Índice general

Lección 11. ¿Es $\widehat{\boldsymbol{\beta}}$ un buen estimador? Insesgadez y varianza

Marcos Bujosa

19 de septiembre de 2026

Resumen

¿Es $\widehat{\boldsymbol{\beta}}$ un buen estimador de $\boldsymbol{\beta}$? Para el modelo lineal simple: $\hat{\beta}_2$ es insesgado y calculamos su varianza, *reutilizando* la fórmula de la lección 7 sobre variables aleatorias.

"¿Es bueno un estimador? ¿Está bien centrado? ¿Cuánto se dispersa?"

- (slides) — (html) — (pdf)

1 De dónde venimos: una pregunta pendiente

Recordemos cómo cerraba la lección 10: *"¿es $\widehat{\boldsymbol{\beta}}$ un buen estimador del verdadero $\boldsymbol{\beta}$?"*

$\hat{\beta}_1, \hat{\beta}_2$ —calculados primero con argumentos geométricos sobre los vectores de datos— se reinterpretaron después como *estimadores* de β_1, β_2 por el método de los momentos.

Hoy, para el modelo lineal simple —constante *1* más un regresor *no constante* X—, demostraremos dos propiedades de $\hat{\beta}_2$:

1. *Insesgadez*: $\mathrm{E}(\hat{\beta}_2) = \beta_2$.
2. *Varianza*: una fórmula exacta de $Var[\hat{\beta}_2 \mid \boldsymbol{X}]$.

Nos centramos en $\hat{\beta}_2$ (la pendiente, de mayor interés económico); los mismos argumentos, sobre la fórmula $\hat{\beta}_1$, permiten probar la insesgadez de $\hat{\beta}_1$ —no las desarrollamos aquí para no ser repetitivos.

Ningún ingrediente geométrico nuevo: la fórmula de $\hat{\beta}_2$ que obtuvimos por geometría, la reutilizamos hoy tal cual —evaluada sobre variables aleatorias en lugar de sobre datos: *una fórmula aritmética vieja, aplicada sobre un objeto nuevo.*

De dónde venimos: una pregunta pendiente

Esta sesión cierra un hilo que llevamos arrastrando desde hace varias lecciones. En la lección 9 ("El puente") dimos sentido, por primera vez, a la pregunta "¿$\hat{\beta}_2$ estima bien al verdadero β_2 poblacional?", al construir el modelo poblacional $Y = \beta_1\, 1 + \beta_2 X + U$ y reinterpretar las fórmulas de la lección 7 como estimadores por el método de los momentos. En la lección 10 (regresión múltiple) esa misma pregunta se generalizó a k regresores, pero se dejó explícitamente sin respuesta

—una sola frase en la recapitulación, sin fórmula ni supuestos generalizados—, precisamente para retomarla hoy con calma, en el caso más sencillo posible: $k = 2$, es decir, *dos* regresores en total —la constante *1* y un único regresor no constante X.

La estrategia de esta sesión no introduce ninguna geometría nueva. Vamos a reutilizar, tal cual, la fórmula de $\hat{\beta}_2$ que obtuvimos en la lección 7 mediante la "vía alternativa" (proyección sobre generadores ortogonales, vía Cauchy–Schwarz). Esa fórmula, una vez concluido el cálculo geométrico, quedó escrita como una expresión puramente aritmética: una suma de productos dividida por una suma de cuadrados. Precisamente por ser aritmética —y ya no geométrica—, podemos evaluarla en cualquier conjunto de $2n$ números que le demos como argumento. Hoy le daremos, en lugar de datos fijos $x_1, \ldots, x_n, y_1, \ldots, y_n$, las variables aleatorias de una muestra $X_1, \ldots, X_n, Y_1, \ldots, Y_n$. El resultado de esa sustitución ya no es un número: es una nueva variable aleatoria, el *estimador* $\hat{\beta}_2(\boldsymbol{X}, \boldsymbol{Y})$.

Conviene ser honestos, desde el principio, sobre qué NO vamos a hacer. No vamos a definir un nuevo producto escalar $\langle \cdot, \cdot \rangle$ sobre "vectores cuyas componentes son variables aleatorias", ni vamos a apelar de nuevo a Cauchy–Schwarz para justificar la fórmula de $\hat{\beta}_2$: esa demostración geométrica ya se hizo, una vez, en la lección 7, sobre datos fijos, y no hay que repetirla ni reinventarla sobre un objeto distinto. Lo único que hacemos hoy es sustituir números por variables aleatorias dentro de una fórmula ya obtenida —una operación mucho más modesta que una nueva demostración geométrica, y que no exige comprobar ninguna propiedad de producto escalar—. Es la diferencia entre demostrar un teorema y usar la fórmula que el teorema produjo: hoy solo hacemos lo segundo.

Conviene también ser explícitos, desde el principio, sobre el alcance de esta sesión: demostramos insesgadez y la expresión de la varianza *solo* para $\hat{\beta}_2$ en la regresión con dos regresores (el modelo lineal simple). La generalización a k regresores existe (y la citaremos, sin demostrarla, al final de la sesión), pero requiere álgebra matricial que para este curso he preferido no desarrollar.[1] Lo que sí haremos, como cierre, es comprobar que la fórmula general, evaluada en el caso particular $k = 2$, coincide con lo que aquí probamos sin matrices.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, secciones 2.4-2.5: insesgadez y varianza de los estimadores MCO en regresión simple (Teoremas 2.1 y 2.2).

2 El marco: muestra aleatoria

Recordemos (lección 9): los n datos son n copias idénticas e independientes de Y; hoy, del par (X, U).

$\boldsymbol{X}$: la *versión aleatoria* de $\mathbf{X} = [\mathbf{1};\ \boldsymbol{x}]$, con columnas *1* ($n$ copias) y $X_1, \ldots, X_n$. Análogamente $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ y $\boldsymbol{U} = (U_1, \ldots, U_n)$. Para cada i: $Y_i = \beta_1\, 1 + \beta_2 X_i + U_i$, con los tres supuestos de la lección 9, condicionando ahora en $\boldsymbol{X}$.

Lo nuevo es el **muestreo aleatorio**: las n copias (X_i, U_i) son independientes entre sí. Dos consecuencias:

- $E[U_i \mid \boldsymbol{X}] = 0$: $U_i \perp$ funciones de X_i (*por definición*) y de las demás copias (*por independencia*).

[1] para no saturar al estudiante.

- Perturbaciones de observaciones distintas, incorreladas: $Cov[U_i, U_j \mid \boldsymbol{X}] = 0 \;\; (i \neq j)$.

  Condicionar en $\boldsymbol{X}$: la esperanza condicional de la lección 9, sobre las n copias a la vez.

El marco: muestra aleatoria

La lección 9 introdujo, de pasada, una frase que hoy se convierte en el ingrediente central de la sesión: "asumiremos que nuestros n datos son el resultado de observar la realización de n copias idénticas e independientes de esa variable Y". Allí esa frase servía únicamente para justificar por qué tenía sentido tratar cada dato y_i como la realización de una variable aleatoria. Hoy la necesitamos con mucho más detalle, porque para hablar de la *distribución* de $\hat{\beta}_2$ —de cuánto varía de una muestra a otra— necesitamos imaginar que la muestra entera podría haber sido distinta.

Formalicemos esa idea. Desde la lección 4 venimos usando $\boldsymbol{x} = (x_1, \ldots, x_n)$ para el vector de datos observados (y, desde la lección 6, para el vector de datos observados de un *regresor*), y en la lección 10 agrupamos los regresores observados en la matriz de datos $\mathbf{X}$ (en el modelo lineal simple, $\mathbf{X} = [\mathbf{1};\ \boldsymbol{x}]$, dos columnas). Hoy introducimos su análogo aleatorio: la matriz[2] $\boldsymbol{X}$. Este es el objeto sobre el que condicionaremos todos los momentos a lo largo de la sesión. Cada columna de $\boldsymbol{X}$ corresponderá al muestreo de uno de los regresores poblacionales. En el modelo lineal simple son $\mathit{1}$ y X; por tanto, la segunda columna de $\boldsymbol{X}$ está formada por las n variables aleatorias $X_1, \ldots, X_n$: cada una copia idéntica e independiente de la variable poblacional X de la lección 9.[3] Es la misma relación dato/variable aleatoria que venimos usando desde la lección 9 (entre y_i e Y), aplicada componente a componente a los n elementos de la muestra correspondiente a cada columna.

Una precisión de uso. Durante casi toda la sesión no necesitaremos ninguna operación matricial con $\boldsymbol{X}$: bastarán sumatorios sobre las X_i, y "condicionar en $\boldsymbol{X}$" significará siempre proyectar sobre el espacio de funciones cuyo argumento son, simultáneamente, todas las variables aleatorias presentes en $\boldsymbol{X}$. La estructura de matriz de $\boldsymbol{X}$ reaparecerá puntualmente al justificar el divisor $n - k$ (sección "MCO es BLUE; estimando σ^2", donde hablaremos de las columnas de $\mathbf{X}$) y, con más de 2 columnas, en la generalización final a k regresores. Es decir: el símbolo $\boldsymbol{X}$ significa lo mismo al principio y al final de la sesión; lo único que cambia es cuántas columnas tiene.

Para cada observación i, el modelo poblacional de la lección 9 se aplica sin ningún cambio, con subíndice i añadido:

$$Y_i = \beta_1 \, \mathit{1} + \beta_2 X_i + U_i, \qquad i = 1, \ldots, n,$$

con los tres supuestos del Modelo Clásico de Regresión Lineal aplicados a cada par (X_i, U_i): linealidad, independencia lineal ($\mathrm{Var}(X) \neq 0$) y homocedasticidad $Var[U_i \mid \boldsymbol{X}] = \sigma^2 \, \mathit{1}$; más la exogeneidad estricta $E[U_i \mid \boldsymbol{X}] = \mathit{0}$, que en la lección 9 obteníamos gratis.[4]

[2] Denotada en negrita cursiva pero sin letra de palo seco (reservada para las matrices de datos observados).

[3] Para decir que son copias idénticas basta con que tengan la misma función de distribución: $F_{X_i}(x) = \mathrm{P}(X_i \leq x)$ para todo $i = 1, \ldots, n$. Esto no las obliga a ser la misma función sobre Ω; al contrario: dos copias independientes suelen asignar valores distintos para un mismo suceso elemental. Lo que sí comparten, por tener la misma distribución, es el conjunto de valores que toman con probabilidad positiva —su imagen, salvo en sucesos de probabilidad nula, donde ninguna de las dos está sujeta a restricción alguna—. Por otro lado, la independencia entre variables aleatorias implica que su función de distribución de probabilidad conjunta equivale al producto de sus distribuciones individuales. Para el caso de n variables, esto se expresa como: $F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$.

[4] Ese "gratis" merece hoy una letra pequeña que en la lección 9 no hacía falta. Allí, $U = Y - E[Y|X]$ era ortogonal *por construcción* a toda función de X, y de ahí salía $E[U \mid X] = \mathit{0}$ sin coste alguno. Hoy, sin embargo, no estamos condicionando en X_i, sino en *toda* la matriz $\boldsymbol{X}$, es decir, en las n copias a la vez. Y la definición de U_i solo garantiza,

Sobre el supuesto $\mathrm{Var}(X) \neq 0$ conviene una precisión, porque es fácil que el símbolo "X" despiste una vez que ya solo hablamos de $X_1, \ldots, X_n$. X, sin subíndice, sigue denotando —como en la lección 9— la variable poblacional de la que $X_1, \ldots, X_n$ son copias. El supuesto $\mathrm{Var}(X) \neq 0$ se enuncia sobre esa variable poblacional, y se aplica a cada copia X_i porque, al ser copias idénticas, cada X_i tiene la misma distribución que X (y, en particular, la misma varianza).

En resumen: la exogeneidad estricta sigue sin ser un supuesto, pero hoy descansa sobre dos cosas y no sobre una. La definición de U_i da $U_i \perp$ funciones de X_i; el muestreo aleatorio da $U_i \perp$ funciones de las demás copias. Es la primera consecuencia que le sacamos al muestreo aleatorio; la segunda viene a continuación. Nótese, además, que la homocedasticidad de cada U_i también se condiciona a *toda* la matriz $\boldsymbol{X}$, no solo a X_i: es más fuerte que condicionar solo al regresor de esa observación, y es lo que hace falta para todo lo que sigue; pero aquí no hay corolario que valga: la estamos *imponiendo* directamente en esa forma más fuerte.

La segunda consecuencia del muestreo aleatorio es aún más directa: si (X_i, U_i) y (X_j, U_j) son independientes para $i \neq j$, entonces, en particular, U_i y U_j son independientes, y dos variables independientes tienen covarianza cero:[5]

$$Cov[U_i, U_j \mid \boldsymbol{X}] = 0 \qquad (i \neq j).$$

No se trata, por tanto, de un supuesto añadido *ex novo*: es la traducción, en términos de covarianzas, de algo que ya habíamos dado por sentado desde la lección 9 al hablar de copias "idénticas e independientes" (lo que su curso de estadística del cuatrimestre anterior llamaba, seguramente, *muestreo aleatorio simple*). Hoy es la primera vez que necesitamos esta consecuencia de forma explícita, porque es la primera vez que trabajamos con *varias* perturbaciones $U_1, \ldots, U_n$ simultáneamente, en lugar de con una sola U genérica.

---

por sí sola, la ortogonalidad frente a las funciones de X_i: nada dice, de entrada, sobre las funciones de X_j con $j \neq i$. ¿Por qué vale entonces $E[U_i \mid \boldsymbol{X}] = 0$? Por el muestreo aleatorio, y se puede ver con la misma geometría de siempre. Lo que hay que comprobar es que U_i es ortogonal a toda función de la matriz $\boldsymbol{X}$. Empecemos por las funciones más sencillas: los productos $a(X_i)b(X_{-i})$, donde X_{-i} abrevia "las demás copias", X_j con $j \neq i$. Separemos $b(X_{-i})$ en su parte constante y su parte centrada, exactamente como hicimos con los vectores de datos en la lección 4, escribiendo $\mathrm{E}(b)$ por $\mathrm{E}\big(b(X_{-i})\big)$:

$$a(X_i)\, b(X_{-i}) \;=\; \underbrace{\mathrm{E}(b)\, a(X_i)}_{\text{función de } X_i} \;+\; a(X_i)\, \big(b(X_{-i}) - \mathrm{E}(b)\, \mathit{1}\big).$$

U_i es ortogonal al primer sumando por definición: es una función de X_i. Y es ortogonal al segundo por la independencia. En efecto, (X_i, U_i) es independiente de X_{-i}, y la independencia se traduce, en nuestro lenguaje, en que *la parte centrada de cualquier función de X_{-i} es ortogonal a cualquier función de (X_i, U_i)* —es el hecho, conocido del curso de probabilidad, de que variables independientes están incorreladas, aplicado aquí a $U_i\, a(X_i)$ y a $b(X_{-i})$—:

$$\big\langle U_i\, a(X_i),\; b(X_{-i}) - \mathrm{E}(b)\, \mathit{1} \big\rangle_P = \mathrm{Cov}\big(U_i\, a(X_i),\, b(X_{-i})\big) = 0.$$

Así que U_i es ortogonal a todos los productos $a(X_i)\, b(X_{-i})$ y, por linealidad del producto escalar, a sus combinaciones lineales. Lo único que admitimos sin demostración —es el paso que pertenece a un curso de probabilidad— es que con esas combinaciones se aproxima cualquier función de $\boldsymbol{X}$. Con ello, U_i es ortogonal a toda función de $\boldsymbol{X}$, que es lo que dice $E[U_i \mid \boldsymbol{X}] = 0$.

[5] *Notación.* Esta es la primera vez en el curso que aparece una covarianza *condicionada*. Sigue el mismo convenio que ya fijamos en la lección 9 para la esperanza y la varianza: $\mathrm{Cov}(\cdot,\cdot)$, con paréntesis y en recta, es un *número* (la covarianza no condicionada); $Cov[\cdot,\cdot \mid \cdot]$, con corchetes y en cursiva, es una *variable aleatoria* —una función de aquello sobre lo que se condiciona—. Su definición es la análoga de la varianza condicional: $Cov[Z, W \mid \boldsymbol{X}] = E\big[(Z - E[Z \mid \boldsymbol{X}])(W - E[W \mid \boldsymbol{X}]) \mid \boldsymbol{X}\big]$. La usaremos en serio al calcular la varianza de $\hat{\beta}_2$.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, sección 2.5, supuesto SLR.2 (muestreo aleatorio).

3 $\hat{\beta}_2$: la misma fórmula, aplicada a variables aleatorias

Estimación vs. estimador (distinción hoy imprescindible):

- $\hat{\beta}_2(\boldsymbol{x}, \boldsymbol{y})$, la *estimación*: un número.
- $\hat{\beta}_2(\boldsymbol{X}, \boldsymbol{Y})$, el *estimador*: una variable aleatoria, la MISMA fórmula sobre la muestra aleatoria.

Recordando la "vía alternativa" (lección 7):

$$\hat{\beta}_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i (x_i - \mu_{\boldsymbol{x}}) y_i}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2} = \sum_i w_i y_i; \quad \text{donde } w_i = \frac{(x_i - \mu_{\boldsymbol{x}})}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2}.$$

Sustituimos x_i, y_i por X_i, Y_i:

$$\hat{\beta}_2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_i (X_i - \overline{X}) Y_i}{\sum_j (X_j - \overline{X})^2} = \sum_i W_i Y_i; \quad \text{donde } W_i = \frac{(X_i - \overline{X})}{\sum_j (X_j - \overline{X})^2}.$$

donde $\overline{X} = \frac{1}{n} \sum_i X_i$ es la *media muestral*; y $\sum_i (X_i - \overline{X}) = 0$.

$\hat{\beta}_2$: la misma fórmula, aplicada a variables aleatorias

Antes de seguir, conviene fijar una distinción que hasta ahora el curso no había requerido. Cuando escribimos $\hat{\beta}_2 = \sigma_{\boldsymbol{xy}} / \sigma^2_{\boldsymbol{x}} = \frac{\sum_i (x_i - \mu_{\boldsymbol{x}}) y_i}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2}$ (lección 7) y evaluamos esa fórmula empleando dos vectores de datos concretos $\boldsymbol{x}$ e $\boldsymbol{y}$, obtenemos un *número* —la *estimación* $\hat{\beta}_2$ de β_2. Pero esa fórmula con sumatorios, aplicada a dos muestras $\boldsymbol{X}$ y $\boldsymbol{Y}$ resulta ser una *función de variables aleatorias*; es decir, ella misma es una variable aleatoria[6]: en tal caso es el *estimador* $\hat{\beta}_2$ de β_2.[7]

Detengámonos, con todo el cuidado que merece, en qué significa exactamente "sustituir datos por variables aleatorias en una fórmula". Una fórmula como $\mu_{\boldsymbol{x}} = f(x_1, \ldots, x_n) = \frac{1}{n} \sum_i x_i$ es, matemáticamente, una función que toma n números y devuelve un solo número: *la media*. Pero nada impide evaluar esa MISMA fórmula con otro tipo de argumento (si la operación sigue estando bien definida): si en lugar de n números le damos n variables aleatorias $X_1, \ldots, X_n$, el resultado $f(X_1, \ldots, X_n) = \frac{1}{n} \sum_i X_i$ ya no es un número, ahora es una variable aleatoria. A esta nueva variable aleatoria la llamamos *estimador de la media* o sencillamente *media muestral*, y la denotamos con $\overline{X}$. Su definición emplea la MISMA fórmula aritmética que en la lección 4 usamos para definir la media de un vector de datos, pero ahora la fórmula es aplicada a las n variables aleatorias de la muestra (de ahí el nombre de *media muestral*).

Dicha observación —sencilla pero sustancial— guiará el desarrollo de la presente sesión. Conviene enunciarla explícitamente para evitar atribuirle una complejidad que no tiene: nos limitamos

[6] Una función de $\omega \in \Omega$, donde Ω el espacio de sucesos.

[7] En la literatura estadística y econométrica es costumbre usar la misma notación para ambos objetos (la estimación y el estimador), pero recuerde que son muy distintos (un número el primero y una variable aleatoria el segundo).

a definir nuevas variables aleatorias mediante la aplicación de fórmulas preexistentes (desarrolladas en su momento en $\mathbb{R}^n$) a las variables aleatorias de una muestra. Dichas variables aleatorias se conocen como *estadísticos* o *estadísticos muestrales*.

Con esto aclarado, retomamos la fórmula de la "vía alternativa" de la lección 7. Allí, cambiando a los generadores ortogonales $\mathbf{1}$ y $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$, obtuvimos, para dos vectores de datos $\boldsymbol{x}$ e $\boldsymbol{y}$,

$$\hat{\beta}_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{\langle \boldsymbol{y}, (\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1})\rangle_s}{\langle (\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}), (\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1})\rangle_s} = \frac{\sum_i (x_i - \mu_{\boldsymbol{x}}) y_i}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2}.$$

El lado izquierdo de esta igualdad ($\langle \cdot, \cdot \rangle_s$) es geometría: un cociente de dos productos escalares, con todo su significado y papel en la proyección ortogonal. El lado derecho, en cambio, ya no necesita esa maquinaria para ser evaluado: es una simple expresión aritmética (los factores $1/n$ de cada producto escalar se cancelan al dividir, así que ni siquiera hace falta llevar la cuenta del divisor). Hoy nos quedamos con esta expresión aritmética —prescindiendo de la geometría subyacente— y sustituimos directamente las observaciones x_i, y_i por las variables aleatorias X_i, Y_i; y reescribimos la expresión como una suma ponderada:[8]

$$\hat{\beta}_2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_i (X_i - \overline{X}) Y_i}{\sum_j (X_j - \overline{X})^2} = \sum_{i=1}^{n} \underbrace{\left(\frac{(X_i - \overline{X})}{\sum_j (X_j - \overline{X})^2} \right)}_{W_i} Y_i = \sum_{i=1}^{n} W_i Y_i.$$

Escribimos los pesos aleatorios W_i con mayúscula por ser estadísticos (i.e., variables aleatorias).

Respecto a la media muestral, fijémonos que como $\overline{X} = \frac{1}{n}\sum_i X_i$, entonces $\sum_i X_i = n\overline{X}$ y que $\sum_{i=1}^n \overline{X} = \underbrace{\overline{X} + \cdots + \overline{X}}_{n \text{ veces}} = n\overline{X}$; así

$$\sum_i (X_i - \overline{X}) = \sum_i X_i - \sum_i \overline{X} = n\overline{X} - n\overline{X} = 0.$$

Usaremos esta propiedad de la *media muestral* en la siguiente transparencia.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, sección 2.4 (donde $\hat{\beta}_2$ se escribe, con la misma idea, como $\sum w_i y_i$ mediante sumatorios).

4 Dos propiedades de los pesos

Dos identidades algebraicas sobre los pesos. Ninguna es nueva.

Ambas valen para *cualquier* lista de números y, por tanto, siguen siendo ciertas al sustituir x_i por X_i: son hechos aritméticos, no probabilísticos:

- la primera es la condición $\sum_i (X_i - \overline{X}) = 0$ de la transparencia anterior (la misma cuenta que en la lección 4 daba media cero al vector en desviaciones);

[8] El nombre correcto de la operación debería ser *forma bilineal definida sobre el anillo de las variables aleatorias*. Pero ese nombre aquí no ayuda; dado el nivel matemático del curso. Tan solo advertir que el nombre de *suma ponderada* es abusivo, pues en la expresión $\sum_{i=1}^n W_i Y_i$ no hay "números" que hagan de ponderaciones en la suma de las Y_i; lo que aparece son las variables aleatorias W_i, que son funciones (no números). Por el mismo motivo, tampoco es correcto el nombre de *combinación lineal* de las Y_i.

$$\sum_i W_i = 0;$$

- la segunda es la razón por la que, en la "vía alternativa" de la lección 7, el coeficiente de $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$ resultaba ser exactamente $\hat{\beta}_2$.

$$\sum_i W_i X_i = 1 \qquad \Big(\text{pues } \frac{\sum_i (X_i - \overline{X})X_i}{\sum_j (X_j - \overline{X})^2} = \frac{\sum_i (X_i - \overline{X})^2}{\sum_j (X_j - \overline{X})^2}\Big).$$

Dos propiedades de los pesos

Antes de sustituir el modelo poblacional en $\hat{\beta}_2 = \sum_i W_i Y_i$, necesitamos dos propiedades de los pesos W_i que, en realidad, ya conocemos aunque nunca las hayamos escrito con este nombre. La primera es, literalmente, $\langle \boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}, \mathbf{1}\rangle_s = 0$ (lección 4: el vector en desviaciones tiene media cero), dividida por el denominador común de los W_i. La segunda es lo que, en la "vía alternativa" de la lección 7, hacía que al proyectar $\boldsymbol{y}$ sobre $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$ el coeficiente obtenido fuera la pendiente $\hat{\beta}_2$ (allí no aislamos unos "pesos"; nos limitamos a escribir la fórmula como cociente de productos escalares).

Conviene subrayar, antes de nada, un punto que puede pasar desapercibido y que es clave para no reintroducir la confusión entre estimador y estimación de la transparencia anterior: estas dos propiedades NO son afirmaciones probabilísticas. Son identidades *algebraicas*, del mismo tipo que "$2 + 3 = 5$": se cumplen para cualquier elección de n números $x_1, \ldots, x_n$ —iguales o distintos entre sí, positivos o negativos, no importa—, porque se deducen únicamente de la propia definición de $\overline{X}$ y W_i, nunca de ninguna propiedad estadística de esos números. Por eso, cuando sustituimos x_i por la variable aleatoria X_i, la identidad sigue siendo cierta sin necesidad de ningún argumento adicional (salvo, quizá, en el caso extremo de que todos los X_i coincidieran, en cuyo caso $\sum_j (X_j - \overline{X})^2 = 0$ y W_i ni siquiera estaría definido).[9]

Primera propiedad: $\sum_i W_i = 0$. Como $W_i = \frac{(X_i - \overline{X})}{\sum_j (X_j - \overline{X})^2}$ y el denominador es común a todos los términos, $\sum_i W_i = \frac{\sum_i (X_i - \overline{X})}{\sum_j (X_j - \overline{X})^2}$; y donde el numerador sabemos que es nulo: $\sum_i (X_i - \overline{X}) = 0$.

Segunda propiedad: $\sum_i W_i X_i = 1$. Usando de nuevo que $\sum_i (X_i - \overline{X}) = 0$

$$\sum_i (X_i - \overline{X})X_i = \sum_i (X_i - \overline{X})(X_i - \overline{X} + \overline{X}) = \sum_i (X_i - \overline{X})(X_i - \overline{X}) + \overline{X} \underbrace{\sum_i (X_i - \overline{X})}_{=0};$$

y donde hemos sacado factor común $\overline{X}$ en el sumatorio $\sum_i (X_i - \overline{X})\overline{X}$ para comprobar que es cero.

Por tanto $\sum_i W_i X_i = \frac{\sum_i (X_i - \overline{X})X_i}{\sum_j (X_j - \overline{X})^2} = \frac{\sum_i (X_i - \overline{X})^2}{\sum_j (X_j - \overline{X})^2} = 1$.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, sección 2.4 (donde estas dos propiedades se verifican mediante sumatorios, sin distinguir explícitamente su estatus puramente algebraico).

[9] Merece una precisión, porque es fácil pasarse de frenada: el supuesto $\text{Var}(X) \neq 0$ *no* basta para excluir ese caso extremo con probabilidad uno. Si X es continua, sí: el suceso "las n copias toman el mismo valor" tiene probabilidad nula. Pero si X es discreta no: con X de Bernoulli de parámetro $1/2$, por ejemplo, $\text{Var}(X) = 1/4 \neq 0$ y, sin embargo, la probabilidad de que las n copias coincidan vale $2^{1-n} > 0$. Lo correcto, en general, es decir que $\text{Var}(X) \neq 0$ garantiza que esa probabilidad es *menor que uno* y que tiende a cero al crecer n; y que todo lo que sigue se entiende condicionado a que la muestra observada no sea una de esas degeneradas —igual que un programa estadístico se niega a estimar la pendiente cuando el regresor no varía en la muestra—.

5 Sustituyendo el modelo poblacional

$$\hat{\beta}_2(\boldsymbol{X},\boldsymbol{Y}) = \sum_i W_i Y_i = \sum_i W_i(\beta_1\, 1 + \beta_2 X_i + U_i)$$

$$= \beta_1 \underbrace{\sum_i W_i}_{=0} + \beta_2 \underbrace{\sum_i W_i X_i}_{=1} + \sum_i W_i U_i$$

$$\boxed{\hat{\beta}_2(\boldsymbol{X},\boldsymbol{U}) = \beta_2\, 1 + \sum_{i=1}^{n} W_i U_i}$$

(escrito ahora en función de $\boldsymbol{U}$, pues $\boldsymbol{Y}$ ha sido sustituida).

Mediante manipulación algebraica (sustitución y las dos identidades de la transparencia anterior): *el estimador es el parámetro verdadero MÁS una "suma ponderada" de las perturbaciones*. Todo lo que sigue sale de estudiar ese segundo término.

Sustituyendo el modelo poblacional

Con las dos propiedades de la transparencia anterior —propiedades puramente algebraicas, válidas para cualquier valor que tomen $X_1, \ldots, X_n$—, el cálculo es puramente mecánico. Partimos de $\hat{\beta}_2 = \sum_i W_i Y_i$ (sección "$\hat{\beta}_2$: la misma fórmula, aplicada a variables aleatorias") y sustituimos el modelo poblacional $Y_i = \beta_1\, 1 + \beta_2 X_i + U_i$ (sección "El marco: muestra aleatoria") en cada sumando:

$$\hat{\beta}_2(\boldsymbol{X},\boldsymbol{U}) = \sum_i W_i(\beta_1\, 1 + \beta_2 X_i + U_i) = \beta_1 \sum_i W_i + \beta_2 \sum_i W_i X_i + \sum_i W_i U_i.$$

El primer término se anula ($\sum_i W_i = 0$); el segundo se reduce exactamente a $\beta_2\, 1$ ($\sum_i W_i X_i = 1$):

$$\hat{\beta}_2(\boldsymbol{X},\boldsymbol{U}) = \beta_2\, 1 + \sum_{i=1}^{n} W_i U_i.$$

Esta identidad es el corazón algebraico de toda la sesión, y merece leerse con detenimiento: dice que el estimador $\hat{\beta}_2$ es la variable aleatoria constante $\beta_2\, 1$ (donde β_2 es el verdadero parámetro poblacional) *más* una "suma ponderada" de las perturbaciones $U_1, \ldots, U_n$ —con los pesos W_i que ya conocemos—. Todo lo que necesitamos demostrar sobre $\hat{\beta}_2$ (que su valor esperado sea β_2, y que su dispersión sea tal o cual) se reduce, a partir de aquí, a estudiar el comportamiento de esa suma ponderada $\sum_i W_i U_i$. Y como los W_i, según acabamos de comprobar, son —al margen de toda probabilidad— sencillamente funciones de $X_1, \ldots, X_n$, al condicionar en $\boldsymbol{X}$ podremos *sacarlos fuera* de la esperanza y de la varianza condicionales. Conviene ser preciso sobre qué autoriza ese paso, porque es tentador decir que "condicionar en $\boldsymbol{X}$ convierte los W_i en números": no es así. Los W_i siguen siendo variables aleatorias —funciones de ω— antes y después de condicionar; lo que ocurre es que son funciones *de aquello sobre lo que se condiciona*, y el Lema de la sección siguiente demuestra que tales factores salen fuera del operador $E[\cdot \mid \boldsymbol{X}]$. El resultado *se parece* al de una combinación lineal con coeficientes fijos, pero la justificación no es esa: es el Lema, que a su vez sale de la caracterización geométrica de la esperanza condicional (lección 9).

Merece la pena notar, de pasada, la coherencia con el ejemplo numérico de la lección 7: allí, con $\boldsymbol{u} = (1, -2, 0, 2, -1)$ diseñado deliberadamente con $\sum w_i u_i = 0$ (covarianza muestral nula con el regresor), obtuvimos la *estimación* $\hat{\beta}_2 = \beta_2$ exactamente. La fórmula de hoy explica *por qué* ese resultado era una casualidad de diseño y no lo esperable en general: $\hat{\beta}_2 = \beta_2\, \mathit{1} + \sum_i W_i U_i$, y solo cuando esa suma se anula por construcción (como en aquel ejemplo, ya con datos y no con variables aleatorias) coinciden estimador y parámetro. En una muestra genuinamente aleatoria, $\sum_i W_i U_i$ no será cero, sino una variable aleatoria con su propia media y su propia dispersión —que es lo que estudiamos en las dos transparencias siguientes.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, ecuación (2.35) y comentario adyacente.

6 Insesgadez: $\mathrm{E}(\hat{\beta}_2) = \beta_2$

Cambiemos de pregunta:

- de "¿Cuánto vale la estimación $\hat{\beta}_2$ con los datos disponibles $\boldsymbol{x}$ e $\boldsymbol{y}$?"
- a "¿cuánto vale el estimador $\hat{\beta}_2$, en promedio, sobre todas las muestras posibles?" (su proyección sobre $\mathit{1}$, lección 9).

  Lema (demostrado en los apuntes): si $c(\boldsymbol{X})$ es una función de $\boldsymbol{X}$,

$$E\big[c(\boldsymbol{X})Z \mid \boldsymbol{X}\big] = c(\boldsymbol{X})E[Z \mid \boldsymbol{X}].$$

Aplicándolo con $c(\boldsymbol{X}) = W_i$ (pues cada W_i es función de $\boldsymbol{X}$, sección anterior):

$$E[\hat{\beta}_2 \mid \boldsymbol{X}] = \beta_2\, \mathit{1} + \sum_i W_i\, E[U_i \mid \boldsymbol{X}] = \beta_2\, \mathit{1} + \sum_i W_i\, \mathit{0} = \beta_2\, \mathit{1}.$$

Ley de las esperanzas iteradas: Como $E[Z \mid \boldsymbol{X}]$ es la proyección ortogonal de Z sobre el espacio de las funciones de $\boldsymbol{X}$ (y $\mathit{1}$ es una de ellas), entonces $Z - E[Z \mid \boldsymbol{X}] \perp \mathit{1}$. Por tanto,

$$\mathrm{E}(Z) = \mathrm{E}(E[Z \mid \boldsymbol{X}]).$$

$$\boxed{\mathrm{E}(\hat{\beta}_2) = \mathrm{E}(E[\hat{\beta}_2 \mid \boldsymbol{X}]) = \beta_2.}$$

$\hat{\beta}_2$ no acierta en cada muestra — pero no se equivoca *sistemáticamente* en ninguna dirección.

Insesgadez: $\mathrm{E}(\hat{\beta}_2) = \beta_2$

Partimos de $\hat{\beta}_2 = \beta_2\, \mathit{1} + \sum_i W_i U_i$ (transparencia anterior) y tomamos esperanza condicionada a $\boldsymbol{X}$ —el primer paso de la sesión que de verdad usa probabilidad—.

Antes de seguir, conviene explicitar algo que en la lección 9 quedó implícito en la propia definición de $E[Z \mid \boldsymbol{X}]$ como la función de $\boldsymbol{X}$ que minimiza la distancia $\|Z - g(\boldsymbol{X})\|_P$. Exactamente

igual que en $\mathbb{R}^n$ (lección 4 para una recta, y lección 6 al pasar a un plano: minimizar la distancia de un vector a un subespacio equivale a que el residuo sea ortogonal a *todo* el subespacio —a cada uno de sus generadores—), esa minimización es equivalente a esta caracterización, que usaremos repetidamente de aquí en adelante:

> *Caracterización de* $E[Z \mid \boldsymbol{X}]$. Es la única variable aleatoria (salvo, como advertimos en la lección 9, en un conjunto de sucesos de probabilidad nula) que cumple, a la vez:
>
> 1. *Pertenencia*: es una función de $\boldsymbol{X}$.
> 2. *Ortogonalidad del residuo*: $\mathrm{E}\Big((Z - E[Z \mid \boldsymbol{X}])h(\boldsymbol{X})\Big) = 0$ para *toda* función $h(\boldsymbol{X})$.

De esta caracterización salen, con el mismo tipo de comprobación (verificar que el candidato pertenece al espacio y que el residuo es ortogonal), dos propiedades que usaremos continuamente sin volver a citarlas: $E[\,\cdot \mid \boldsymbol{X}]$ es *lineal*, $E[aZ + bW \mid \boldsymbol{X}] = a\,E[Z \mid \boldsymbol{X}] + b\,E[W \mid \boldsymbol{X}]$, y *deja fijas* las funciones de $\boldsymbol{X}$, $E[g(\boldsymbol{X}) \mid \boldsymbol{X}] = g(\boldsymbol{X})$ —en particular, $E[\beta_2\, 1 \mid \boldsymbol{X}] = \beta_2\, 1$—. Y con esta misma caracterización podemos demostrar el siguiente lema:

> *Lema (sacar fuera de la esperanza condicionada las funciones de variables sobre las que se condiciona).* Si $c(\boldsymbol{X})$ es una función de la lista de variables aleatorias $\boldsymbol{X}$,
>
> $$E[c(\boldsymbol{X})Z \mid \boldsymbol{X}] = c(\boldsymbol{X})E[Z \mid \boldsymbol{X}].$$

Demostración. Comprobamos que el candidato $c(\boldsymbol{X})E[Z \mid \boldsymbol{X}]$ cumple las DOS propiedades que caracterizan a $E[c(\boldsymbol{X})Z \mid \boldsymbol{X}]$:

- *Pertenencia*: el producto $c(\boldsymbol{X})E[Z \mid \boldsymbol{X}]$ es función de $\boldsymbol{X}$ (producto de funciones de $\boldsymbol{X}$).
- *Ortogonalidad*: para cualquier $h(\boldsymbol{X})$,

$$\mathrm{E}\Big(\big(c(\boldsymbol{X})Z - c(\boldsymbol{X})E[Z \mid \boldsymbol{X}]\big)\,h(\boldsymbol{X})\Big) = \mathrm{E}\Big((Z - E[Z \mid \boldsymbol{X}])\,\underbrace{c(\boldsymbol{X})h(\boldsymbol{X})}_{\text{función de } \boldsymbol{X}}\Big) = 0$$

porque $c(\boldsymbol{X})h(\boldsymbol{X})$ es, ella misma, función de $\boldsymbol{X}$, y ya sabemos que $Z - E[Z \mid \boldsymbol{X}]$ es ortogonal a *cualquiera* de ellas.

Como $c(\boldsymbol{X})E[Z \mid \boldsymbol{X}]$ cumple las dos propiedades que caracterizan a $E[c(\boldsymbol{X})Z \mid \boldsymbol{X}]$, concluimos $E[c(\boldsymbol{X})Z \mid \boldsymbol{X}] = c(\boldsymbol{X})E[Z \mid \boldsymbol{X}]$. ■

Aplicamos el lema con $c(\boldsymbol{X}) = W_i$ y con $Z = U_i$:

$$E[\hat{\beta}_2 \mid \boldsymbol{X}] = \beta_2\, 1 + \sum_i E[W_i U_i \mid \boldsymbol{X}] = \beta_2\, 1 + \sum_i W_i\, E[U_i \mid \boldsymbol{X}] = \beta_2\, 1 + \sum_i W_i\, 0 = \beta_2\, 1,$$

donde hemos usado la exogeneidad estricta $E[U_i \mid \boldsymbol{X}] = 0$ de la sección "El marco: muestra aleatoria".

Falta un último paso, pasar de la esperanza condicional a la esperanza incondicional: queremos encontrar $\mathrm{E}(\hat{\beta}_2)$, no $E[\hat{\beta}_2 \mid \boldsymbol{X}]$. Aquí es donde entra la Ley de las Esperanzas Iteradas, que —como puede comprobarse— es, en realidad, la *misma* idea que ya usamos en la lección 9 para deducir $\mathrm{E}(U) = 0$ a partir de $U \perp 1$. En general: la función constante 1 es, trivialmente, una función

de $\boldsymbol{X}$; aplicando la ortogonalidad del residuo (recordada más arriba) con $h(\boldsymbol{X}) = \mathit{1}$, obtenemos $Z - E[Z \mid \boldsymbol{X}] \perp \mathit{1}$ para *cualquier* Z

$$\mathrm{E}\Big(\big(Z - E[Z \mid \boldsymbol{X}]\big)\,\mathit{1}\Big) = 0 \iff \mathrm{E}(Z) - \mathrm{E}(E[Z \mid \boldsymbol{X}]) = 0,$$

es decir, $\mathrm{E}(E[Z \mid \boldsymbol{X}]) = \mathrm{E}(Z)$ —sin ninguna restricción sobre Z—. Esta es la Ley de las Esperanzas Iteradas, una consecuencia inmediata de nuestra propia definición geométrica de esperanza condicional.

Aplicándola a $Z = \hat{\beta}_2$, y usando que $E[\hat{\beta}_2 \mid \boldsymbol{X}] = \beta_2\,\mathit{1}$ es constante:

$$\mathrm{E}(\hat{\beta}_2) = \mathrm{E}(E[\hat{\beta}_2 \mid \boldsymbol{X}]) = \mathrm{E}(\beta_2\,\mathit{1}) = \beta_2.$$

$\hat{\beta}_2$ es, por tanto, un *estimador insesgado* de β_2: en promedio, sobre todas las muestras posibles que podríamos haber observado, ni sobreestima ni subestima el verdadero parámetro poblacional.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, Teorema 2.1 (insesgadez de $\hat{\beta}_1, \hat{\beta}_2$).

7 Varianza: cuán dispersas están las estimaciones

Desarrollando el cuadrado (apuntes) y aplicando el *Lema* anterior:

$$Var[\hat{\beta}_2 \mid \boldsymbol{X}] = Var\Big[\textstyle\sum_i W_i U_i \,\Big|\, \boldsymbol{X}\Big] = \sum_i W_i^2\, Var[U_i \mid \boldsymbol{X}] + \sum_{i \neq j} W_i W_j\, Cov[U_i, U_j \mid \boldsymbol{X}]$$

Homocedasticidad + $Cov[U_i, U_j \mid \boldsymbol{X}] = \mathit{0}$ (“El marco”) $\implies$

$$Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \sigma^2 \sum_i W_i^2.$$

$$\boxed{Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \frac{\sigma^2}{\sum_i (X_i - \overline{X})^2} = \frac{\sigma^2}{n\,S^2}.}$$

donde $S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$ es la *varianza muestral* del regresor (divisor n, como en la lección 5).

Más dispersión en X, o más observaciones: más precisión. Más ruido (σ^2): menos precisión.

Versión interactiva: cuaderno 4 en MyBinder (muchas muestras, muchas rectas; el histograma de $\hat{\beta}_2$ y el efecto de la dispersión de $\boldsymbol{x}$).

Varianza: cuán dispersas están las estimaciones

La insesgadez nos dice que $\hat{\beta}_2$ está bien “centrado” en β_2. Pero un estimador insesgado puede, aun así, ser muy poco fiable en una muestra concreta si su dispersión es enorme. Necesitamos, por tanto, calcular $Var[\hat{\beta}_2 \mid \boldsymbol{X}]$.

Partimos de $\hat{\beta}_2 = \beta_2\,\mathit{1} + \sum_i W_i U_i$. Como $\beta_2\,\mathit{1}$ es una variable aleatoria constante: $Var[\hat{\beta}_2 \mid \boldsymbol{X}] = Var\big[\sum_i W_i U_i \mid \boldsymbol{X}\big]$. Nos queda, por tanto, desarrollar la varianza de una suma de n sumandos. Lo haremos en dos pasos: primero calculamos la varianza y la covarianza de cada sumando por separado

(aquí interviene el Lema), y después desarrollamos la suma completa. Calculemos primero, para un único sumando, $Var[W_iU_i \mid \boldsymbol{X}]$. Ya vimos que $E[W_iU_i \mid \boldsymbol{X}] = 0$ (Lema, con $c(\boldsymbol{X}) = W_i$). Por tanto, la definición de varianza condicional ($Var[Z \mid \boldsymbol{X}] = E[(Z - E[Z \mid \boldsymbol{X}])^2 \mid \boldsymbol{X}]$) se simplifica exactamente como en la lección 9 ($Var[U \mid X] = E[U^2 \mid X]$, porque también allí $E[U \mid X] = 0$):

$$Var[W_iU_i \mid \boldsymbol{X}] = E[(W_iU_i)^2 \mid \boldsymbol{X}] = E[W_i^2U_i^2 \mid \boldsymbol{X}] = W_i^2\, E[U_i^2 \mid \boldsymbol{X}] = W_i^2\, Var[U_i \mid \boldsymbol{X}],$$

donde la penúltima igualdad es, de nuevo, el Lema (con $c(\boldsymbol{X}) = W_i^2$, sobre $Z = U_i^2$), y la última vuelve a usar que $E[U_i \mid \boldsymbol{X}] = 0$.

Por el mismo argumento, para $i \neq j$ (usando la covarianza condicional definida en "El marco: muestra aleatoria", $Cov[Z, W \mid \boldsymbol{X}] = E[(Z - E[Z \mid \boldsymbol{X}])(W - E[W \mid \boldsymbol{X}]) \mid \boldsymbol{X}]$, análoga a la varianza):

$$Cov[W_iU_i, W_jU_j \mid \boldsymbol{X}] = E[W_iU_iW_jU_j \mid \boldsymbol{X}] = W_iW_j\, E[U_iU_j \mid \boldsymbol{X}] = W_iW_j\, Cov[U_i, U_j \mid \boldsymbol{X}],$$

usando el Lema con $c(\boldsymbol{X}) = W_iW_j$.

Con estas dos piezas ya podemos desarrollar la suma entera. Abreviando $Z_i = W_iU_i$ —de modo que $E[Z_i \mid \boldsymbol{X}] = 0$ para todo i, como acabamos de comprobar—, la definición de varianza condicional da

$$Var\Big[\sum_i Z_i \,\Big|\, \boldsymbol{X}\Big] = E\Big[\Big(\sum_i Z_i\Big)^2 \,\Big|\, \boldsymbol{X}\Big] = E\Big[\sum_i \sum_j Z_iZ_j \,\Big|\, \boldsymbol{X}\Big] = \sum_i \sum_j E[Z_iZ_j \mid \boldsymbol{X}],$$

donde solo hemos desarrollado el cuadrado de la suma y usado después la linealidad de la esperanza *condicional* (recordada junto al Lema, en la sección "Insesgadez"), aplicada ahora a una suma de n^2 términos. Basta ya con separar los términos de la diagonal ($i = j$) del resto: como todas las Z_i tienen esperanza condicional nula, $E[Z_iZ_j \mid \boldsymbol{X}]$ es $Var[Z_i \mid \boldsymbol{X}]$ cuando $i = j$, y $Cov[Z_i, Z_j \mid \boldsymbol{X}]$ cuando $i \neq j$. Sustituyendo en ambos casos las dos piezas calculadas más arriba:

$$Var\Big[\sum_i W_iU_i \,\Big|\, \boldsymbol{X}\Big] = \sum_i W_i^2\, Var[U_i \mid \boldsymbol{X}] + \sum_{i \neq j} W_iW_j\, Cov[U_i, U_j \mid \boldsymbol{X}].$$

Ahora entran en juego, por fin, la homocedasticidad ($Var[U_i \mid \boldsymbol{X}] = \sigma^2\, 1$ para todo i, supuesto de la lección 9) y la ausencia de covarianza cruzada ($Cov[U_i, U_j \mid \boldsymbol{X}] = 0$ para $i \neq j$, consecuencia del muestreo aleatorio, sección "El marco: muestra aleatoria"). Con ambas, la doble suma se reduce a

$$Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \sigma^2 \sum_i W_i^2.$$

Calculemos $\sum_i W_i^2$. Como $W_i = \frac{X_i - \overline{X}}{\sum_j (X_j - \overline{X})^2}$:

$$\sum_i W_i^2 = \sum_i \left(\frac{X_i - \overline{X}}{\sum_j (X_j - \overline{X})^2}\right)^2 = \frac{\sum_i (X_i - \overline{X})^2}{\Big(\sum_j (X_j - \overline{X})^2\Big)^2} = \frac{1}{\sum_j (X_j - \overline{X})^2}.$$

Sustituyendo:

$$Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \frac{\sigma^2}{\sum_i (X_i - \overline{X})^2}.$$

Esta expresión admite una segunda lectura, útil para su interpretación. Como $\sum_i (X_i - \overline{X})^2 = n \cdot \frac{1}{n}\sum_i (X_i - \overline{X})^2 = n\, S^2$ (la varianza muestral del regresor, con la convención estadística del curso, $S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$), podemos escribir, equivalentemente,

$$Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \frac{\sigma^2}{n\, S^2}.$$

Esta segunda forma hace explícitos los tres ingredientes que determinan la precisión de $\hat{\beta}_2$. (i) Cuanto mayor sea σ^2 (más ruido en la perturbación), menor será la precisión. (ii) Cuanto mayor sea n (más observaciones), mayor será la precisión. (iii) Cuanto mayor sea S^2 (más dispersión en los valores del regresor), mayor será la precisión: una muestra donde X apenas varía da muy poca información sobre la pendiente que relaciona X con Y.[10]

Nótese, por último, que este resultado es *condicional* a $\boldsymbol{X}$: distintas muestras, con distintos valores del regresor, tendrían distinta precisión. Si quisiéramos la varianza *incondicional* $\mathrm{Var}(\hat{\beta}_2)$, tendríamos que aplicar la *ley de la varianza total*, que no es sino la identidad pitagórica de la lección 9 (allí enunciada para Y, condicionando en X), aplicada ahora a $Z = \hat{\beta}_2$ condicionando en $\boldsymbol{X}$. Para cualquier variable aleatoria Z con varianza finita, $E[Z \mid \boldsymbol{X}]$ es la proyección ortogonal de Z sobre las funciones de $\boldsymbol{X}$, así que $Z = E[Z \mid \boldsymbol{X}] + (Z - E[Z \mid \boldsymbol{X}])$ descompone Z en dos partes ortogonales entre sí. Aplicando Pitágoras a esa descomposición, y la ley de las esperanzas iteradas para escribir $\mathrm{E}((Z - E[Z \mid \boldsymbol{X}])^2)$ como $\mathrm{E}(Var[Z \mid \boldsymbol{X}])$, se obtiene

$$\mathrm{Var}(\hat{\beta}_2) = \mathrm{E}(Var[\hat{\beta}_2 \mid \boldsymbol{X}]) + \mathrm{Var}(E[\hat{\beta}_2 \mid \boldsymbol{X}]),$$

[10] *Un apunte geométrico, antes de seguir.* La fórmula de W_i no es una construcción de esta lección: ya estaba latente en la propia definición de c_2, el coeficiente que la "vía alternativa" de la lección 7 nos dio como pendiente. Recordemos aquella fórmula:

$$c_2 = \frac{\left\langle \boldsymbol{y},\, (\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}) \right\rangle_s}{\|\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}\|_s^2} = \hat{\beta}_2$$

Si desarrollamos ambos productos escalares como sumas —cancelando el factor $1/n$ entre numerador y denominador, la misma observación que hicimos al abrir esta sesión, al pasar de $\langle \cdot, \cdot \rangle_s$ a la expresión con sumatorios—, obtenemos

$$c_2 = \frac{\frac{1}{n}\sum_i (x_i - \mu_{\boldsymbol{x}}) y_i}{\frac{1}{n}\sum_j (x_j - \mu_{\boldsymbol{x}})^2} = \frac{\sum_i (x_i - \mu_{\boldsymbol{x}}) y_i}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2} = \sum_i \underbrace{\left(\frac{(x_i - \mu_{\boldsymbol{x}})}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2} \right)}_{w_i} y_i.$$

Es decir: los pesos $w_i = (x_i - \mu_{\boldsymbol{x}}) / \sum_j (x_j - \mu_{\boldsymbol{x}})^2$ —aquí evaluados sobre un vector de datos fijo $\boldsymbol{x}$, no sobre la muestra aleatoria $\boldsymbol{X}$— no son más que c_2 descompuesto sumando a sumando. No hay ninguna fórmula nueva: es el mismo cociente de productos escalares de la lección 3 (Cauchy–Schwarz), el que la lección 7 aplicó al generador $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$, ahora simplemente mirada componente a componente. Esta reconexión nos regala, sin ningún cálculo adicional, la cantidad $\sum_i w_i^2$ que acabamos de necesitar para la varianza. Como el denominador $\sum_j (x_j - \mu_{\boldsymbol{x}})^2$ es común a todos los w_i:

$$\sum_i w_i^2 = \sum_i \frac{(x_i - \mu_{\boldsymbol{x}})^2}{\left(\sum_j (x_j - \mu_{\boldsymbol{x}})^2\right)^2} = \frac{\sum_i (x_i - \mu_{\boldsymbol{x}})^2}{\left(\sum_j (x_j - \mu_{\boldsymbol{x}})^2\right)^2} = \frac{1}{\sum_j (x_j - \mu_{\boldsymbol{x}})^2},$$

la misma expresión de unas líneas más arriba —allí evaluado sobre la muestra aleatoria $\boldsymbol{X}$; aquí, sobre el vector de datos $\boldsymbol{x}$—. Evaluada en la muestra observada, la varianza de $\hat{\beta}_2$ es, pues, σ^2 dividido entre el cuadrado de la longitud *euclídea* (lección 2, sin el factor $1/n$) del vector en desviaciones del regresor, $\|\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}\|_e^2$ —el mismo vector con el que, en la lección 7, calculamos la propia pendiente—. Cuanto más "largo" sea ese vector en desviaciones, menor la varianza: la misma intuición de "más dispersión, más precisión" que ya leímos en la fórmula con S^2, ahora vista como una propiedad exacta de la longitud de un vector concreto en $\mathbb{R}^n$.

donde el segundo sumando es cero (pues $E[\hat{\beta}_2 \mid \boldsymbol{X}] = \beta_2\, \mathit{1}$ es constante, sección "Insesgadez", y la varianza de una constante es cero). Pero el primer sumando, $\mathrm{E}(\sigma^2/(nS^2))$, es la esperanza de un *cociente*, y no se simplifica en general (la esperanza de $1/S^2$ no es $1/\mathrm{E}(S^2)$). Por eso, en la práctica —y en el resto de este curso—, trabajaremos siempre con la fórmula *condicional* recién obtenida, evaluada en la muestra efectivamente observada: es, de hecho, exactamente lo que hace cualquier programa estadístico al reportar el error estándar de un coeficiente.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, Teorema 2.2 (varianza de $\hat{\beta}_1, \hat{\beta}_2$ bajo los supuestos SLR.1-SLR.5).

8 MCO es BLUE; estimando σ^2

Sin demostrarlo (Gauss–Markov): de entre todos los estimadores *lineales e insesgados* de β_2, $\hat{\beta}_2$ tiene la menor varianza — MCO es *BLUE* (*Best Linear Unbiased Estimator*).

σ^2 (desconocido) se estima con la *cuasivarianza de los residuos*, un *tercer* producto escalar, tras el euclídeo y el estadístico:

$$\langle \cdot, \cdot \rangle_{n-k} = \frac{1}{n-k} \langle \cdot, \cdot \rangle_e, \qquad \mathfrak{s}^2 = \langle \widehat{\boldsymbol{e}}, \widehat{\boldsymbol{e}} \rangle_{n-k} = \frac{\mathrm{SRC}}{n-k}.$$

Tres divisores, tres motivos distintos: 1 (geometría pura, sin más), n (para que $\|\mathbf{1}\|_s = 1$, lección 4), $n-k$ (para que $\mathfrak{s}^2$ sea un estimador insesgado de σ^2).

MCO es BLUE; estimando σ^2

Con insesgadez y varianza ya establecidas, cerramos con dos observaciones importantes que, por su naturaleza, no vamos a demostrar en este curso, pero que conviene conocer con precisión.

La primera es el *Teorema de Gauss–Markov*: bajo los tres supuestos del Modelo Clásico de Regresión Lineal (y el muestreo aleatorio de esta sesión), $\hat{\beta}_2$ no es solo insesgado; es, además, el estimador de *menor varianza* entre todos los estimadores *lineales* e *insesgados*. Lineal quiere decir de la forma $\sum_i C_i Y_i$, donde $Y_1, \ldots, Y_n$ son las variables aleatorias de la muestra (no los datos observados $y_1, \ldots, y_n$) y los pesos C_i no dependen de $\boldsymbol{Y}$ (van con mayúscula porque, igual que los W_i, pueden ser funciones de $\boldsymbol{X}$). Este resultado se resume con el acrónimo *BLUE* (*Best Linear Unbiased Estimator*, "el mejor estimador lineal insesgado"). No lo demostraremos: la demostración general requiere comparar $\hat{\beta}_2$ con un estimador lineal insesgado arbitrario, con maquinaria matricial que este curso no desarrolla. Pero merece una mención explícita, porque es la justificación teórica última de por qué, entre todas las rectas que podríamos ajustar de otras maneras, elegimos la de mínimos cuadrados.

La segunda observación es más práctica: la fórmula $Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \sigma^2 / \sum_i (X_i - \overline{X})^2$ depende de σ^2, la varianza de la perturbación poblacional, que no conocemos. Para usarla en la práctica (algo que haremos en la próxima sesión teórica, al construir errores estándar) necesitamos estimar σ^2 a partir de los residuos $\widehat{\boldsymbol{e}}$. Estos residuos observados se interpretan aquí como la realización, sobre nuestra muestra concreta, del vector aleatorio de residuos $\widehat{\boldsymbol{E}}$, cuyas componentes son $\hat{E}_i = Y_i - \hat{\beta}_1\, \mathit{1} - \hat{\beta}_2 X_i$; es el mismo paso de variable aleatoria a dato que hemos dado varias veces en esta

sesión. Esas componentes aproximan, sin ser idénticas a ellas, a las perturbaciones U_i del modelo poblacional.[11]

La estimación natural sería $\langle \widehat{\boldsymbol{e}}, \widehat{\boldsymbol{e}} \rangle_s = \mathrm{SRC}/n$ —la varianza de los residuos, con nuestra convención estadística habitual—. Pero puede demostrarse (no lo haremos aquí) que el estimador subyacente a esta fórmula —es decir, la misma fórmula, evaluada sobre la muestra aleatoria $\widehat{\boldsymbol{E}}$ en lugar de sobre los datos observados $\widehat{\boldsymbol{e}}$— es *sesgado*: subestima σ^2 en promedio.[12] La razón de este sesgo es geométrica, y ya la conocemos en parte: por construcción (ecuaciones normales, lección 10), los residuos $\widehat{\boldsymbol{e}}$ no son un vector arbitrario de $\mathbb{R}^n$, sino que viven siempre en el subespacio ortogonal a las columnas de $\mathbf{X}$. Como ese subespacio de regresores tiene dimensión k (hay k regresores, incluida la constante), su complemento ortogonal tiene dimensión $n-k$ —la misma cuenta informal de dimensión que empleamos ya en la lección 4 para $\mathcal{L}(\mathbf{1})^\perp$, de dimensión $n-1$, el caso particular $k=1$—. Dicho de otro modo: aunque $\widehat{\boldsymbol{e}}$ tiene n componentes, solo $n-k$ de ellas son realmente "libres"; las k restantes quedan fijadas por las k condiciones de ortogonalidad $\mathbf{X}^\top \widehat{\boldsymbol{e}} = \mathbf{0}$. Dividir por $n-k$, en vez de por n, es la manera de tener en cuenta esta pérdida de k grados de libertad, y es lo que hace insesgado al estimador resultante. El estimador que sí resulta insesgado usa, por tanto, ese divisor distinto: $n-k$ (el número de observaciones menos el número total de coeficientes estimados, incluida la constante) en vez de n.[13]

Esto nos lleva, de forma natural, a introducir un *tercer* producto escalar en la familia que el curso ha ido construyendo desde la lección 2: junto al euclídeo $\langle \cdot, \cdot \rangle_e$ (divisor 1) y al estadístico $\langle \cdot, \cdot \rangle_s = \langle \cdot, \cdot \rangle_n$ (divisor n), definimos

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle_{n-k} = \frac{1}{n-k} \langle \boldsymbol{v}, \boldsymbol{w} \rangle_e .$$

Sigue siendo, técnicamente, un producto escalar legítimo (multiplicar por una constante positiva no rompe simetría, linealidad ni positividad),[14] y con él definimos la *cuasivarianza de los residuos*

$$\mathfrak{s}^2 = \langle \widehat{\boldsymbol{e}}, \widehat{\boldsymbol{e}} \rangle_{n-k} = \frac{\mathrm{SRC}}{n-k},$$

que sí es un estimador insesgado de σ^2 (afirmación que no demostramos). Merece la pena subrayar que el motivo último de cada uno de los tres divisores es distinto. El divisor 1 no obedece a ningún motivo particular (es la geometría euclídea sin más). El divisor n se eligió, en la lección 4, para que $\|\mathbf{1}\|_s = 1$: un motivo puramente geométrico. El divisor $n-k$ se elige aquí por un motivo *estadístico*

[11] Esta es la posición de Eros con la que abríamos la lección 9: ni conocemos U_i (eso sería la posición del dios), ni carecemos de toda información sobre ella (la posición del ignorante); $\hat{E}_i$ es nuestro mejor sustituto calculable, construido para que, bajo los supuestos del modelo, su comportamiento medio reproduzca el de U_i.

[12] Seguimos aquí, por simplicidad, la misma convención del resto de esta sección: escribimos las fórmulas con los datos observados $\widehat{\boldsymbol{e}}$, aunque hablemos de propiedades —como el sesgo o la insesgadez— que, en sentido estricto, son propiedades del estimador (la variable aleatoria subyacente), no de un número concreto. Es el mismo tipo de abuso de notación, ya señalado al distinguir estimador de estimación, habitual en toda la literatura econométrica; introducir aquí toda la notación necesaria para evitarlo tendría un coste en claridad mayor que el beneficio.

[13] Esta idea ya la usamos, sin justificarla del todo, en la práctica de laboratorio que sigue a la lección 10, al motivar por qué dividir por $n-k$ (y no por n) produce estimadores insesgados. Precisemos ahora, con algo más de rigor, el motivo exacto de ese divisor.

[14] El subíndice $n-k$ indica por qué número se divide. Con ese mismo criterio los tres productos escalares del curso admiten una notación unificada: $\langle \cdot, \cdot \rangle_e = \langle \cdot, \cdot \rangle_1$ (se divide por 1), $\langle \cdot, \cdot \rangle_s = \langle \cdot, \cdot \rangle_n$ (por n) y $\langle \cdot, \cdot \rangle_{n-k}$ (por $n-k$). Mantendremos los subíndices e y s, ya habituales, pero conviene saber que la familia es una sola.

—lograr insesgadez—, aunque, como acabamos de ver, admite también una lectura geométrica: es la dimensión del subespacio donde vive $\widehat{\boldsymbol{e}}$. El símbolo $\mathfrak{s}^2$ (con la ese gótica) se reserva para esta cantidad, para no confundirla con la varianza muestral S^2 (introducida hoy) ni con la s^2 de los manuales clásicos de estadística, que suele denotar esto mismo con otra convención tipográfica.

La expresión $n-k$ recibe, en la jerga habitual de la estadística, el nombre de *grados de libertad* del modelo: el número de observaciones menos el número de parámetros ya estimados. En esta sesión, con dos regresores en total, $k=2$, así que $\mathfrak{s}^2 = \text{SRC}/(n-2)$. Volveremos sobre los grados de libertad, con más detalle, en la próxima sesión teórica.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, Teorema 2.2 (parte final: BLUE) y cap. 3, sección 3.5 (Teorema de Gauss-Markov, versión general con k regresores).

- Wooldridge, J. M. (2020), cap. 3, sección 3.6 (estimación insesgada de σ^2, grados de libertad).

9 ¿Y con más regresores?

Sin demostrarlo (generalización matricial de lo probado hoy; recordando la notación de la lección 10 y su extensión aleatoria de "El marco"):

$$E[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \boldsymbol{\beta}\, \boldsymbol{1}, \qquad Var[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

En la diagonal de esta matriz están $Var[\hat{\beta}_1 \mid \boldsymbol{X}], \ldots, Var[\hat{\beta}_k \mid \boldsymbol{X}]$; fuera de la diagonal, las covarianzas $Cov[\hat{\beta}_i, \hat{\beta}_j \mid \boldsymbol{X}]$.

Evaluada en una muestra ya observada, y con σ^2 sustituido por su estimación $\mathfrak{s}^2$ (transparencia anterior) —la fórmula que de hecho calculan los programas estadísticos—:

$$\mathfrak{s}^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Para $k=2$, el elemento $(2,2)$ de esta fórmula matricial se reduce exactamente a lo que hoy hemos demostrado sin matrices (y los otros tres elementos dan, de propina, lo que no hemos calculado).

¿Y con más regresores?

Todo lo demostrado en esta sesión se limita, deliberadamente, al caso con dos regresores: la constante *1* y un único regresor no constante X. Con k regresores (lección 10), el resultado análogo existe, y su enunciado —sin demostración— es la generalización matricial natural de lo que hoy hemos probado:

$$E[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \boldsymbol{\beta}\, \boldsymbol{1}, \qquad Var[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

Aquí $\boldsymbol{X}$ es la misma matriz aleatoria de regresores que fijamos al abrir la sesión, solo que ahora con k columnas en lugar de dos; $\mathbf{X}$ sigue denotando, como en la lección 10, la matriz de datos *observados*. La primera igualdad generaliza la insesgadez de la sección "Insesgadez": $\boldsymbol{\beta}\, \boldsymbol{1}$ denota el vector cuyas componentes son las variables aleatorias constantes $\beta_1\, \boldsymbol{1}, \ldots, \beta_k\, \boldsymbol{1}$ —el análogo, para un vector de parámetros, de la constante $\beta_2\, \boldsymbol{1}$ que obtuvimos en la sección "Insesgadez"—. La segunda generaliza la varianza de la sección "Varianza": es una matriz $k \times k$ que ocupa el lugar que antes ocupaba el escalar $1/\sum_i (X_i - \overline{X})^2$; en su diagonal contiene las varianzas $Var[\hat{\beta}_1 \mid \boldsymbol{X}], \ldots, Var[\hat{\beta}_k \mid \boldsymbol{X}]$ (con

$Var[\hat{\beta}_2 \mid \boldsymbol{X}]$ como caso particular, ya conocido, en la posición $(2,2)$), y fuera de la diagonal, las covarianzas $Cov[\hat{\beta}_i, \hat{\beta}_j \mid \boldsymbol{X}]$ entre cada par de coeficientes.

Al evaluar esta fórmula en una muestra ya observada, sustituimos la matriz aleatoria $\boldsymbol{X}$ por la matriz de datos $\mathbf{X}$, con lo que queda por calcular $(\mathbf{X}^\top\mathbf{X})^{-1}$. Multiplicada por $\mathfrak{s}^2$ (la estimación de σ^2 de la transparencia anterior, pues σ^2 sigue siendo desconocido) da $\mathfrak{s}^2(\mathbf{X}^\top\mathbf{X})^{-1}$: la fórmula que Gretl (y cualquier otro programa estadístico) usa internamente para calcular los errores estándar que veremos en la próxima sesión teórica.

Es razonable preguntarse si esta fórmula, presentada aquí sin demostración, es compatible con lo que sí hemos demostrado hoy. La respuesta es sí, y podemos comprobarlo con un cálculo elemental. Para $k = 2$ (constante más un único regresor no constante), la matriz $\boldsymbol{X}^\top\boldsymbol{X}$ (análogo aleatorio de la $\mathbf{X}^\top\mathbf{X}$ de la lección 10) es

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{pmatrix} n\,1 & n\overline{X} \\ n\overline{X} & \sum_i X_i^2 \end{pmatrix}.$$

Usando la fórmula general para invertir una matriz 2×2, obtenemos[15]

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{1}{n\sum_i (X_i - \overline{X})^2} \begin{pmatrix} \sum_i X_i^2 & -n\overline{X} \\ -n\overline{X} & n\,1 \end{pmatrix}.$$

El elemento $(2,2)$ de esta matriz —el que corresponde a la varianza condicional del segundo coeficiente, $\hat{\beta}_2$— vale $n\,1/(n\sum_i (X_i - \overline{X})^2) = 1/\sum_i (X_i - \overline{X})^2$. Multiplicando por σ^2:

$$\sigma^2\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{22} = \frac{\sigma^2}{\sum_i (X_i - \overline{X})^2},$$

exactamente la fórmula de la sección "Varianza". Y, de propina, la misma matriz nos da las otras dos cantidades que esta sesión no ha llegado a calcular —al habernos centrado, como anunciamos al abrir, en $\hat{\beta}_2$—: el elemento $(1,1)$, multiplicado por σ^2, da $Var[\hat{\beta}_1 \mid \boldsymbol{X}] = \sigma^2 \sum_i X_i^2/(n\sum_i (X_i - \overline{X})^2)$; y el elemento $(1,2) = (2,1)$, multiplicado por σ^2, da $Cov[\hat{\beta}_1, \hat{\beta}_2 \mid \boldsymbol{X}] = -\sigma^2\overline{X}/\sum_i (X_i - \overline{X})^2$. La generalización matricial, aunque no la hayamos demostrado, es consistente con todo lo que sí hemos probado en esta sesión —y nos regala, sin coste adicional, dos fórmulas más que no habíamos calculado en la lección.

Para profundizar: Wooldridge, J. M. (2020), Apéndice E.2 (fórmulas matriciales de insesgadez y varianza para el modelo de regresión múltiple).

[15] $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, con $a = n\,1$, $b = c = n\overline{X}$, $d = \sum_i X_i^2$, y usando que $ad - bc = n\sum_i X_i^2 - n^2\overline{X}^2 = n\sum_i (X_i - \overline{X})^2$ (la "fórmula de cálculo" de la varianza, lección 5, evaluada sobre $\boldsymbol{X}$ y multiplicada por n).

10 Recapitulación y guiño a la sesión siguiente

Propiedad	Resultado (regresión simple, $k = 2$)	Supuestos usados
Insesgadez	$\mathrm{E}(\hat{\beta}_2) = \beta_2$	Linealidad + $\mathrm{Var}(X) \neq 0$ + muestreo aleatorio
Varianza	$Var[\hat{\beta}_2 \mid \boldsymbol{X}] = \frac{\sigma^2}{\sum_i (X_i - \overline{X})^2}$	+ Homocedasticidad (+ $Cov[U_i, U_j \mid \boldsymbol{X}] = 0$, del muestreo aleatorio)
Eficiencia	MCO es BLUE (mejor entre los lineales e insesgados)	Los tres + muestreo aleatorio (Gauss–Markov, sin demostrar)

Casi nada nuevo: la vía alternativa (lección 7), la esperanza condicional y la independencia entre observaciones (lección 9). Lo único nuevo: evaluar la fórmula sobre variables aleatorias, no sobre datos; eso, y solo eso, convierte $\hat{\beta}_2$ en un estimador.

Nota: los mismos argumentos, sobre $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$, prueban la insesgadez de $\hat{\beta}_1$.

Lección 12: ya conocemos la media y la varianza de $\hat{\beta}_2$; para intervalos de confianza y contrastes hace falta algo más: la *forma* de su distribución.

Recapitulación y guiño a la sesión siguiente

El recorrido de esta sesión, visto en conjunto, es notablemente corto en ingredientes nuevos. Hemos demostrado dos propiedades —insesgadez y varianza de $\hat{\beta}_2$— con piezas que ya teníamos. La fórmula de la "vía alternativa" de la lección 7 nos dio $\hat{\beta}_2$ como combinación lineal de los datos. La esperanza condicional de la lección 9 nos permitió tomar esperanzas y varianzas "condicionando en $\boldsymbol{X}$". Y la independencia entre observaciones anunciada en esa misma lección 9 ("copias idénticas e independientes").

Conviene decir con claridad qué es lo único genuinamente nuevo, porque es fácil confundirlo con algo más sofisticado: tomar una fórmula ARITMÉTICA ya obtenida —geométricamente, en la lección 7, sobre datos fijos— y evaluarla en variables aleatorias en lugar de en datos. Eso, sin más, es lo que convierte una estimación (un número, calculado a partir de una muestra concreta) en un estimador (una variable aleatoria, cuya distribución depende de qué muestra hubiéramos observado). No hemos definido ningún producto escalar nuevo sobre "vectores de variables aleatorias", ni hemos vuelto a demostrar Cauchy–Schwarz sobre un objeto distinto. Sencillamente, hemos sustituido números por variables aleatorias dentro de una fórmula que ya sabíamos cierta, y hemos reservado la maquinaria probabilística de la lección 9 (esperanza condicional, independencia) para el momento en que de verdad hacía falta: al tomar esperanzas y varianzas de esa suma ponderada.

La tabla resume las dos propiedades centrales, junto con los supuestos exactos que cada una requiere: la insesgadez solo necesita linealidad (más el muestreo aleatorio, que es lo que eleva la exogeneidad estricta al condicionamiento en toda la matriz $\boldsymbol{X}$); la varianza necesita, además, homocedasticidad y la ausencia de covarianza cruzada entre perturbaciones (también consecuencia del muestreo aleatorio). Esta jerarquía de supuestos —cada propiedad exige un subconjunto distinto de los tres supuestos del Modelo Clásico— es importante tenerla presente: en las próximas sesiones, cuando alguno de estos supuestos se ponga en duda (más adelante, al tratar la heterocedasticidad, por ejemplo), sabremos exactamente qué propiedad de $\hat{\beta}_2$ queda en entredicho y cuál sigue intacta.

Queda, sin embargo, una limitación importante que esta sesión no resuelve. Saber que $\hat{\beta}_2$ está

bien centrado en β_2 (insesgadez) y saber cuánto se dispersa alrededor de ese centro (varianza) NO basta para responder preguntas del tipo "¿es $\hat{\beta}_2 = 0{,}5$ una estimación compatible con $\beta_2 = 0$?" o "¿en qué rango de valores podemos confiar razonablemente que está el verdadero β_2?". Para responderlas —son los contrastes de hipótesis y los intervalos de confianza que ocuparán buena parte del resto del curso— no basta con la media y la varianza de $\hat{\beta}_2$: hace falta conocer, además, la *forma* completa de su distribución. Esa es la tarea de la próxima sesión teórica: enunciar (sin demostrarla desde cero, apoyándonos en un supuesto adicional sobre la distribución de la perturbación U) la distribución de $\hat{\beta}_2$, y a partir de ella construir el error estándar, el intervalo de confianza y el contraste t.

Para profundizar: Wooldridge, J. M. (2020), cap. 2, sección 2.6 (unidades de medida y forma funcional; transición natural hacia la inferencia del capítulo 4).