respuesta de factores invariantes en el tiempo. La desventaja de los modelos de efectos aleatorios está en imponer la ortogonalidad entre estos factores tiempo-invariantes y las demás covariables. La combinación de las bondades de ambos enfoques es posible a través de modelos de efectos aleatorios con variables instrumentales.

Referencias Bibliográficas

Baltagi, Badi (1995). Econometric Analysis of Panel Data. 1ª Edición. Wiley. USA.

Greene, William (2007). Econometric Analysis. 6th Ed. (3ª en Español). Prentice Hall. USA.

Klugman, Jeni (2002). A Sourcebook for Poverty Reduction Strategies. The World Bank. USA.

White, Howard (2011). An introduction to the use of randomized control trials to evaluate development interventions. WP 9. International Initiative for Impact Evaluation. New Delhi. India.

Wooldridge, Jeffrey (2002). Econometric Analysis of Cross Section and Panel Data. The MIT Press. Reino Unido.