Econometría I

Carlos Velasco[1]

[1]Departamento de Economía
Universidad Carlos III de Madrid

Econometría I
Máster en Economía Industrial
Universidad Carlos III de Madrid
Curso 2007/08

Outline

1. ¿Qué es la Econometría?
2. Etapas del Análisis Económico Empírico
3. La Estructura de los Datos Económicos
   - Datos de Corte Transversal
   - Series Temporales
   - Datos fusionados de Sección Cruzada
   - Datos de Panel o Longitudinales
4. Causalidad y análisis ceteris paribus

Objetivos

1. Distinción entre modelo económico y modelo econométrico.
2. Distinción entre variables endógenas y exógenas.
3. Distinción entre diferentes procesos generadores de datos.

Bibliografía

Wooldridge (2006). Capítulo 1.
Goldberger (2001). Capítulo 1.

¿Qué es la Econometría?

Ciencia basada en el desarrollo de modelos probabilísticos y métodos de inferencia estadística para estudiar fenómenos económicos, teniendo en cuenta la particular naturaleza de los datos económicos (**observacionales**) para estimar relaciones causales entre variables económicas, el contraste de teorías económicas y la evaluación y diseño de políticas gubernamentales y empresariales.

Aplicaciones de la Econometría

1. Predicción de variables macroeconómicas (tasas de inflación y desempleo, crecimiento del PIB, tipos de interés, etc.).
2. Modelización macroeconómica: relación entre inflación y desempleo, producción y masa salarial, oferta y demanda.
3. Modelización microeconómica: relación entre inversión en educación y salario; producción y factores productivos (f. de producción); proporción de gasto en un bien y renta; gastos en publicidad y ventas; programas de reciclaje y prob. de desempleo.
4. Finanzas: volatilidad condicional; valoración de activos.
5. Historia, sociología, psicología, etc.

Etapas del Análisis Económico Empírico

1. Modelo económico.
2. Datos y variables. Proceso Generador de Datos.
3. Modelo econométrico.

Modelo Económico

Un modelo económico describe relaciones entre diferentes variables mediante ecuaciones matemáticas. El modelo explica el comportamiento de un conjunto de variables $Y = (Y_1, \ldots, Y_m)'$ en términos de otras variables $X = (X_1, \ldots, X_k)'$ que se determinan fuera del modelo.

- $Y_1, \ldots, Y_m$ son las **variables endógenas**: se obtienen como la solución de un sistema de ecuaciones

$$\begin{aligned} f_1(Y_1, \ldots, Y_m, X_1, \ldots, X_k) &= 0 \\ &\vdots \\ f_m(Y_1, \ldots, Y_m, X_1, \ldots, X_k) &= 0 \end{aligned}$$

  donde cada función $f_1, \ldots, f_m$ puede depender de todas o sólo parte de las variables Y y las X.
- $X_1, \ldots, X_k$ son las **variables exógenas.**
- Las funciones $f_1, \ldots, f_m$ representan el comportamiento de los agentes económicos determinado en base a una optimización.

Ejemplos

Modelo económico para el crimen

Becker (1968): "Crime and punishment: an economic approach" *Journal of Political Economy* 76, pp. 169-217.
Modelo económico: describe la participación de los individuos en actividades criminales mediante un análisis de maximización de utilidad y asignación óptima de recursos. Modelo:

$$Y = f\left(X_1, X_2, X_3, X_4, X_5, X_6, X_7\right)$$

Y	$\equiv$	horas dedicadas a actividades delictivas
X_1	$\equiv$	"salario" por hora de la actividad delictiva
X_2	$\equiv$	"salario" por hora en empleo legal
X_3	$\equiv$	otras rentas
X_4	$\equiv$	probabilidad de ser detenido
X_5	$\equiv$	probabilidad de ser condenado después de ser detenido
X_6	$\equiv$	duración esperada de la condena en caso de ser condenado
X_7	$\equiv$	edad

Variables y Datos

Las variables de un modelo económico $Z = (Y', X')'$ representan aspectos del comportamiento de los agentes a nivel individual o agregado.
El economista observa, directamente o de forma aproximada, el comportamiento y las características de los agentes, que organiza en lo que llamamos datos, la evidencia empírica,

$$\text{Datos} \quad \rightarrow \quad \mathcal{Z}_n = \{z_1, \ldots, z_n\}$$

donde

$$z_i = \begin{pmatrix} y_{1i} \\ \vdots \\ y_{mi} \\ x_{1i} \\ \vdots \\ x_{ki} \end{pmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix} y \\ \vdots \\ y \end{matrix}}\right\} y_i \\ \left.\vphantom{\begin{matrix} x \\ \vdots \\ x \end{matrix}}\right\} x_i \end{array}$$

Proceso Generador de Datos

- Dependiendo de la naturaleza de los datos (cómo se han recogido las observaciones, relación entre ellas, etc.) el economista utiliza un modelo teórico que puede explicar la forma en que se generan esos datos. Este modelo, en general desconocido, es probabilístico y se conoce como Proceso Generador de Datos (PGD).
- Debido a la naturaleza aleatoria del modelo, se considera que el vector de variables de interés, $Z = (Y', X')'$ es un vector de variables aleatorias (v.a.s) o un vector de procesos estocásticos. La distribución conjunta de Z es el PGD, y cada dato (observación i ó t) es una realización de esta v.a..

Modelos Econométricos

Un modelo econométrico se construye para cuantificar y contrastar las relaciones entre variables económicas postuladas por un modelo económico a partir de la evidencia empírica (los datos).

Características de un modelo econométrico:

1. Reconoce el carácter estocástico que gobierna las relaciones entre variables.
2. Postula una forma funcional que depende de parámetros, los cuales se definen (identifican) a partir de la información que proporciona la teoría económica, o el sentido común, y/o supuestos probabilísticos no contrastables.
3. El modelo debe tener en cuenta que hay otros muchos factores que afectan a la decisión y que en general no son observables o identificables.

Para un conjunto de datos $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ es de esperar que la inmensa mayoría de ellos no cumpla lo que especifica un modelo económico. Es decir, para muchos $i = 1, \ldots, n$

$$\begin{aligned} f_1 \left(y_{1i}, \ldots, y_{mi}, x_{1i}, \ldots, x_{ki} \right) &\neq 0 \\ &\vdots \\ f_m \left(y_{1i}, \ldots, y_{mi}, x_{1i}, \ldots, x_{ki} \right) &\neq 0, \end{aligned}$$

para cualquier conjunto de funciones $f_1, \ldots, f_m$ no trivial ($\neq 0$).
Sin embargo, al ser las variables estocásticas, siempre se puede encontrar un conjunto de funciones que satisfaga dicha relación en media,

$$\begin{aligned} E \left[f_1 \left(Y_1, \ldots, Y_m, X_1, \ldots X_k \right) \right] &= 0 \\ &\vdots \\ E \left[f_m \left(Y_1, \ldots, Y_m, X_1, \ldots X_k \right) \right] &= 0. \end{aligned}$$

El hecho de que las variables X sean exógenas, se puede reconocer explícitamente, exigiendo que las funciones f_j cumplan que

$$\begin{aligned} E\left[f_1\left(Y_1,\ldots,Y_m,X_1,\ldots X_k\right)|X_1=x_1,\ldots,X_k=x_k\right] &= 0 \\ &\vdots \\ E\left[f_m\left(Y_1,\ldots,Y_m,X_1,\ldots X_k\right)|X_1=x_1,\ldots,X_k=x_k\right] &= 0 \end{aligned}$$

para todo vector $(x_1,\ldots,x_k)'$.

■ Esto se expresa en término de un **error** (aleatorio), $E\left[\varepsilon_j\right]=0$,

$$f_j\left(Y_1,\ldots,Y_m,X_1,\ldots X_k\right) = \varepsilon_j,\ j=1,\ldots,m,$$

de tal forma que los errores $(\varepsilon_1,\ldots,\varepsilon_m)$ son inobservables y satisfacen

$$E\left[\varepsilon_j|X_1=x_1,\ldots,X_k=x_k\right]=0,\ \ j=1,\ldots,m$$

para todo $(x_1,\ldots,x_k)$. Para valores fijos de $X_1,\ldots,X_k$, los términos de error gobiernan la aleatoriedad de la relación entre las variables, haciendo que las relaciones sean sólo exactas en media.

■ Para cuantificar estas relaciones a partir de los datos, se propone una **forma funcional** para las funciones f_j que depende de un vector de **parámetros**. El modelo econométrico general se expresa como

$$\begin{aligned} g_1\left(Y_1, \ldots, Y_m, X_1, \ldots X_k; \theta_1\right) &= \varepsilon_1 \\ &\vdots \\ g_m\left(Y_1, \ldots, Y_m, X_1, \ldots X_k; \theta_m\right) &= \varepsilon_m, \end{aligned}$$

donde $g_1, \ldots, g_m$ son funciones conocidas y $\theta_1, \ldots, \theta_m$ son parámetros desconocidos.

■ La naturaleza del modelo y la interpretación de los parámetros dependen de los **supuestos** que se hagan sobre el término de error en relación a las variables exógenas.

Los supuestos habituales son

$$\varepsilon_1, \ldots, \varepsilon_m \text{ son independientes de } X_1, \ldots, X_k$$

$$\Downarrow$$

$$E\left[\varepsilon_j | X_1, \ldots, X_k\right] = 0, \quad j = 1, \ldots, m$$

$$\Downarrow$$

$$E\left[\varepsilon_j \varphi_j\left(X_1, \ldots, X_k\right)\right] = 0, \quad j = 1, \ldots, m, \ \varphi_j : \mathbb{R}^k \to \mathbb{R}.$$

En función del supuesto que se considere válido estaremos ante diferentes modelos (regresión clásica, regresión, correlación) con diferentes implicaciones en el significado e interpretación de los parámetros (problema de la *identificación*) y con consecuencias sobre los métodos de *estimación* apropiados dado un conjunto de datos.

Las ambigüedades inherentes al modelo económico se resuelven especificando un modelo econométrico como

$$
\begin{aligned}
crimen &= \beta_0 + \beta_1 salario_\ell + \beta_2 otrarenta + \beta_3 frecde \\
&\quad + \beta_4 freccon + \beta_5 durmed + \beta_1 edad + \varepsilon
\end{aligned}
$$

donde

$$
\begin{aligned}
crimen &\equiv \text{medida de actividad criminal } \rightarrow Y \\
salario_\ell &\equiv \text{"salario" por hora en empleo legal } \rightarrow X_2 \\
otrarenta &\equiv \text{otros ingresos} \rightarrow X_3 \\
frecde &\equiv \text{frecuencia de detenciones por infracciones anteriores} \rightarrow \\
freccon &\equiv \text{frecuencia de condenas } \rightarrow X_5 \\
durmed &\equiv \text{duración media de las sentencias anteriores} \rightarrow X_6 \\
\varepsilon &\equiv \text{término de error} \\
\beta_j,\ j &= 0, \ldots, 6 \quad \text{parámetros del modelo.}
\end{aligned}
$$

Ejemplos

Productividad y cursos de formación

El objetivo es estudiar los efectos de los programas de formación en la productividad de los trabajadores. Es obvio que elementos como educación, experiencia y cursos de formación, entre otros, afectan a dicha productividad, y que en general los trabajadores reciben salarios en relación a su productividad, lo que llevaría a un modelo como

$$salario = f\left(educ,\ exper, formación\right)$$

donde

$$\begin{aligned} salario &\equiv \text{salario por hora} \\ educ &\equiv \text{años de escolarización} \\ exper &\equiv \text{años de experiencia laboral} \\ formación &\equiv \text{semanas empleadas en cursos de formación} \end{aligned}$$

- Un modelo econométrico para este problema sería

$$salario = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 formación + \varepsilon$$

  donde el término ε contiene factores como"habilidad", calidad de la educación, antecedentes familiares y todos los factores que pueden afectar al salario de un trabajador.
- Si estamos interesados en el efecto de los programas de formación, β_3 es el parámetro de interés.
  La interpretación del parámetro β_3 dependerá de la relación entre la variable *formación* y los factores contenidos en ε, ya que la decisión (o la elegibilidad) en tomar cursos puede depender de dichos factores (al igual que del nivel de educación y/o exper.).
- Si dichos factores son *independientes* la *formación*, el parámetro β_3 nos informará de la influencia de la formación sobre el salario del trabajador, manteniendo todos los demás factores que afectan al salario constantes.

Son datos estáticos, referidos a un periodo fijo, sobre individuos de una población, generalmente provenientes de encuestas sobre familias, empresas, etc.

Muestreo Aleatorio: Se considera el experimento entrevistar a una persona al azar en una población de individuos. El espacio muestral, Ω, de sucesos individuales, es el conjunto de todos los individuos de la población. Se considera la v.a.

$$Z = Z(\omega), \quad \omega \in \Omega; \;\; Z : \Omega \to \mathbb{R}^{m+k},$$

que asigna a cada elemento del espacio muestral, cada individuo, un vector de números en $\mathbb{R}^{m+k}$. La función Z tiene una función de distribución conjunta relacionada con el comportamiento y características de los individuos en la población.

También se pueden consider las observaciones de $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ como las realizaciones IID de n v.a. independientes $\{Z_1, \ldots, Z_n\}$.

526 trabajadores en el año 1976: salario, años de educación, años de experiencia laboral, indicador de mujer e indicador de casado.

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
⋮	⋮	⋮	⋮	⋮	⋮
526	3.50	14	5	1	0

Series Temporales

- Son observaciones de una variable o varias variables a lo largo de varios periodos de tiempo (años, meses, etc.). Ejemplos: precio de activos, tipos de interés, tipos de cambio, tasa de inflación, oferta de dinero, PIB, tasa de desempleo, ventas de una empresa, etc.
- Las observaciones **no son independientes:** la evolución temporal puede explotarse con fines predictivos.
- La **frecuencia** con la que se observan los datos es muy importante.

$$\text{Datos} \quad \rightarrow \quad \mathcal{Z}_T = \{z_1, \ldots, z_T\}$$

con

$$z_t = \left(y_t', x_t'\right)' = \left(y_{1t}, \ldots, y_{mt}, x_{1t}, \ldots x_{kt}\right)'.$$

Puerto Rico: salario mínimo, tasa media de cobertura, desempleo y PNB.

obsno	year	avgmin	avgcov	unemp	gnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	20.7	16.0	925.0
⋮	⋮	⋮	⋮	⋮	⋮
38	1987	3.35	58.2	16.8	4496.7

- Características de ambos tipos de datos, sección cruzada y series temporales.
- Diferente secciones cruzadas de la misma población obtenidas en diferentes instantes del tiempo.
- Permite aumentar el tamaño muestral.
- Sobre todo permite estudiar cambios en las variables más relevantes.

Datos fusionados de Sección Cruzada

Ejemplo

Precios de viviendas para dos años:

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85.500	42	1600	3	2.0
2	1993	67.300	36	1440	3	2.5
3	1993	134.000	38	2000	4	2.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
250	1993	243.600	41	2600	4	3.0
251	1995	65.000	16	1250	2	1.0
252	1995	182.400	20	2200	4	2.0
253	1995	97.500	15	1540	3	2.0
⋮	⋮	⋮	⋮	⋮	⋮	
5320	1995	57.200	16	1100	2	1.5

Paneles de datos (o datos longitudinales)

- Consisten en una serie temporal por cada unidad de sección cruzada. La longitud de las series temporales (T) suele ser mucho más corta que el número de unidades de sección cruzada (n).
- Ejemplos: dado un panel de individuos, podemos tener la serie histórica de salarios, educación y empleo durante 10 años (alguna variable puede que no cambien en el tiempo!); dado un panel de empresas, podemos tener observaciones de sus variables principales (producción, demanda de factores, tamaño, etc.) durante el mismo periodo temporal. En vez de típicas secciones cruzadas podemos tener países o regiones.

$$\text{Datos} \quad \rightarrow \quad \mathcal{Z}_{Tn} = \{z_{i1}, \ldots, z_{iT}, \ i = 1, \ldots, n\}$$

con

$$z_{it} = \left(y_{it}', x_{it}'\right)' = \left(y_{1it}, \ldots, y_{mit}, x_{1it}, \ldots x_{kit}\right)'.$$

Paneles de datos (cont)

- Son más difíciles de conseguir.
- Aportan más información y permite responder a preguntas que las secciones cruzadas no pueden
- Permiten incluir una estructura temporal en el razonamiento económico

Datos de panel de dos años de estadísticas de delincuencia en las ciudades:

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64330	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
299	150	1986	25	543000	4.5	520
300	150	1990	32	546200	5.2	493

Causalidad y análisis ceteris paribus

- Objetivo: descubrir si una variable tiene un **efecto causal** sobre otras variables.
- La relación causal es muy diferente de una simple asociación o correlación
- **Ceteris paribus:** otros factores (relevantes) siendo iguales, es un concepto clave en el análisis causal. Sin este concepto no se pueden medir efectos causales.
- Práctica: imposible realizar experimentos c.p. en Economía.
- Técnicas Econométricas: simulan tales experimentos a partir de datos observacionales.

- La cantidad de fertilizante es sólo uno de los factores que afectan al rendimiento del cultivo.
- Por lo tanto debemos plantearnos la cuestión como un problema ceteris paribus.
- Posible experimento: asignar diferentes cantidades de fertilizante sobre diferentes parcelas, medir el rendimiento (datos de corte transversal) y utilizar técnicas estadísticas para medir el grado de asociación.
- Sin embargo, no es posible elegir parcelas exactamente iguales que sólo cambien por el nivel de fertilizante.
- La clave es cómo se eligen los niveles de fertilizantes en cada parcela, independientemente de otros factores, o con cierta conexión.

- Problema: *si elegimos una persona al azar entre la población, y le damos un año más de educación, ¿cuánto aumentaría su salario?*
- Se trata también de una cuestión **ceteris paribus:** todos los otros factores constantes.
- Posible experimento realizado por un planificador social: seleccionar un grupo de personas, y atribuir a cada uno un nivel aleatorio de educación y midamos sus salarios.
- En caso de asignación independiente, el experimento servirá para medir la relación causal, en caso contrario no.
- En la práctica esto es posible, y podremos controlar otros factores importantes, como *experiencia*, pero otros serán más problemáticos, como el nivel de *habilidad*, que influye tanto en el nivel de educación como en el de salarios.
- Incluso si no cotrolamos por experiencia es posible diseñar métodos para medir efectos causales.