Índice general

Lección 9. El puente: de los datos a las variables aleatorias

Marcos Bujosa

19 de septiembre de 2026

Resumen

"Ninguno de los dioses filosofa ni desea hacerse sabio, porque ya lo es [...] Tampoco los ignorantes filosofan ni desean hacerse sabios [...] Eros está en el medio de la sabiduría y de la ignorancia." (Platón, *El Banquete*, 203e–204a)

La econometría habita en ese mismo espacio intermedio: entre la ignorancia absoluta y el conocimiento total de la población. Esta lección da el salto desde la geometría de las muestras (lo que vemos) al espacio de las variables aleatorias (el modelo subyacente que queremos inferir).

- (slides) — (html) — (pdf)

1 El puente epistemológico: entre Eros y los datos

> "Ninguno de los dioses filosofa ni desea hacerse sabio, porque ya lo es [...] Tampoco los ignorantes filosofan ni desean hacerse sabios [...] Eros está en el medio de la sabiduría y de la ignorancia." — Diotima a Sócrates (Platón, *El Banquete*)

- *El Dios*: Conoce toda la población (certeza absoluta). No necesita inferir nada.
- *El Ignorante*: No tiene datos ($n = 0$). No puede afirmar nada.
- *El Econometrista*: Como Eros, habita el espacio intermedio. Tiene una *muestra imperfecta* ($n < N$) y el *deseo de inferir* la estructura verdadera del mundo.

Hasta ahora: *geometría exacta* sobre una muestra finita $\boldsymbol{y} \in \mathbb{R}^n$. A partir de hoy: la muestra $\boldsymbol{y}$ se concibe como *una realización particular* (una "sombra sensible") de un modelo abstracto superior: una *variable aleatoria* Y.

Advertencia, desde ya: lo que sigue es una **analogía**, no una demostración. Y solo es válida para variables aleatorias con esperanza y varianza definidas — *no todas la tienen.*

El puente epistemológico: entre Eros y los datos

Esta analogía platónica encierra el problema epistemológico de la econometría. En las primeras lecciones de este curso nos hemos dedicado a describir muestras: hemos calculado medias, varianzas

y rectas de regresión para conjuntos de datos concretos ($\boldsymbol{y} \in \mathbb{R}^n$). En ese nivel, no había incertidumbre: si medimos la covarianza de nuestra muestra, el número que obtenemos es un hecho geométrico exacto.

Sin embargo, en economía y política rara vez nos interesa la muestra por sí misma. Si analizamos una muestra de 1.000 salarios, no lo hacemos porque nos importen exclusivamente esos 1.000 trabajadores concretos; lo hacemos porque queremos descubrir el mecanismo subyacente que determina los salarios en *toda la población* o en el *proceso de generación de salarios*. Queremos descubrir la idea abstracta detrás de los datos concretos.

El dios que conociera el mecanismo exacto no necesitaría la estadística. El ignorante absoluto sin datos no podría aplicarla. El econometrista actúa en el medio: usa datos contingentes (la muestra) para aproximarse racionalmente al mecanismo subyacente (el modelo poblacional).

El vector de datos $\boldsymbol{y} \in \mathbb{R}^n$ que hemos usado hasta ahora es un conocimiento cerrado, sin resquicios de ignorancia — sabemos exactamente cuánto vale cada y_i, y por tanto sabemos exactamente cuánto vale $\mu_{\boldsymbol{y}}$, $\sigma^2_{\boldsymbol{y}}$, o cualquier otra cantidad calculada a partir de él; pero es un conocimiento incompleto, limitado a la muestra de datos de que disponemos. Cuando aspiramos a decir algo sobre el salario "en general", o sobre el precio de la vivienda "en general" — no solo de los pocos casos que observamos, sino de la relación subyacente que sospechamos que existe y que ni observamos ni observaremos jamás en su totalidad —, no estamos en la posición del sabio. Pero tampoco estamos en la del ignorante total: no partimos de la nada, tenemos una muestra, un vector concreto, una teoría económica que orienta la búsqueda. Estamos, como Eros, en el medio: con datos parciales pero con el deseo (y las herramientas) de inferir algo que va más allá de una colección de datos parciales.

La *variable aleatoria* es el objeto matemático que formaliza esa posición intermedia. A partir de esta lección, el vector de datos $\boldsymbol{y}$ deja de ser nuestro único objeto de estudio. Pasaremos a considerarlo como una *realización particular*, una "fotografía" momentánea, de un ente matemático abstracto superior: una *variable aleatoria* Y. Asumiremos que nuestros n datos son el resultado de observar la realización de n copias idénticas e independientes de esa variable Y.[1] Y la esperanza matemática $\mathrm{E}(Y)$ no es, como veremos, algo que calculemos con exactitud a partir de una tabla de números: es un número teórico, generalmente desconocido, que solo podemos *inferir* —nunca poseer del todo— a partir de nuestra muestra. Ahí radica el "deseo filosófico" del econometrista: no tiene la certeza del sabio (no conoce $\mathrm{E}(Y)$), pero tampoco actúa a ciegas como el ignorante (tiene un vector de datos y un método, la geometría que hemos construido, para intentar acercarse a esa certeza inalcanzable).

Al dar este paso, la intuición geométrica que hemos construido en $\mathbb{R}^n$ no desaparece; todo lo que sigue en esta sesión es una *analogía*, deliberadamente construida en paralelo con la geometría de las lecciones 4 a 8, pero *no* es una demostración de que las variables aleatorias se comporten como los vectores de $\mathbb{R}^n$. Hay claras diferencias y excepciones, que iremos señalando conforme aparezcan. Y hay, además, una restricción teórica muy importante que conviene decir en voz alta desde ya: no todas las variables aleatorias tienen esperanza matemática definida, y entre las que la tienen, no todas tienen varianza definida. Aquí trabajaremos exclusivamente con variables aleatorias con

[1] por realización de una variable aleatoria entenderemos observar un único valor tomado por dicha variable aleatoria de entre todos los que puede tomar.

esperanza y varianza definidas, pues son solo estas las que permiten replicar la estructura geométrica de las lecciones anteriores.

Para quien quiera ver hasta qué punto la intuición cotidiana sobre "el valor esperado" puede fallar cuando la esperanza no está bien definida, recomiendo el vídeo "The Two Envelope Problem — a Mystifying Probability Paradox": un ejercicio mental sencillo de enunciar, y sorprendentemente resbaladizo, porque la esperanza matemática involucrada no está bien definida.

Para profundizar:

- Platón, *Banquete*, 203e–204a (discurso de Diotima sobre la naturaleza de Eros).
- Wooldridge, J. M. (2020), Apéndice C.1–C.2: repaso de variable aleatoria, función de distribución y esperanza matemática desde la estadística clásica.

2 Un vector es una función; y una variable aleatoria también

Un vector $\boldsymbol{y} \in \mathbb{R}^n$ (lista ordenada de números —Lección 2) es literalmente una *función* que asigna un número a cada índice $i \in \{1, \ldots, n\}$.

Una *variable aleatoria* Y es también una función — pero definida sobre un conjunto Ω (de "sucesos elementales"), generalmente más grande y sin una descripción explícita.

$\mathbb{R}^n$	Variables aleatorias
dominio $\{1, \ldots, n\}$	dominio Ω
$\boldsymbol{y}(i) = y_i; \quad i \in \{1, \ldots, n\}$	$Y(\omega) = y; \quad \omega \in \Omega$
vector = lista de n números	v.a. = función sobre Ω

Lo que *sobrevive* al cambio:

La *combinación lineal* (lección 2) de variables aleatorias $\beta_1 \mathit{1} + \beta_2 X$ sigue siendo una variable aleatoria, exactamente como $\beta_1 \mathbf{1} + \beta_2 \boldsymbol{x}$ era un vector.

Nota: *1* es la variable aleatoria constante que asigna 1 a todo suceso elemental de Ω.

Un vector es una función; y una variable aleatoria también

En la lección 2 definimos un vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ como una lista ordenada de números; pero esta definición nos permite afirmar que un vector de $\mathbb{R}^n$ es, sencillamente, una *función* que asigna a cada índice i, del conjunto $\{1, 2, \ldots, n\}$, un número real y_i.

Una *variable aleatoria* Y es, formalmente, el mismo tipo de objeto: una función que asigna un número real a cada elemento de un conjunto. La diferencia está en el dominio. En $\mathbb{R}^n$, el dominio es el conjunto finito y perfectamente conocido $\{1, \ldots, n\}$ (los índices de nuestras n observaciones). En el caso de una variable aleatoria, el dominio es un conjunto Ω —llamado convencionalmente "espacio de sucesos elementales"— que puede ser enorme, incluso infinito, y cuyos elementos concretos casi nunca se describen (salvo en los cursos de probabilidad para presentar ejemplos triviales como el lanzamiento de un dado o una moneda). $Y(\omega)$ denota el valor numérico que la variable aleatoria Y asigna al elemento $\omega \in \Omega$. Como las variables aleatorias pueden sumarse y multiplicarse por escalares para obtener nuevas variables aleatorias, también forman un espacio vectorial (aunque ya no es el espacio $\mathbb{R}^n$ de las primeras lecciones).

Aquí aparece ya la primera diferencia real entre ambos mundos, y conviene señalarla sin rodeos: mientras que en $\mathbb{R}^n$ conocemos perfectamente el dominio (los índices $1, \ldots, n$), en el caso de las variables aleatorias casi nunca sabemos, ni necesitamos saber, qué es exactamente Ω. Cuando decimos "el precio de una vivienda, P, es una variable aleatoria", no estamos pensando en un Ω concreto con sucesos elementales identificables uno a uno; estamos usando la maquinaria matemática de las variables aleatorias para modelizar de manera abstracta nuestra incertidumbre sobre ese precio, sin necesidad de precisar más.

Pese a esta diferencia importante, lo que sobrevive intacto a este cambio es lo que permite mantener la maquinaria que hemos usado hasta hoy: la *combinación lineal*, introducida ya en la lección 2. Del mismo modo que $\beta_1 \mathbf{1} + \beta_2 \boldsymbol{x}$ era, en $\mathbb{R}^n$, otro vector (obtenido sumando dos vectores tras multiplicarlos por escalares), la expresión $\beta_1\, \mathit{1} + \beta_2 X$ —donde X es una variable aleatoria y $\mathit{1}$ denota la variable aleatoria *constante* igual a 1 para todo suceso[2]— es otra variable aleatoria. Esta es la observación que permitirá, en unas transparencias, escribir el modelo de regresión $Y = \beta_1\, \mathit{1} + \beta_2 X + U$ con total naturalidad: es la misma operación algebraica de siempre, solo que ahora los objetos que combinamos ya no son listas de n números, sino funciones sobre un dominio Ω (casi nunca especificado explícitamente).

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice B.3: propiedades de linealidad de la esperanza matemática.

3 El producto escalar de la probabilidad

En $\mathbb{R}^n$: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_s = \sum_{j=1}^{n} \frac{x_j y_j}{n}$.

Con variables aleatorias, el papel de $\langle \cdot, \cdot \rangle_s$ lo juega la *esperanza del producto*:

$$\langle X, Y \rangle_P \;=\; \mathrm{E}(XY).$$

Advertencia honesta: $\langle \cdot, \cdot \rangle_P$ es solo un *semi*-producto escalar. $\mathrm{E}(X^2) = \|X\|_P^2 = 0$ NO implica $X = \mathit{0}$ (podría ser solo cero "casi seguro"; no lo desarrollaremos).

Nota: $\mathit{0}$ es la variable aleatoria constante que asigna 0 a todo suceso elemental de Ω.

Pero Pitágoras, Cauchy–Schwarz y la desigualdad triangular (lección 3) solo exigían simetría, linealidad y positividad — *no* positividad estricta. Así que *siguen siendo válidos*, sin cambiar una coma.

El producto escalar de la probabilidad

Toda la geometría que hemos construido desde la lección 3 —ortogonalidad, ángulos, Pitágoras, proyecciones— dependía de un único ingrediente: un producto escalar $\langle \cdot, \cdot \rangle$ con tres propiedades abstractas (simetría, linealidad en cada argumento, positividad). En la lección 4 vimos que, para

[2] es decir, $\mathit{1}(\omega) = 1$ para todo ω del dominio Ω.

hacer estadística en $\mathbb{R}^n$, se usa el producto escalar $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_s = \frac{1}{n} \langle \boldsymbol{x}, \boldsymbol{y} \rangle_e$, con el factor $\frac{1}{n}$ elegido precisamente para que $\|\mathbf{1}\|_s = 1$.

Para dar el salto de hoy necesitamos el mismo tipo de ingrediente: un producto escalar entre variables aleatorias. La elección natural —y la que hace que todo el paralelismo funcione— es la *esperanza del producto*:

$$\langle X, Y \rangle_P \;=\; \mathrm{E}(XY),$$

donde XY denota el producto *punto a punto* de las dos funciones X y Y, es decir, la función que a cada $\omega \in \Omega$ le asigna $X(\omega)Y(\omega)$. Es el análogo, en el mundo de las variables aleatorias, del producto componente a componente $\boldsymbol{x} \odot \boldsymbol{y}$ que en la lección 7 usamos para escribir la "fórmula de cálculo" de la varianza y la covarianza. Como en $\mathbb{R}^n$, la norma asociada a este producto escalar se define del mismo modo: $\|X\|_P = \sqrt{\langle X, X \rangle_P} = \sqrt{\mathrm{E}(X^2)}$.

- En cuanto a las propiedades relacionadas con la probabilidad, si denotamos con *1* la variable aleatoria constante 1 (aquella que asigna el valor 1 a todo suceso ω: $1(\omega) = 1$), también ocurre que $\|1\|_P = 1$, puesto que $\|1\|_P^2 = \langle 1, 1 \rangle_P \;=\; \mathrm{E}(1^2) = \mathrm{E}(1) = 1$.[3]
- En cuanto a las propiedades relevantes que permiten replicar el esquema geométrico: $\langle \cdot, \cdot \rangle_P$ hereda (aquí solo las enunciamos sin demostración alguna):
  - *Simetría*: $\mathrm{E}(XY) = \mathrm{E}(YX)$, ya que, como el producto de números reales es conmutativo: $X(\omega)Y(\omega) = Y(\omega)X(\omega)$.
  - *Linealidad*: $\mathrm{E}\Big((aX + bZ)Y\Big) = a\mathrm{E}(XY) + b\mathrm{E}(ZY)$, pues la esperanza matemática es una operación lineal.
  - *Positividad*: $\mathrm{E}(X^2) \geq 0$, porque $X^2(\omega) \geq 0$ para todo ω, y la esperanza de una función no negativa no puede ser negativa.

  Las tres se cumplen, pero hay una diferencia real que conviene señalar. En $\mathbb{R}^n$, un producto escalar propio exige además que $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ si y solo si $\boldsymbol{x} = \mathbf{0}$ (positividad *estricta*); y en efecto, $\langle \boldsymbol{x}, \boldsymbol{x} \rangle_s = \frac{1}{n} \sum x_i^2 = 0$ obliga a que cada componente de $\boldsymbol{x}$ sea cero. Con $\langle \cdot, \cdot \rangle_P$ esto ya no es cierto: $\mathrm{E}(X^2) = 0$ no implica que X sea la variable aleatoria nula en todo Ω; solo implica que X vale cero *con probabilidad uno*. Es decir, el conjunto de ω para los que $X(\omega) \neq 0$ tiene, en el sentido de la probabilidad, "tamaño" cero, aunque pudiera no estar vacío. A un objeto así, que cumple simetría, linealidad y positividad pero no positividad estricta, se le llama *semi*-producto escalar (o semi-norma, cuando hablamos de $\sqrt{\mathrm{E}(X^2)}$). No desarrollaremos esta sutileza —pertenece a un curso de probabilidad más formal—, pero es importante saber que existe, para no dar por sentado que el paralelismo con $\mathbb{R}^n$ es perfecto en todos sus detalles.

¿Qué salvamos, entonces, de todo el edificio construido en la lección 3? La buena noticia es que *todo*. Repasando las demostraciones de la lección 3 —Pitágoras, la desigualdad de Cauchy–Schwarz, la desigualdad triangular—, ninguna de ellas usó en ningún momento la propiedad de positividad *estricta*. Todas se apoyaban únicamente en simetría, linealidad y positividad (no estricta). Por tanto, esas tres demostraciones siguen siendo válidas, palabra por palabra, con $\langle \cdot, \cdot \rangle_P$ en lugar de $\langle \cdot, \cdot \rangle_s$: seguiremos pudiendo hablar de vectores ortogonales ($\langle X, Y \rangle_P = \mathrm{E}(XY) = 0$), de la norma

[3] usted ya conocía esta condición porque la ha visto en los cursos de probabilidad: *la integral de la función de densidad es igual a* 1 y, por tanto, $\mathrm{E}(1) = \int 1 \cdot f(x)\,dx = \int f(x)\,dx = 1$.

de una variable aleatoria ($\|X\|_P = \sqrt{\mathrm{E}(X^2)}$), y de proyecciones ortogonales, con el mismo aparato geométrico.

Para no sobrecargar la notación, se suele escribir sencillamente $\mathrm{E}(XY)$ en lugar de $\langle X, Y\rangle_P$ (así lo hace, de hecho, toda la literatura econométrica): pero conviene tener presente que la operación $\mathrm{E}(XY)$ tiene el mismo papel que $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_s$ en $\mathbb{R}^n$.

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice B.3: propiedades de linealidad de la esperanza.

4 La esperanza: dos caras de la misma moneda

En $\mathbb{R}^n$, el vector constante uno, $\mathbf{1}$, tuvo un papel especial; ahora lo tiene la v.a. cte. *1*.

	$\mathbb{R}^n$	Variables aleatorias
Escalar	$\mu_{\boldsymbol{y}} = \langle \boldsymbol{y}, \mathbf{1}\rangle_s$	$\mathrm{E}(Y) = \langle Y, 1\rangle_P$
Proyección	$\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$ (vector *constante*)	$E[Y \mid 1] = \mathrm{E}(Y)\ 1$ (función *constante*)

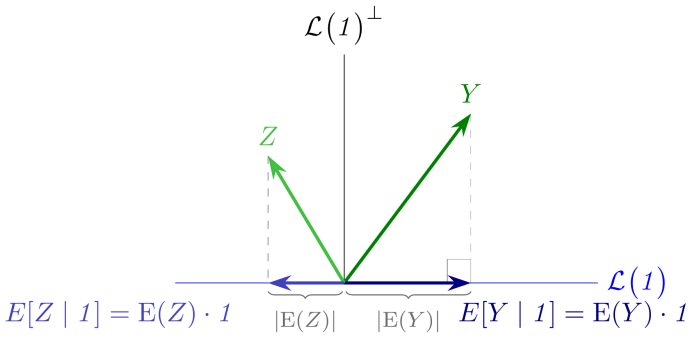

Figura 1: Proyección de la variable aleatoria Y sobre el subespacio generado por la variable constante *1*. El escalar de la proyección es $\mathrm{E}(Y)$ (la esperanza matemática o valor esperado), y el vector proyectado es la v.a. constante $E[Y|1] = \mathrm{E}(Y)\ 1$ (la esperanza condicional).

Por Pitágoras (como lección 5): $\mathrm{Var}(Y) = \left\|Y - E[Y|1]\right\|_P^2 = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$.

La esperanza: dos caras de la misma moneda

Empecemos por el concepto más simple: la proyección sobre las variables aleatorias constantes.

En la lección 4 proyectamos el vector de datos $\boldsymbol{y}$ sobre la recta generada por $\mathbf{1}$. Aquí hacemos lo mismo: proyectamos la variable aleatoria Y sobre el subespacio unidimensional generado por *1* (la variable aleatoria que siempre toma el valor 1).

Cuidado con la notación: En muchos manuales de estadística se usa el símbolo $\mathrm{E}(Y)$ tanto para referirse al *número* (el valor esperado, un escalar) como a la *variable aleatoria constante* (el vector

proyectado). Para no perder el rigor geométrico que tanto nos ha costado construir, haremos una distinción crucial:

- Usaremos $\mathrm{E}(Y)$ estrictamente como *un número* (el escalar que resulta del producto escalar $\langle Y, 1\rangle_P$).
- Usaremos $E[Y|1]$ como la *variable aleatoria proyectada* (el vector en el espacio).

La relación algebraica es idéntica a la que teníamos en el espacio euclídeo ($\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$): la variable proyectada es igual al escalar multiplicado por el vector generador; es decir, es una variable aleatoria constante:

$$E[Y|1] = \mathrm{E}(Y)\ 1$$

La diferencia entre el vector Y y su proyección es ortogonal al vector constante 1, lo que significa que *su esperanza es cero*[4]: $\mathrm{E}\big(Y - E[Y|1]\big) = 0$.

Y igual que hicimos en las lecciones 4 y 5, si aplicamos el teorema de Pitágoras a este triángulo rectángulo, vemos que el cuadrado de la hipotenusa, $\mathrm{E}(Y^2)$, es la suma de los cuadrados de los catetos, $(\mathrm{E}(Y))^2 + \mathrm{Var}(Y)$. Despejando, obtenemos la fórmula clásica de la varianza de la variable aleatoria Y (también llamada varianza poblacional): $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$.

La geometría no ha cambiado; solo hemos cambiado de espacio (ahora los objetos son variables aleatorias en lugar de listas de números).

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice B.3: definición y propiedades básicas de $\mathrm{E}(Y)$.

5 Esperanza condicional: proyección sobre un espacio ENORME

La *Esperanza Condicional*, $E[Y|X]$, es la proyección ortogonal de Y sobre el subespacio generado por las funciones de la variable X (¡Un espacio enorme!).

Descomposición del vector de datos $\boldsymbol{y}$	Descomposición variable aleatoria Y
Vector ajustado: $\widehat{\boldsymbol{y}}$	*Esperanza Condicional*: $E[Y \mid X]$
Residuo: $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$	*Perturbación*: $U = Y - E[Y \mid X]$

La esperanza condicional: proyección sobre un espacio ENORME

La regresión lineal simple que vimos en las lecciones 6, 7 y 8 consistía en proyectar los datos sobre el plano $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$. El análogo poblacional es proyectar la variable Y sobre el espacio de funciones de las variables aleatorias X (y la variable aleatoria 1 es, trivialmente, una de ellas). El vector resultante es la *Esperanza Condicional*, $E[Y|X]$; es decir, la variable aleatoria (de dicho espacio) más próxima a Y.

A la diferencia, $Y - E[Y|X]$, la llamaremos *perturbación* U (la componente de Y ortogonal a las funciones de X), y la interpretaremos en los modelos econométricos como "el resto de cosas que afectan a Y pero que son ortogonales a las funciones de X".

[4] puesto que $\big\langle (Y - E[Y|1]), 1 \big\rangle_P = \mathrm{E}\big((Y - E[Y|1])\ 1\big) = \mathrm{E}\big(Y - E[Y|1]\big) = 0$.

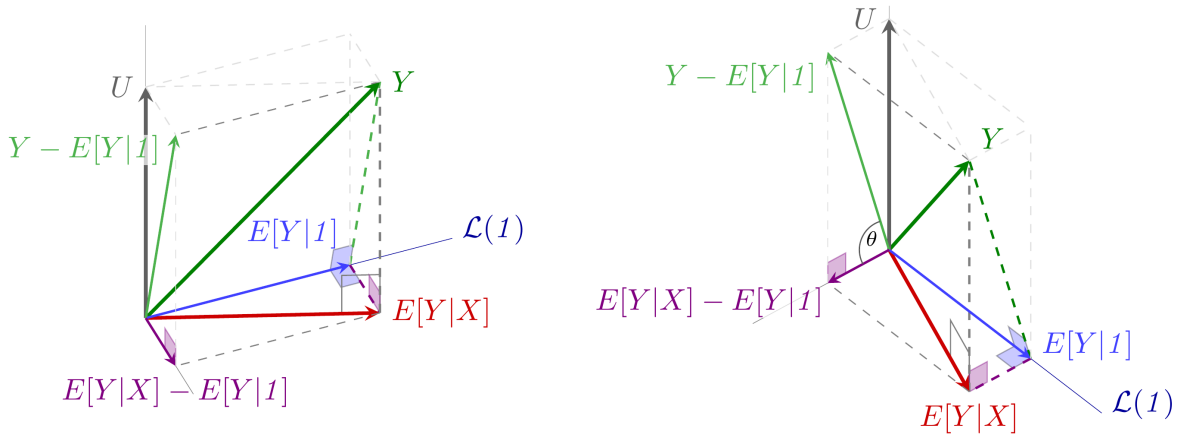

Figura 2: Proyección de Y sobre el espacio vectorial generado por las funciones de *1* y X. La perturbación U es ortogonal a dichas funciones. A la izquierda, el triángulo rectángulo de Pitágoras para las varianzas.

¡Atención, un abismo geométrico! En las primeras lecciones solo teníamos dos vectores de $\mathbb{R}^n$ ($\mathbf{1}$ y $\boldsymbol{x}$) que formaban un plano "muy dócil" de dos dimensiones. Pero en el mundo de las variables aleatorias, el subespacio sobre el que proyectamos no contiene únicamente las combinaciones lineales de *1* y X correspondientes al modesto plano bidimensional $\mathcal{L}(1, X)$.[5] También incluye todas las combinaciones lineales de toda una gigantesca familia de funciones que incluyen: X^2, $\log(X)$, e^X, $-\frac{1}{2}\exp\left(\sqrt{1+\frac{\pi}{7}\ln X + 5\sin^3(e^X)}\right)$, funciones a trozos, o cualquier inimaginable transformación punto a punto de X, siempre y cuando tenga varianza finita… En los cursos de probabilidad se dice que contiene *cualquier "función medible"* de X.

Consecuentemente, el subespacio sobre el que proyectamos para obtener $E[Y|X]$ no es un plano; es ¡un monstruo de dimensión infinita!

La *esperanza condicional* $E[Y|X]$ (la "sombra" de Y sobre dicho espacio) es, por definición, la función $g(X)$ (de entre todas las de ese enorme espacio) que resulta más próxima a Y en la norma $\|\cdot\|_P$ — es decir, la que hace mínima $\mathrm{E}\Big(\big(Y - g(X)\big)^2\Big)$ entre todas las funciones g posibles. Y aquí está la consecuencia que conviene subrayar con toda claridad: como el espacio sobre el que proyectamos es tan vasto, *no hay ninguna razón, en general, para que esa función más próxima sea una recta.* $E[Y|X]$ puede ser, en principio, cualquier curva: puede subir y luego bajar, puede tener escalones, puede curvarse de maneras arbitrariamente complicadas. La única propiedad que le exigimos, por definición, es ser la función de X que minimiza la distancia a Y — no que sea una combinación lineal de *1* y X.

La figura de la transparencia ilustra esta idea con el lenguaje visual de las lecciones 6 y 8. La variable aleatoria Y (verde) se descompone en dos partes ortogonales: su proyección $E[Y|X]$ (rojo)

[5] Al escribir $\mathcal{L}(1, X)$ estamos reutilizando la misma notación de lecciones anteriores, $\mathcal{L}(\cdot)$; pero ahora para denotar el subespacio vectorial formado por las combinaciones lineales de las variables aleatorias *1* y X (en lugar de por vectores de $\mathbb{R}^n$).

sobre el espacio de funciones de X, y $U = Y - E[Y|X]$ (gris), la componente perpendicular a dicha proyección —marcada, como en aquellas figuras, con un ángulo recto—, que llamamos *perturbación*. La diferencia esencial respecto a aquellas figuras es que aquí el "plano" sobre el que proyectamos ya no es un plano de dos direcciones. Es, informalmente, un espacio con tantísimas direcciones que ni siquiera podemos dibujarlo con fidelidad; la figura debe leerse como un esquema conceptual, no como una representación literal.

Esta observación no es un mero tecnicismo: es la que da sentido genuino al *primer supuesto* del Modelo Clásico de Regresión, que presentaremos en la próxima transparencia. Decir que "$E[Y|X]$ es una función *lineal* de X" —es decir, que $E[Y|X] = \beta_1\, 1 + \beta_2 X$— no es una trivialidad, como podría parecer si pensáramos (erróneamente) que $E[Y|X]$ vive por definición en el plano $\mathcal{L}(1, X)$ de las combinaciones lineales 1 y X. Es, por el contrario, un supuesto sustantivo: estamos afirmando que, de entre todas las infinitas formas funcionales posibles que $E[Y|X]$ podría tener, la función de X y 1 que mejor se aproxima a Y resulta ser la recta $\beta_1\, 1 + \beta_2 X$.

Para profundizar:

- Bujosa, M., sección "Esperanza condicional": definición de $E[Y|X]$ como proyección ortogonal sobre el subespacio probabilístico generado por X, con la advertencia (que aquí simplificamos) de que dicha proyección es una *clase de equivalencia* de variables aleatorias.
- Wooldridge, J. M. (2020), Apéndice B.4 y capítulo 2, sección 2.6: introducción intuitiva a la esperanza condicional y su papel en la especificación del modelo de regresión.

6 El Modelo Clásico de Regresión Lineal (MCRL)

El MCRL consiste en *imponer tres supuestos* para que el modelo sea manejable:

- *Linealidad de la esperanza condicional*: $E[Y|X]$ cae en el plano $\mathcal{L}(1, X)$:

$$E[Y|X] = \beta_1\, 1 + \beta_2 X.$$

  La exogeneidad sale gratis: $U = Y - E[Y|X]$ es, **por definición**, ortogonal a toda función de X. Así que aquí $E[U|X] = 0$ no es un supuesto: es un **corolario**.

  Por tanto $Y = \beta_1\, 1 + \beta_2 X + U$, con $\mathrm{E}(U) = 0$ y $\mathrm{Cov}(U, X) = 0$.

- *Independencia lineal de los regresores* ($\mathrm{Var}(X) \neq 0$)*:* X no es constante con probabilidad 1.
- *Homocedasticidad (varianza condicional constante)*: la "dispersión" de U no depende de X.

$$Var[U|X] = E[U^2|X] = \sigma^2\, 1$$

Nomenclatura:

- Y regresando
- X y 1 regresores
- U perturbación ("otras cosas").

Parámetros: Nótese que en el modelo no hay "sombreros"

- β_1 y β_2 son parámetros poblacionales desconocidos

(no son los $\hat{\beta}_1$ y $\hat{\beta}_2$ de la lección 7).

El Modelo Clásico de Regresión Lineal (MCRL)

Esta es una de las transparencias nucleares del curso. El "Modelo Clásico de Regresión Lineal" es simplemente un conjunto de reglas de juego geométricas que imponemos sobre un mecanismo poblacional que en realidad es desconocido.

Hemos definido la *perturbación* $U = Y - E[Y|X]$ como la diferencia entre Y y su proyección sobre el (enorme) espacio de las funciones de X. Por construcción, U es ortogonal (en el sentido de $\langle \cdot, \cdot \rangle_P$) a $E[Y|X]$, y a cualquier función de X. Con esta definición, podemos escribir siempre, sin ningún supuesto adicional:

$$Y \;=\; E[Y|X] + U.$$

Esta descomposición es, como todas las que hemos hecho en $\mathbb{R}^n$, puramente algebraica: se cumple *siempre*, sea $E[Y|X]$ una recta, una curva, o lo que sea. Ahora bien, el *Modelo Clásico de Regresión Lineal* consiste en añadir tres supuestos que, cuando se cumplen, simplifican radicalmente esta descomposición general y la reducen a algo mucho más manejable.

Antes de enunciarlos, conviene advertir que *tres* es un supuesto menos de los que encontrará en casi cualquier manual —Wooldridge, por ejemplo, enumera cinco para la regresión simple—. La diferencia no es un descuido ni un atajo: es consecuencia directa de haber definido U como lo que queda al proyectar. Al final de esta sección explicamos la equivalencia con detalle (apéndice "Tres supuestos aquí, cinco en Wooldridge").

1. *Linealidad:* Suponemos que

   $$E[Y|X] = \beta_1\, \mathit{1} + \beta_2 X,$$

   para ciertos números β_1, β_2. Con esto la descomposición $Y = E[Y|X] + U$ se convierte en

   $$Y \;=\; \beta_1\, \mathit{1} + \beta_2 X + U.$$

   Adviértase la ausencia de sombrero sobre β_1 y β_2. En la lección 7 calculamos $\hat{\beta}_1, \hat{\beta}_2$ a partir de un vector de datos concreto $\boldsymbol{y}, \boldsymbol{x} \in \mathbb{R}^n$, mediante las condiciones de ortogonalidad muestrales. Aquí, β_1 y β_2 son *parámetros poblacionales*: números fijos, pero *desconocidos*, que describen la relación teórica subyacente entre las variables aleatorias X e Y.

   ¿Cómo de realista es este supuesto?: Recuerde que el espacio sobre el que se proyecta Y es de dimensión infinita, así que la esperanza condicional $E[Y|X]$ podría ser casi cualquier función de X (por muy extraña o complicada que fuera). El Supuesto 1 impone que, "milagrosamente", la sombra de Y (la variable aleatoria función de X más próxima a Y) cae sobre el plano formado por la constante *1* y la variable X. Es decir, que $E[Y|X]$ es sencillamente la combinación lineal $\beta_1\, \mathit{1} + \beta_2 X$. ¿Es esto realista? Depende. Si sabemos (teorema en los cursos de probabilidad) que X e Y provienen de una distribución conjunta Normal, la matemática garantiza que la esperanza condicional siempre es lineal. Si no lo son, este supuesto supone una restricción fortísima (estamos imponiendo un corsé lineal a la realidad, pues estamos

imponiendo que de todas las funciones posibles entre $\mathit{1}$ y X, la que mejor se aproxima a Y es una combinación lineal de los regresores).

Corolario inmediato — la exogeneidad estricta ($E[U|X] = \mathit{0}$)[6]. Obsérvese que NO la enumeramos entre los supuestos, y la razón es que no hay nada que suponer: es una consecuencia de cómo hemos definido U. En efecto, puesto que $U = Y - E[Y|X]$ es, por construcción, aquello que queda tras proyectar ortogonalmente Y sobre el espacio de *todas* las funciones de X, la perturbación U es ortogonal a cualquier función de X; en particular, es simultáneamente ortogonal a $\mathit{1}$ y a X. De esa doble ortogonalidad se siguen dos consecuencias que emplearemos sin descanso en el resto del curso:

- Que $U \perp \mathit{1}$ implica que $\mathrm{E}(U\,\mathit{1}) = \mathrm{E}(U) = 0$.
- Y, recordando *el truco del vector de media nula* que vimos en la lección 6 para $\mathbb{R}^n$: si un vector tiene media cero, su producto escalar con otro vector es directamente la *covarianza* entre ambos. Usando el mismo razonamiento que en $\mathbb{R}^n$, y como U tiene valor esperado cero y es ortogonal a X ($\mathrm{E}(UX) = 0$), tenemos que $\mathrm{Cov}(U, X) = \mathrm{E}(UX) - \mathrm{E}(U)\mathrm{E}(X) = 0$; por tanto, la perturbación no está correlada con los regresores.

2. *Independencia lineal de los regresores con probabilidad* 1 *(*$\mathrm{Var}(X) \neq 0$*):* Este supuesto impone que, con probabilidad 1, X no toma un único valor, es decir, que X en desviaciones no es cero, y por tanto $\|X - \mathrm{E}(X)\,\mathit{1}\|_P^2 = \mathrm{Var}(X) \neq 0$[7]

   Con un único regresor, esta es toda la sustancia del segundo supuesto: una condición sobre la varianza de una sola variable. En la lección 10, al generalizar el modelo a varios regresores, significará algo más que eso, pues: dos o más regresores pueden tener cada uno varianza no nula por separado y, aun así, ser linealmente dependientes entre sí.

3. *Homocedasticidad:* Asumimos que la varianza de U, condicionada a X, es una función constante $\sigma^2\,\mathit{1}$. Por tanto, la dispersión de U no depende de X. Este supuesto no es estrictamente necesario para encontrar la recta que define a la esperanza condicional (como veremos), pero será fundamental más adelante para poder hacer inferencia y construir intervalos de confianza de forma sencilla.

   Fíjese que, usando $E[U|X] = \mathit{0}$ (el corolario del supuesto 1), tenemos que $Var[U|X] = E\big[(U - E[U|X])^2|X\big] = E[U^2|X]$, que es lo que aparece en la transparencia.

   Conviene notar, además, que al enunciar este tercer supuesto estamos asumiendo *de paso* algo que NO está garantizado en general: que la variable aleatoria U^2 tiene esperanza (recuerde que no toda variable aleatoria tiene esperanza definida). Que U tenga norma finita —es decir,

[6] Donde $\mathit{0}$ es la variable aleatoria constante cero: $\mathit{0} = 0\,\mathit{1}$ (i.e., a todo suceso elemental le asigna el valor 0).

[7] Técnicamente se debería precisar qué significa exactamente esta condición, porque aquí es fácil cometer un desliz. Podría parecer natural traducirla como "X no es un múltiplo de $\mathit{1}$", es decir, que X no sea, literalmente, la función constante $a\,\mathit{1}$ sobre todo el dominio Ω. Pero esa traducción es demasiado permisiva: recordemos la advertencia hecha al introducir $\langle\cdot,\cdot\rangle_P$ (transparencia "El producto escalar de la probabilidad"), sobre variables aleatorias iguales "casi seguro" pero no idénticas en todo punto. Una variable aleatoria puede coincidir con un valor fijo *a con probabilidad uno* —y por tanto tener varianza cero— sin ser, en sentido estricto, la función constante que para todo suceso elemental toma el valor a; basta con que difiera de a en un conjunto de sucesos elementales con probabilidad nula (de medida nula). La condición rigurosa, por tanto, nunca es "no ser literalmente un múltiplo de $\mathit{1}$", sino, sin rodeos, $\mathrm{Var}(X) \neq 0$: que X no sea igual, con probabilidad uno, a ningún valor fijo.

que $\mathrm{E}(U^2)$ exista, lo que ya necesitábamos para poder hablar de $\|U\|_P$— no es una propiedad que se herede sin más al multiplicar variables aleatorias entre sí. No entraremos en el detalle, que pertenece a un curso de probabilidad más formal; basta con saber que el supuesto de homocedasticidad da por sentada esa existencia.

- Apéndice: tres supuestos aquí, cinco en Wooldridge Si abre el manual de Wooldridge por el capítulo 2 encontrará *cinco* supuestos donde nosotros hemos puesto tres. Conviene entender por qué, porque no se trata de que uno de los dos textos se deje algo: son dos formas distintas de escribir el mismo modelo, y cada una paga un precio distinto.

  Los cinco supuestos de Wooldridge para la regresión simple (los llama SLR, por *Simple Linear Regression*) son:

  - *SLR.1 — Lineal en los parámetros*: en la población, $y = \beta_0 + \beta_1 x + u$.
  - *SLR.2 — Muestreo aleatorio*: se dispone de una muestra de n observaciones independientes extraídas de esa población.
  - *SLR.3 — Variación muestral en el regresor*: los x_i no son todos iguales.
  - *SLR.4 — Media condicional nula*: $\mathrm{E}(u \mid x) = 0$.
  - *SLR.5 — Homocedasticidad*: $\mathrm{Var}(u \mid x) = \sigma^2$.

  (En regresión múltiple los renombra MLR.1 a MLR.5, con MLR.3 convertido en "no colinealidad perfecta"; añade además MLR.6, normalidad, cuando llega a la inferencia. Volveremos sobre ello en las lecciones 10 y 12.)

  La correspondencia con los nuestros es casi inmediata: nuestro supuesto 2 ($\mathrm{Var}(X) \neq 0$) es SLR.3, y nuestro supuesto 3 (homocedasticidad) es SLR.5. SLR.2 no aparece en nuestra lista porque en este curso el muestreo aleatorio no es un supuesto *del modelo poblacional*, sino el marco que introduciremos en la lección 11, cuando pasemos de la población a la muestra. Queda, pues, una sola diferencia real: Wooldridge escribe *dos* supuestos —SLR.1 y SLR.4— donde nosotros escribimos *uno*, el de linealidad.

  Y la razón es la definición de u. Wooldridge *postula* la ecuación $y = \beta_0 + \beta_1 x + u$ con unos β que tienen un significado previo —típicamente causal o estructural—, y define u como "todo lo demás". Con esa definición, nada garantiza que u sea ortogonal a x: hay que *exigirlo*, y ese es exactamente el contenido de SLR.4. Nosotros hacemos el camino inverso: primero proyectamos, definimos $U = Y - E[Y|X]$, y solo después preguntamos si esa proyección resulta ser una recta. Con nuestra definición, la ortogonalidad de U frente a toda función de X es un teorema, no una exigencia; y todo el contenido sustantivo se concentra en un único sitio: que la *mejor* función de X resulte ser lineal.

  Dicho en una línea: **SLR.1 + SLR.4 juntos equivalen a nuestro supuesto 1**. No hemos suprimido nada; hemos reagrupado.

  ¿Qué se pierde al reagrupar? Esta es la parte honesta. La presentación de Wooldridge tiene una ventaja que la nuestra no tiene: como su u no está definido por proyección, la condición $\mathrm{E}(u \mid x) = 0$ *puede fallar*, y poder enunciar su fallo es justamente lo que permite hablar de **endogeneidad**: variable relevante omitida, simultaneidad entre y y x, error de medida en

el regresor. Son los problemas que motivan las *variables instrumentales* y buena parte de la econometría aplicada moderna. En nuestro marco esos fenómenos no se pueden ni siquiera formular: con U definido como el residuo de la proyección, $\text{Cov}(U, X) = 0$ se cumple *siempre*, y decir "la perturbación está correlada con el regresor" sería una contradicción en los términos.

Asumimos ese coste deliberadamente, porque este curso no estudia endogeneidad, variable omitida, simultaneidad ni variables instrumentales. Si los estudiara, la simplificación no compensaría: habría que volver a separar SLR.1 de SLR.4 y, además, redefinir U. Quien continúe con un curso de econometría avanzada encontrará allí esa separación, y le convendrá recordar esta página para saber que no hay contradicción entre ambos textos, sino dos puntos de partida distintos.

Para profundizar:

- Wooldridge, J. M. (2020), capítulo 2, secciones 2.1 y 2.5: presentación clásica del modelo de regresión simple y sus supuestos (SLR.1 a SLR.5).
- Wooldridge, J. M. (2020), capítulo 15: variables instrumentales — el capítulo que este curso no recorre y que es el que exige SLR.4 como supuesto separado.

7 Cuidado: Esta descomposición ortogonal no implica causalidad

La descomposición $Y = E[Y|X] + U$ *es un hecho algebraico* (una proyección ortogonal). *Siempre existe* (si las varianzas de X e Y son finitas).

Que el álgebra permita proyectar Y sobre las funciones de X NO significa que X cause o "explique" Y.

- *Sentido económico razonable:* $Consumo = f(Renta) + U$. Tiene sentido pensar que la Renta determina el Consumo, y que U agrupa otros factores (gustos, necesidades).
- *Causalidad absurda (pero aceptable algebraicamente):* $Superficie = f(Precio) + U$. Podemos proyectar la superficie de una casa sobre su precio. La matemática funcionará perfectamente. Pero asumir que "variaciones del precio cambian el tamaño de la casa" es absurdo.

La causalidad NO emana de los datos ni de la geometría. Debe provenir de una Teoría (Económica, Política, Sociológica, etc.) previa.

Cuidado: Esta descomposición ortogonal no implica causalidad

Esta es la advertencia más importante de todo el curso, y el motivo principal por el que hemos pospuesto hablar de modelos y causalidad hasta la novena lección.

La descomposición de una variable aleatoria en su proyección ortogonal $E[Y|X]$ y la perturbación U es un *hecho geométrico* sin interpretación *per se*. Pasa siempre. Y como tal, es completamente mudo respecto a la dirección de la causalidad.

En econometría se utilizan nombres muy sugerentes que pueden llevar a engaño: a Y se la llama "variable explicada" y a X "variable explicativa". Los nombres *regresando* y *regresor* son mejores, porque son neutros. Y la probabilidad está llena de términos con el mismo sesgo. *Esperanza*

sugiere que desvela lo que cabe esperar, cuando es solo la función de los regresores más cercana al regresando. *Variable aleatoria* no es una variable, sino una regla fija que asigna un número a cada elemento de Ω; y el *suceso elemental* no tiene nada que ver con que suceda algo. No se deje engañar por los nombres. Si hacemos una regresión del Consumo sobre la Renta Disponible, los nombres encajan con nuestra intuición macroeconómica: la renta "causa" (explica) el consumo.

Pero si invertimos el papel de las variables —buscando, por ejemplo, explicar la superficie de una vivienda en función de su precio, en lugar de al revés—, la maquinaria de proyección ortogonal calculará igualmente una recta, hallará un residuo perfectamente ortogonal al precio, y nos devolverá un R^2 positivo. Sin embargo, defender la relación causal en esa dirección (implicando que las alteraciones en el precio modifican la superficie de la vivienda) es ridículo. Es, de hecho, el mismo ejemplo —deliberadamente absurdo pero algebraicamente impecable— que ya vimos en la lección 8.

La estadística es ciega a la dirección temporal o lógica del universo. Consideremos cualquiera de estos cuatro casos:

- X "causa" Y;
- Y "causa" X;
- una tercera variable causa ambas;
- o no hay ninguna relación sustantiva entre X e Y.

Pues bien, la descomposición algebraica, por sí sola, es completamente indiferente a cuál de esos escenarios describe realmente el fenómeno que estamos estudiando.

Esta es la razón por la que en la lección 6 pospusimos deliberadamente la distinción entre "regresor" (una noción que atañe a la estructura del modelo, pero sin ningún compromiso causal) y "variable explicativa" (un concepto que sí lleva, implícita, la pretensión de que X desempeña algún papel explicativo o causal, aunque sea parcial, en la generación de Y). Cobramos hoy esa promesa: podemos llamar a X "regresor" con toda tranquilidad, en cualquier circunstancia, porque es una etiqueta puramente algebraica; pero solo deberíamos llamarlo "variable explicativa" —con esa carga causal— cuando dispongamos de alguna razón, *externa a la propia matemática*, que justifique semejante lectura.

¿De dónde puede venir dicha lectura? No de la geometría, ni del álgebra: únicamente de una *teoría económica previa*. Si, antes de mirar ningún dato, tenemos razones teóricas sólidas para pensar que la renta disponible de un consumidor influye en cuánto consume —y no al revés, ni por alguna tercera vía—, entonces sí podemos aventurar una lectura causal de la descomposición $Consumo = E[Consumo \mid Renta] + U$. Pero si, por el contrario, no tenemos ninguna teoría previa que distinga qué variable debería hacer de X y cuál de Y —como en el ejemplo, deliberadamente absurdo pero matemáticamente impecable, de regresar la superficie de una vivienda sobre su precio—, entonces la descomposición algebraica sigue siendo perfectamente válida, pero carece de cualquier interpretación causal razonable.

Conviene, además, prevenir una lectura equivocada de la transparencia anterior. Que hayamos enunciado *tres* supuestos en lugar de los cuatro o cinco habituales, y que la exogeneidad haya resultado ser un corolario gratuito, NO significa que el modelo clásico exija menos de la realidad,

ni que se cumpla más a menudo. Significa exactamente lo contrario de lo que podría parecer: todo el peso —todo lo que puede fallar— se ha concentrado en un único sitio, el supuesto 1. Es ahí, y solo ahí, donde afirmamos algo fuerte sobre el mundo: que de entre las infinitas funciones de X que podríamos haber usado para aproximar Y, la mejor resulta ser precisamente una recta. Cuando, en la última parte del curso, pongamos a prueba la *especificación* de un modelo, será ese supuesto el que estemos contrastando.

Cerramos esta advertencia recuperando, de forma explícita, el mantra con el que abrimos el curso en la lección 1: *correlación no implica causalidad*. Hoy, con el aparato construido en esta sesión, podemos precisar ese mantra con mayor rigor: tampoco la existencia de una esperanza condicional bien definida —ni, por supuesto, el cumplimiento de los tres supuestos del modelo clásico que hemos presentado— implica, por sí sola, ninguna relación causal. Los tres supuestos garantizan que $\beta_1\, 1 + \beta_2 X$ es una buena *aproximación probabilística* de la variable Y en función de la variable X; no garantizan, en ningún caso, que esa aproximación sea el reflejo de una relación causa-efecto.

La econometría proporciona el método para *cuantificar* la magnitud de un efecto, pero dicho método solo es útil si el efecto corresponde a una relación teóricamente válida. La justificación sobre la dirección causal del efecto solo puede provenir de un modelo teórico previo.

Para profundizar:

- Angrist, J. y Pischke, J.-S. (2014), *Mastering 'Metrics'*, capítulo 1: discusión accesible sobre la distinción entre asociación estadística y causalidad, con ejemplos económicos.
- Bujosa, M., sección "Regresión como descomposición ortogonal": discusión, con los mismos ejemplos, de por qué la descomposición ortogonal por sí sola no valida ninguna teoría causal.

8 El salto: método de los momentos

En el modelo poblacional, los verdaderos parámetros β_2 y β_1 dependen de la geometría de la *población* (del producto escalar probabilístico $\langle \cdot, \cdot \rangle_P$):

$$\beta_2 \;=\; \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \qquad\qquad \beta_1 \;=\; \mathrm{E}(Y) - \beta_2 \mathrm{E}(X).$$

Problema: Eros no conoce la población entera; aunque conoce algunos datos.

Solución (El Método de los Momentos): Sustituir los momentos poblacionales desconocidos ($\mathrm{E}(\cdot)$) por sus análogos muestrales ($\frac{1}{n}\sum$, es decir, usando el producto escalar estadístico $\langle \cdot, \cdot \rangle_s$ sobre los datos).

$$\hat{\beta}_2 \;=\; \frac{\sigma_{\boldsymbol{xy}}}{\sigma^2_{\boldsymbol{x}}}, \qquad\qquad \hat{\beta}_1 \;=\; \mu_{\boldsymbol{y}} - \hat{\beta}_2 \mu_{\boldsymbol{x}}.$$

¡Son las mismas fórmulas que derivamos en la lección 7! Mínimos Cuadrados Ordinarios (MCO) equivale al Método de los Momentos en el Modelo Clásico de Regresión Lineal.

El salto: método de los momentos

Llegamos al momento en que todo el andamiaje construido hoy se pone al servicio de algo muy concreto: entender, con una mirada nueva, las fórmulas que ya calculamos en la lección 7.

Bajo el Modelo Clásico de Regresión, hemos asumido que la "sombra" cae en el plano: $E[Y|X] = \beta_1\, \mathit{1} + \beta_2 X$. Partamos de las dos condiciones poblacionales que obtuvimos como corolario en la transparencia del MCRL, a partir de $E[U \mid X] = \mathit{0}$: $\mathrm{E}(U) = 0$ y $\mathrm{Cov}(U, X) = 0$. Sustituyamos en ellas $U = Y - \beta_1\, \mathit{1} - \beta_2 X$ (el primer supuesto del modelo, despejando U) y resolvamos el sistema resultante[8] para β_1, β_2.

De $\mathrm{Cov}(U, X) = 0$, usando la linealidad de la covarianza y que $\mathrm{Cov}(\beta_1\, \mathit{1}, X) = 0$ (una variable aleatoria constante no covaría con nada):

$$\mathrm{Cov}(U, X) = \mathrm{Cov}(Y - \beta_1\, \mathit{1} - \beta_2 X,\, X) = \mathrm{Cov}(Y, X) - \beta_2 \mathrm{Var}(X) = 0,$$

de donde

$$\beta_2 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

Y de $\mathrm{E}(U) = 0$:

$$\mathrm{E}(U) = \mathrm{E}(Y) - \beta_1 - \beta_2 \mathrm{E}(X) = 0 \quad \Longrightarrow \quad \beta_1 = \mathrm{E}(Y) - \beta_2 \mathrm{E}(X).$$

Estas dos fórmulas son, letra por letra, el análogo poblacional exacto de las que obtuvimos en la lección 7: $\hat{\beta}_2 = \sigma_{\boldsymbol{xy}}/\sigma_{\boldsymbol{x}}^2$ y $\hat{\beta}_1 = \mu_{\boldsymbol{y}} - \hat{\beta}_2 \mu_{\boldsymbol{x}}$, con $\mathrm{E}(\cdot)$, $\mathrm{Cov}(\cdot, \cdot)$ y $\mathrm{Var}(\cdot)$ en el lugar de $\mu_\cdot$, $\sigma_{\cdot\cdot}$ y $\sigma_\cdot^2$.

Aquí topamos, sin embargo, con un problema real, y precisamente el que da nombre a esta transparencia. Las cantidades $\mathrm{Cov}(X, Y)$, $\mathrm{Var}(X)$, $\mathrm{E}(Y)$ y $\mathrm{E}(X)$ son *momentos poblacionales*: describen a las variables aleatorias X e Y en su totalidad, no a ninguna muestra concreta. Y, en general, *no los conocemos*; citando a Platón, no somos dioses ni sabios: no conocemos ni X ni Y ni sus valores esperados, no conocemos la covarianza poblacional $\mathrm{Cov}(X, Y)$ ni la varianza poblacional $\mathrm{Var}(X)$. Pero tampoco somos completos ignorantes: tenemos una muestra de tamaño n.

Así pues, esta es la situación filosófica que anunciamos al abrir la sesión con la cita de Diotima. No poseemos β_1, β_2 (la "verdad" poblacional, que sería el conocimiento del sabio); pero tampoco carecemos de todo (no somos el ignorante que no sabe ni que le falta algo). Tenemos una muestra, un vector concreto, con el que podemos intentar *aproximarnos* a esos parámetros desconocidos.

El "salto" que da nombre a esta transparencia consiste en resolver este problema del modo más directo posible: *sustituir cada momento poblacional (desconocido) por su análogo muestral (calculable)*. Donde aparecía $\mathrm{E}(X)$ (desconocido), escribimos $\mu_{\boldsymbol{x}}$ (calculable a partir de la muestra); donde aparecía $\mathrm{Cov}(X, Y)$, escribimos $\sigma_{\boldsymbol{xy}}$; donde aparecía $\mathrm{Var}(X)$, escribimos $\sigma_{\boldsymbol{x}}^2$; y donde aparecía $\mathrm{E}(Y)$, escribimos $\mu_{\boldsymbol{y}}$. El resultado de esta sustitución, aplicada a las dos fórmulas de más arriba, es:

$$\hat{\beta}_2 = \frac{\sigma_{\boldsymbol{xy}}}{\sigma_{\boldsymbol{x}}^2}, \qquad\qquad \hat{\beta}_1 = \mu_{\boldsymbol{y}} - \hat{\beta}_2 \mu_{\boldsymbol{x}}.$$

[8] Es *exactamente* la misma álgebra que hicimos en la lección 7 para obtener $\hat{\beta}_1, \hat{\beta}_2$ a partir de las dos condiciones de ortogonalidad muestrales, solo que ahora con esperanzas en lugar de medias muestrales.

¡Pero estas son, exactamente, las fórmulas que ya obtuvimos en la lección 7! No como consecuencia de ninguna casualidad, sino porque la ruta que hemos seguido —resolver las condiciones poblacionales y sustituir momentos poblacionales por muestrales— reproduce, paso a paso, la misma estructura algebraica que allí seguimos —resolver las dos condiciones de ortogonalidad muestral directamente sobre $\widehat{\boldsymbol{e}}$—. Esta estrategia general, de sustituir momentos teóricos desconocidos por sus análogos muestrales para obtener un estimador, se llama *método de los momentos*, y es, en este caso concreto, exactamente equivalente a la estimación por mínimos cuadrados ordinarios.

Lo que ha cambiado, entre la lección 7 y hoy, no es la fórmula —es idéntica—, sino su *interpretación*. En la lección 7, $\hat{\beta}_2$ era, sencillamente, la pendiente del ajuste que mejor se acomodaba a un vector de datos concreto: un hecho geométrico sobre ese vector, sin ninguna pretensión más allá de él. Hoy, con el modelo poblacional $Y = \beta_1\, 1 + \beta_2 X + U$ ya construido, $\hat{\beta}_2$ pasa a ser algo distinto: un *estimador* del parámetro poblacional (desconocido) β_2 —una aproximación, calculada a partir de nuestra muestra, a una cantidad que presumimos que existe en el mundo económico subyacente pero que nunca observaremos directamente—.

Un ejemplo concreto, para no perder el contacto con el laboratorio de regresión simple con datos reales. Escribamos el modelo poblacional $Price = \beta_1\, 1 + \beta_2 Rooms + U$, con la interpretación —bajo el primer supuesto, y con las reservas causales de la transparencia anterior— de que β_2 mide cómo varía, en promedio, el precio de una vivienda cualquiera ante un incremento de una habitación; no solo de las 506 del Boston metropolitano que observamos. Entonces el coeficiente $\hat{\beta}_2$ que Gretl calculó en aquella práctica no es solo "la pendiente de la recta que mejor se ajusta a esos 506 puntos". Es, además, nuestra mejor *estimación*, con los datos disponibles, de ese parámetro poblacional desconocido.

Para profundizar:

- Bujosa, M., sección "Estimación del Modelo Clásico de Regresión Lineal": discusión extensa (con el ejemplo detallado del precio de la vivienda) de por qué sustituir momentos poblacionales por muestrales constituye "un salto audaz", y de la relación exacta entre $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_s)$ y un espacio de probabilidad.
- Wooldridge, J. M. (2020), Apéndice C.1: introducción al método de los momentos como principio general de estimación.

9 Recapitulación y guiño a la lección siguiente

$\mathbb{R}^n$ (dato conocido)	Variables aleatorias (a inferir)
vector de datos $\boldsymbol{y}$	variable aleatoria Y
$\mu_{\boldsymbol{y}} = \langle \boldsymbol{y}, \mathbf{1}\rangle_s$	$\mathrm{E}(Y) = \langle Y, 1\rangle_P$
$\overline{\boldsymbol{y}}$ (vector constante)	$E[Y \mid 1]$ (función constante)
$\widehat{\boldsymbol{y}}$ (proyección sobre $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$)	$E[Y \mid X]$ (proyección sobre las func. de X y 1)
$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$	$U = Y - E[Y \mid X]$
$\sigma^2_{\boldsymbol{y}} = \|\boldsymbol{y} - \mu_{\boldsymbol{y}}\mathbf{1}\|^2_s$	$\mathrm{Var}(Y) = \|Y - \mathrm{E}(Y)\, 1\|^2_P$
$\sigma^2_{\boldsymbol{y}} = \sigma^2_{\widehat{\boldsymbol{y}}} + \sigma^2_{\widehat{\boldsymbol{e}}}$	$\mathrm{Var}(Y) = \mathrm{Var}(E[Y \mid X]) + \mathrm{Var}(U)$
$\hat{\beta}_1, \hat{\beta}_2$ (ajuste de una muestra)	β_1, β_2 (parámetros poblacionales, a estimar)

Próximos pasos:

- Hasta ahora la variable X era un único regresor (Regresión Simple).
- En economía real intervienen múltiples factores a la vez $(X_1, X_2, \ldots, X_k)$.
- *Lección 10 (Regresión Múltiple)*: Ampliaremos la proyección del plano 2D a subespacios de mayor dimensión. Para manejar tantas condiciones de ortogonalidad, introduciremos una notación muy elegante: las *matrices*.

Recapitulación y guiño a la lección siguiente

Cerramos esta lección con la tabla-diccionario que resume, de un vistazo, todo el camino recorrido. Conviene leerla despacio, fila por fila, porque cada línea condensa una transparencia entera de la sesión.

La primera fila recoge la idea de partida (transparencia 2): tanto el vector de datos $\boldsymbol{y}$ como la variable aleatoria Y son, en el fondo, funciones —solo que definidas sobre dominios de naturaleza radicalmente distinta: uno finito y perfectamente conocido ($\{1, \ldots, n\}$), el otro potencialmente inmenso y, en la práctica, nunca descrito en detalle (Ω).

La segunda y tercera filas recogen la distinción, mantenida durante toda la sesión, entre el valor esperado como número y la esperanza condicional como función constante (transparencia 4). E(Y) es un escalar, como lo era $\mu_{\boldsymbol{y}}$. $E[Y|\mathit{1}]$ es una función constante —la variable aleatoria constante más próxima a Y—, como lo era $\overline{\boldsymbol{y}}$, el vector constante más próximo a $\boldsymbol{y}$.

La cuarta fila recoge, quizá, la diferencia más profunda de toda la sesión (transparencia 5). $\widehat{\boldsymbol{y}}$ era la proyección sobre un plano de solo dos direcciones linealmente independientes; $E[Y|X]$ es la proyección sobre un espacio de funciones que, en general, tiene infinitas. Por eso la linealidad de $E[Y|X]$ es un supuesto sustantivo, y no una consecuencia automática de la definición.

La quinta fila es la descomposición fundamental que ambos mundos comparten: en $\mathbb{R}^n$, $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}$; con variables aleatorias, $Y = E[Y|X] + U$. Ambas son descomposiciones ortogonales que existen *siempre*, sin necesidad de ningún supuesto adicional — y, por eso mismo, ninguna de las dos implica, por sí sola, ninguna relación de causalidad (transparencia 7).

La sexta fila recuerda que la varianza, en ambos mundos, mide lo mismo: la dispersión como el cuadrado del tamaño de la componente no constante.

La séptima fila recuerda que la identidad pitagórica permite descomponer la varianza (cuadrado de la hipotenusa $Y - E[Y|\mathit{1}]$), como suma de varianzas (cuadrados de los catetos horizontal: $E[Y|X] - E[Y|\mathit{1}]$ y vertical: U) —véase la figura de la transparencia 5.

Y la octava, finalmente, recoge la diferencia epistemológica más importante de todas, la que hemos intentado transmitir con la imagen de Diotima. $\hat{\beta}_1, \hat{\beta}_2$ son cantidades *exactas*, conocidas sin margen de error para un vector de datos concreto, el de nuestra muestra incompleta (lección 7). β_1, β_2 son cantidades características de la *población* completa, desconocidas para nosotros, que solo podemos tratar de *estimar* —nunca conocer con certeza— a partir de una muestra incompleta. El salto de la última transparencia consistió en mostrar que ambas cantidades, la exacta y la inferida, comparten la misma fórmula.

El mensaje que debe quedar de esta sesión, por encima de cualquier fórmula concreta, es el siguiente: el lenguaje geométrico que hemos construido con tanto cuidado durante las lecciones 1

a 8 —vectores, producto escalar, proyección ortogonal, ángulos— no era exclusivo de $\mathbb{R}^n$. Es, más bien, un lenguaje general para la *aproximación óptima*, que se puede trasladar —con las honestas advertencias que hemos ido señalando en cada paso— a cualquier conjunto de objetos donde exista una noción razonable de producto escalar. Las variables aleatorias, con la esperanza del producto como su propio producto escalar, son uno de esos conjuntos; y por eso la esperanza condicional $E[Y|X]$ hereda, con toda naturalidad, el papel que en $\mathbb{R}^n$ jugaba la proyección $\widehat{\boldsymbol{y}}$.

Quedan, sin embargo, varias tareas pendientes que las próximas lecciones deberán resolver. La primera, inmediata: hasta ahora hemos trabajado con un único regresor X (si la Renta influye en el Consumo, también lo hacen los tipos de interés, las expectativas de inflación, la edad de los consumidores o su número de hijos). En la *Lección 10 (Regresión Múltiple)* proyectaremos sobre subespacios generados por k regresores a la vez. La lógica geométrica será *la misma* (una condición de ortogonalidad por cada regresor), pero escribir un sistema de k ecuaciones con sus sumatorios correspondientes resultaría extenuante. Para comprimir esa información y hacerla manejable introduciremos el uso de *matrices* ($\mathbf{X}$). Las matrices, como verá, no son más que un truco de notación brillante para expresar múltiples productos escalares de una sola vez.

Para profundizar:

- Wooldridge, J. M. (2020), capítulo 2, sección 2.6: resumen de los supuestos del modelo de regresión simple.