Econometría

Magister en Estadística

La econometría se ocupa de la medición de las relaciones económicas. Es una integración de la economía, la economía matemática y la estadística con el objetivo de proporcionar valores numéricos a los parámetros de las relaciones económicas. Las relaciones de las teorías económicas se expresan generalmente en formas matemáticas y se combinan con la economía empírica. Los métodos econométricos se utilizan para obtener los valores de los parámetros que son esencialmente los coeficientes de la forma matemática de las relaciones económicas. Los métodos estadísticos que ayudan a explicar el fenómeno económico se adaptan como métodos econométricos.

Las relaciones econométricas representan el comportamiento aleatorio de las relaciones económicas que generalmente no se consideran en las formulaciones económicas y matemáticas. Cabe señalar que los métodos econométricos pueden utilizarse en otras áreas como las ciencias de la ingeniería, las ciencias biológicas, las ciencias médicas, las geociencias, las ciencias agrícolas, etc. En pocas palabras, siempre que sea necesario encontrar la relación estocástica en formato matemático, los métodos y herramientas econométricas son de gran ayuda. Las herramientas econométricas son útiles para explicar las relaciones entre

Un modelo es una representación simplificada de un proceso del mundo real. Debe ser representativo en el sentido de que debe contener las características más destacadas de los fenómenos en estudio. En general, uno de los objetivos de la modelización es tener un modelo simple para explicar un fenómeno complejo. Tal objetivo puede conducir a veces a un modelo demasiado simplificado y a veces las suposiciones hechas son poco realistas. En la práctica, generalmente, todas las variables que el experimentador considera relevantes para explicar el fenómeno se incluyen en el modelo.

El resto de las variables se vierten en una cesta llamada ≪perturbaciones≫, donde las perturbaciones son variables aleatorias. Esta es la principal diferencia entre la modelización económica y la modelización econométrica. Esta es también la principal diferencia entre la modelización matemática y la modelización estadística. La modelización matemática es exacta por naturaleza, mientras que la modelización estadística también contiene un término estocástico.

Un modelo económico es un conjunto de supuestos que describe el comportamiento de una economía o, más en general, un fenómeno. Un modelo econométrico consta de

- Un conjunto de ecuaciones que describen el comportamiento. Estas ecuaciones se derivan del modelo económico y tienen dos partes: variables observadas y perturbaciones.
- una declaración sobre los errores en los valores observados de las variables.
- una especificación de la distribución de probabilidad de las perturbaciones.

Formulación y especificación de modelos econométricos

Los modelos económicos se formulan en una forma empíricamente verificable. De un modelo económico pueden derivarse varios modelos econométricos, que difieren en la elección de la forma funcional, la especificación de la estructura estocástica de las variables, entre otros aspectos.

Estimación y prueba de modelos

Los modelos se estiman a partir de un conjunto de datos observados y se someten a pruebas para evaluar su idoneidad. Este proceso implica inferencia estadística y el uso de diversos procedimientos de estimación para determinar los valores numéricos de los parámetros desconocidos. Con base en distintas formulaciones estadísticas, se selecciona el modelo más adecuado.

Uso de modelos

Los modelos obtenidos se utilizan para la predicción y formulación de políticas, lo cual es esencial en la toma de decisiones económicas. Estas predicciones permiten a los responsables de políticas evaluar la calidad del modelo ajustado y tomar medidas necesarias para reajustar las variables económicas relevantes.

¿Qué son los estudios observacionales?

- Son estudios en los que el investigador no asigna los tratamientos, sino que observa las exposiciones naturales.
- Se utilizan cuando no es posible realizar experimentos controlados aleatorizados (RCTs) por razones éticas o prácticas.

Tipos de estudios observacionales

- **Estudios de cohorte**: Se sigue a un grupo de individuos en el tiempo para observar la relación entre una exposición y un resultado.
- **Estudios de casos y controles**: Se comparan sujetos con una condición (casos) con sujetos sin ella (controles) para analizar antecedentes.
- **Estudios transversales**: Se analizan datos en un solo punto en el tiempo, útiles para medir prevalencia.

El marco causal de Rubin (MCR) es un marco estructurado para la inferencia causal, llamado Holland (1986) por el trabajo anterior de Rubin. El MCR consta de dos componentes principales y un tercero opcional. El primer componente consiste en definir los efectos causales utilizando los resultados potenciales en todos los escenarios. Este paso esboza la base científica, que requiere la consideración explícita de las intervenciones que definen los tratamientos cuyos impactos causales se van a estimar.

Para definir los efectos causales, hay tres conceptos básicos: unidades, tratamientos y resultados potenciales. Una unidad es un objeto físico, por ejemplo, una persona, en un momento determinado. Un tratamiento es una acción que puede o no aplicarse a una unidad. Nos centramos en el caso de dos tratamientos, aunque la ampliación a más de dos tratamientos es sencilla en principio, aunque no necesariamente con datos reales. Asociados a cada unidad hay dos resultados potenciales: el valor de una variable de resultado Y en el futuro si se aplica el tratamiento activo, y el valor de Y en el mismo punto futuro si se aplica el tratamiento de control. El objetivo es conocer el efecto causal de aplicar el tratamiento activo en comparación con el control en Y, donde, por definición, el efecto causal es una comparación de los dos resultados potenciales.

Por ejemplo, la unidad podría ser un estudiante universitario actualmente matriculado, el tratamiento activo podría participar en un programa intensivo de tutoría, y el control no podría participar. El resultado Y podría ser la nota media al final del semestre, siendo los dos resultados potenciales la nota media con y sin tutoría intensiva; el efecto causal de recibir tutoría intensiva frente a no recibirla es la comparación de la nota media del estudiante al final del semestre con y sin tutoría.

Sea W el tratamiento asignado a una unidad, donde $W = 1$ denota el tratamiento activo y $W = 0$ el control. Definamos $Y(1)$ como el resultado potencial en el tratamiento activo y $Y(0)$ en el control. El efecto causal del tratamiento activo en relación con el control implica comparar $Y(1)$ con $Y(0)$, a menudo a través de su diferencia $Y(1) - Y(0)$, la diferencia logarítmica $\log[Y(1)] - \log[Y(0)]$, u otra métrica como su ratio. La observación se limita a $Y(1)$ o $Y(0)$ en función de W: $Y_{\text{obs}} = WY(1) + (1 - W)Y(0)$. Un reto fundamental de la inferencia causal, conocido como el problema fundamental de la inferencia causal (véase Holland (1986)), es que sólo vemos el resultado potencial del tratamiento administrado a cada unidad, mientras que el resultado del tratamiento alternativo permanece inobservable.

Por lo tanto, la inferencia causal se enfrenta fundamentalmente al problema de los datos que faltan, con al menos el 50 % de los datos de resultados potenciales no disponibles. Para más detalles sobre el problema fundamental de la inferencia causal, véase Brumback (2021) y Imbens (2024).

El problema fundamental de la inferencia causal se centra en la imposibilidad de observar simultáneamente los resultados potenciales para una misma unidad. Este concepto se puede expresar formalmente como: Para cada unidad i, observamos:

$$Y_i = \begin{cases} Y_i(1) & \text{si } W_i = 1 \\ Y_i(0) & \text{si } W_i = 0 \end{cases} \tag{1}$$

Esto significa que nunca podemos observar el efecto causal individual:

$$\tau_i = Y_i(1) - Y_i(0) \tag{2}$$

Para entender esto mejor, consideremos un ejemplo: si una persona toma un medicamento (W = 1), observamos su resultado bajo tratamiento Y(1), pero no podemos observar qué habría sucedido si no lo hubiera tomado Y(0) en ese mismo momento. Esta situación se conoce como el problema de los contrafactuales. Para superar esta limitación, la inferencia causal moderna se centra en efectos promedio en poblaciones o subpoblaciones, como:

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{3}$$

ATT

ATT mide el impacto medio de un tratamiento en las personas que realmente lo han recibido. A diferencia del efecto promedio del tratamiento (Average Treatment Effect, ATE), que mide el efecto de un tratamiento en toda la población del estudio, el ATT se centra específicamente en aquellos a los que se les administró la intervención o el tratamiento. Esta distinción es crucial, ya que a menudo se alinea más estrechamente con contextos prácticos de toma de decisiones en los que el interés principal radica en comprender el efecto de un tratamiento en aquellos que ya han estado expuestos a el.

La identificación de estos efectos requiere supuestos adicionales, como el de no confusión:

$$(Y_i(0), Y_i(1)) \perp W_i | X_i \tag{4}$$

Este problema fundamental motiva el desarrollo de métodos estadísticos para estimar efectos causales y destaca la importancia del diseño experimental, donde la asignación aleatoria del tratamiento permite estimar efectos causales sin sesgo.

CATE

El efecto condicional medio del tratamiento (CATE) es un constructo estadístico empleado en la inferencia causal para cuantificar el impacto causal medio de un tratamiento o intervención en una variable de resultado de interés. CATE delinea un tipo particular de efecto causal que los investigadores pueden tratar de medir. CATE se refiere al efecto medio del tratamiento o exposición dentro de un subgrupo específico. La validez de la estimación depende de la pertenencia a este subgrupo, distinguiéndola de otros efectos causales, como el efecto promedio del tratamiento (ATE), que se refiere a toda la población en estudio.

CATE

En concreto, CATE significa la diferencia en el valor esperado de la variable de resultado entre dos grupos de individuos, que divergen únicamente en su estado de tratamiento pero que, por lo demás, son comparables en cuanto a sus características observadas o covariables. La naturaleza condicional de CATE pone de manifiesto que el efecto del tratamiento puede variar entre los diferentes subgrupos de la población, dependiendo de sus valores de covariables. La estimación de CATE implica generalmente el empleo de un modelo estadístico que correlacione la variable de resultado con el estado de tratamiento y las covariables, seguido de la comparación de los resultados previstos para las cohortes tratadas y no tratadas dentro de cada subgrupo delineado por las covariables.

Ignorabilidad

La ignorabilidad es una condición que debe cumplirse para que el efecto causal promedio de un tratamiento esté bien definido e identificable a partir de datos observacionales. Esta condición también se conoce como la suposición de no confusión o la condición de independencia.
La ignorabilidad puede expresarse como

$$Y(t) \perp T|X \tag{5}$$

Ignorabilidad

Esto indica que, una vez que tenemos en cuenta un conjunto de covariables observadas X, el resultado potencial $Y(t)$ no depende del tratamiento T. En otras palabras, una vez que controlamos X, el tratamiento asignado T es tan aleatorio como si se hubiera asignado en un experimento aleatorio. Esto permite a los investigadores estimar el efecto causal del tratamiento utilizando datos observacionales. T representa la variable de tratamiento en el conjunto de datos. Por otro lado, t se refiere a un nivel o valor específico del tratamiento.

Ignorabilidad

La suposición de ignorabilidad representa una hipótesis sólida que con frecuencia elude la validación empírica debido a las limitaciones de los datos; sin embargo, su plausibilidad puede establecerse dentro de contextos específicos basados en un conocimiento sustantivo profundo del dominio del estudio.
Tenga en cuenta que cuando se cumple la ignorabilidad, se logra la intercambiabilidad. La intercambiabilidad significa que los grupos de tratamiento y de control son comparables en términos de expectativas de resultados si ambos grupos hubieran recibido el mismo nivel de tratamiento. Esto también implica que no hay factores no medidos que influyan tanto en la selección del tratamiento como en el resultado.

Un factor de confusión es una variable que influye tanto en la variable de tratamiento X como en el resultado Y. La presencia de un factor de confusión Z puede llevar a conclusiones erróneas sobre la relación causal entre X y Y si no se controla adecuadamente. Esto se expresa a menudo a través de diagramas causales o modelos de regresión. Por ejemplo, si Z causa tanto X como Y, entonces la asociación entre X y Y puede deberse parcial o totalmente a Z.

- Variable de tratamiento: Fumar (X).
- Resultado: Cáncer de pulmón (Y).
- Factor de confusión: Factores genéticos o exposición a contaminación ambiental (Z).
- *Modelos*

$$X = \alpha_0 + \alpha_1 Z + \nu \tag{6}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon \tag{7}$$

- Variable de tratamiento: Años de educación (X).
- Resultado: Salario (Y).
- Factor de confusión: Habilidad cognitiva o nivel socioeconómico de la familia (Z).
- *Modelos*

$$X = \gamma_0 + \gamma_1 Z + \nu \tag{8}$$

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + \varepsilon \tag{9}$$

- Una variable moderadora afecta la dirección o intensidad de la relación entre una variable independiente (X) y una dependiente (Y).
- Puede ser cualitativa (género, grupo social) o cuantitativa (nivel de ansiedad, edad).
- Matemáticamente, se expresa como:

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 (X \cdot M) + \varepsilon \tag{10}$$

- La interacción $\beta_3(X \cdot M)$ indica si M modera la relación entre X e Y.

Derivada en la moderación

- El efecto marginal de X sobre Y es:

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 M \tag{11}$$

- Si $\beta_3 \neq 0$, entonces el efecto de X sobre Y cambia con M.
- Geométricamente, M altera la pendiente de la relación entre X e Y.

- Una variable mediadora explica el mecanismo por el cual X afecta a Y.
- Se representa con el siguiente sistema de ecuaciones:

$$M = \alpha_0 + \alpha_1 X + \varepsilon_M \tag{12}$$

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 M + \varepsilon_Y \tag{13}$$

- El efecto total de X sobre Y es:

$$\frac{\partial Y}{\partial X} = \gamma_1 + \gamma_2 \alpha_1 \tag{14}$$

- γ_1 es el efecto directo de X sobre Y.
- $\gamma_2 \alpha_1$ es el efecto indirecto de X sobre Y a través de M.

Sea M un mediador entre X e Y. El efecto directo de X sobre Y, denotado como DE, es la parte del efecto de X sobre Y que no pasa a través de M. Por otro lado, el efecto indirecto, denotado como IE, representa la influencia de X en Y que opera a través de M. Estos efectos pueden expresarse matemáticamente como:

$$\text{DE} = \frac{\partial Y}{\partial X}\Big|_{M}, \tag{15}$$

$$\text{IE} = \frac{\partial M}{\partial X} \cdot \frac{\partial Y}{\partial M}. \tag{16}$$

donde $\frac{\partial Y}{\partial X}\Big|_{M}$ representa el cambio en Y debido a un cambio unitario en X, manteniendo constante M.

Bajo ciertas suposiciones de diferenciabilidad, podemos escribir la ecuación en términos de la regla de la cadena:

$$\frac{dY}{dX} = \frac{\partial Y}{\partial X} + \frac{\partial Y}{\partial M} \cdot \frac{\partial M}{\partial X}. \tag{17}$$

Esto implica que el impacto total de X sobre Y se puede descomponer en el impacto directo de X sobre Y más el efecto indirecto mediado por M.

Un mediador es una variable que explica total o parcialmente el efecto de una variable independiente X (también conocida como variable explicativa o de tratamiento) sobre una variable dependiente Y (variable de resultado o respuesta) en un análisis estadístico. Es decir, el mediador es el mecanismo a través del cual X ejerce su influencia sobre Y.

Dado el modelo:

$$M = aX + u, \tag{18}$$

$$Y = bM + cX + v. \tag{19}$$

Sustituyendo la ecuación de M en la de Y:

$$Y = b(aX + u) + cX + v. \tag{20}$$

Reordenando:

$$Y = abX + cX + bu + v. \tag{21}$$

Tomando la derivada respecto a X:

$$\frac{dY}{dX} = ab + c. \tag{22}$$

Por lo tanto, el efecto total de X sobre Y es TE $= ab + c$, donde c representa el efecto directo y ab el efecto indirecto mediado por M.

Consideremos un modelo más complejo en el que X afecta a Y tanto directa como indirectamente a través de dos mediadores M_1 y M_2:

$$M_1 = aX + u_1, \tag{23}$$
$$M_2 = dM_1 + eX + u_2, \tag{24}$$
$$Y = bM_1 + fM_2 + cX + v. \tag{25}$$

- Encuentre la expresión del efecto total TE de X sobre Y.
- Identifique los efectos directo e indirectos en este caso.

Sustituyendo M_1 en la ecuación de M_2:

$$M_2 = d(aX + u_1) + eX + u_2 = daX + eX + du_1 + u_2. \quad (26)$$

Sustituyendo M_1 y M_2 en la ecuación de Y:

$$Y = b(aX + u_1) + f(daX + eX + du_1 + u_2) + cX + v. \quad (27)$$

Reordenando:

$$Y = abX + fdaX + feX + cX + bu_1 + fdu_1 + fu_2 + v. \quad (28)$$

Tomando la derivada respecto a X:

$$\frac{dY}{dX} = ab + fda + fe + c. \quad (29)$$

Así, el efecto total de X sobre Y es TE $= ab + fda + fe + c$.

- El término c representa el efecto directo de X sobre Y.
- ab es el efecto indirecto a través de M_1.
- $fda + fe$ es el efecto indirecto que involucra ambos mediadores.