Índice general

Lección 10. Regresión múltiple: de dos ortogonalidades a k

Marcos Bujosa

19 de septiembre de 2026

Resumen

Generalizamos la regresión a cualquier número de regresores. La geometría sigue siendo idéntica (la ortogonalidad lo gobierna todo), pero el sistema de ecuaciones crece. Para no asfixiarnos en sumatorios, introducimos una notación matricial natural basada en un *operador selector*. Descubriremos que el álgebra de matrices es solo una compresión elegante de los productos escalares y las combinaciones lineales que ya conocemos, y dejamos planteado, para desarrollar con datos reales en el laboratorio, el concepto de efecto parcial o *ceteris paribus*.

"Cuantos más regresores, mayor el espacio de búsqueda; la pregunta —¿qué punto del subespacio está más cerca de $\boldsymbol{y}$?— no cambia."

- (slides) — (html) — (pdf)

1 De dónde venimos y a dónde vamos: del plano al hiperplano

En la lección 6 proyectamos el vector de datos $\boldsymbol{y}$ sobre el *plano* generado por $\mathbf{1}$ y $\boldsymbol{x}$. La ortogonalidad del residuo frente a **todo el plano** se tradujo en **dos** condiciones.

En economía rara vez un fenómeno depende de un solo factor $\boldsymbol{x}$.

El paso natural: Añadir más regresores. Proyectaremos $\boldsymbol{y}$ sobre el subespacio generado por k regresores:

$$\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_k)$$

El ajuste será una combinación lineal de todos ellos:

$$\widehat{\boldsymbol{y}} \;=\; \hat{\beta}_1\mathbf{1} + \hat{\beta}_2\boldsymbol{x}_2 + \hat{\beta}_3\boldsymbol{x}_3 + \cdots + \hat{\beta}_k\boldsymbol{x}_k.$$

Y, recordando la ecuación con la que abríamos la lección 7, el residuo sigue siendo, sencillamente, la diferencia entre los datos y el ajuste:

$$\boldsymbol{y} \;=\; \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}.$$

El subespacio sobre el que proyectamos los datos ya no es ni una recta ni un plano; es un subespacio de dimensión k. Pero **la geometría es la misma**.

De dónde venimos y a dónde vamos: del plano al hiperplano

Hasta ahora hemos avanzado sumando dimensiones de una en una. En la lección 4 resolvimos la proyección más sencilla: buscar el vector más cercano a los datos $\boldsymbol{y}$ dentro de la recta $\mathcal{L}(\mathbf{1})$. Obtuvimos la media. En la lección 6 dimos un paso más: añadimos una segunda dirección, el regresor $\boldsymbol{x}$, y proyectamos sobre el plano $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$. Obtuvimos la regresión lineal simple.

Hoy damos el salto definitivo. En la realidad empírica, un solo regresor casi nunca es suficiente. Si queremos explicar el salario de una persona, no basta con conocer sus años de educación ($\boldsymbol{x}_2$); necesitamos conocer su experiencia laboral ($\boldsymbol{x}_3$), su antigüedad en la empresa ($\boldsymbol{x}_4$), etc.

Geométricamente, esto equivale a añadir más "ejes" o vectores generadores a nuestro subespacio. Si tenemos $k-1$ regresores no constantes más el vector de unos $\mathbf{1}$, el subespacio sobre el que proyectaremos será $\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$. Este subespacio ya no es un plano bidimensional, sino un subespacio de dimensión k (en el lenguaje coloquial, un "hiperplano").

¿Qué cambia en nuestra manera de afrontar el problema? Conceptualmente, **absolutamente nada**. Seguimos buscando el vector $\widehat{\boldsymbol{y}}$ (el ajuste) de ese subespacio que está más cerca de $\boldsymbol{y}$. Y, como sabemos desde la lección 3, ese punto de mínima distancia es la *proyección ortogonal*: el punto que garantiza que el vector de residuos $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ sea estrictamente perpendicular a todo el subespacio. Dicho de otro modo, seguimos exactamente en la misma ecuación con la que abrimos la lección 7, $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}$, solo que ahora $\widehat{\boldsymbol{y}}$ es una combinación lineal de k regresores en lugar de dos.

Para profundizar:

- Wooldridge, J. M. (2020), capítulo 3, sección 3.1: Motivación del modelo de regresión múltiple.
- Bujosa, M. *Curso de Álgebra Lineal*, capítulo 13, sección "Proyecciones sobre subespacios": libro online.

2 La geometría es la misma

- $\widehat{\boldsymbol{y}}$ sigue siendo la "sombra" de $\boldsymbol{y}$ sobre el subespacio de los regresores.
- $\widehat{\boldsymbol{e}}$ sigue siendo **perpendicular a todas las direcciones** de dicho subespacio.

La geometría es la misma

Para ilustrar la regresión múltiple rescatamos la misma figura geométrica que ya usamos en la lección 8 (al demostrar $R^2 = \cos^2\theta$). Esto no es un reciclaje por falta de imágenes, sino una decisión deliberada: en aquella lección fuimos muy cuidadosos de NO dibujar la componente $\hat{\beta}_2\boldsymbol{x}$ individualmente dentro del plano (algo que sí hicimos en la lección 6). En la lección 8 dibujamos únicamente la proyección $\overline{\boldsymbol{y}}$ sobre los vectores constantes y la proyección $\widehat{\boldsymbol{y}}$ sobre el subespacio formado por los regresores (que en aquel momento correspondía al plano engendrado por $\mathbf{1}$ y $\boldsymbol{x}$). Esa fue una decisión anticipada precisamente para esta transparencia: al no comprometerse con ningún regresor concreto, la figura sigue siendo una representación esquemática perfectamente válida cuando el subespacio pasa de tener dos direcciones a tener k. Lo único que ya no podríamos dibujar, si quisiéramos ser literales, son los $k-1$ ejes individuales $\mathcal{L}(\boldsymbol{x}_2), \ldots, \mathcal{L}(\boldsymbol{x}_k)$ dentro del plano —pero la figura nunca pretendió mostrarlos, así que no hay nada que corregir en ella.

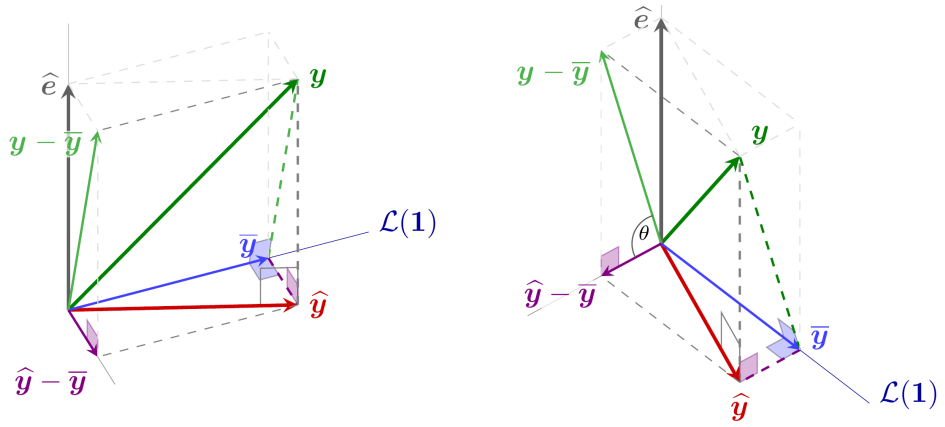

Figura 1: Proyección ortogonal de $\boldsymbol{y}$ sobre un subespacio (aquí representado esquemáticamente por el plano inferior), vista desde dos ángulos distintos. Es la misma figura de la regresión simple (lección 8), solo que ahora el plano inferior que forma el suelo de la figura no representa a $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$, ahora representa a un subespacio vectorial mucho más grande: $\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$ (como ya no es un plano, no lo podemos representar fielmente en la figura. Esta figura es solo un esquema mental; una representación abstracta, no una representación literal).

Esa abstracción nos permite ahora releer la misma figura para k regresores. El subespacio representado por el "suelo" de la figura ya no es un plano bidimensional, sino que representa esquemáticamente el inmenso subespacio $\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$ de k dimensiones.

Todo lo que dedujimos en la lección 8 respecto de esta figura sigue siendo cierto ahora:

1. El vector ajustado $\widehat{\boldsymbol{y}}$ es la proyección ortogonal.
2. El residuo $\widehat{\boldsymbol{e}}$ forma un ángulo de 90^o con el subespacio entero.
3. El triángulo rectángulo formado por $\boldsymbol{y} - \overline{\boldsymbol{y}}$, $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$ y $\widehat{\boldsymbol{e}}$ sigue existiendo, lo que significa que el teorema de Pitágoras sigue funcionando: STC = SEC + SRC ($\sigma^2_{\boldsymbol{y}} = \sigma^2_{\widehat{\boldsymbol{y}}} + \sigma^2_{\widehat{\boldsymbol{e}}}$) es igual de válido en la regresión múltiple.
4. El coeficiente de determinación se define igual: $R^2 = \cos^2\theta$, donde θ sigue siendo el ángulo entre los vectores en desviaciones de los datos y de su ajuste.

Si la geometría es idéntica y todas las conclusiones teóricas se mantienen, ¿dónde está la dificultad de la regresión múltiple? En la aritmética. En la siguiente transparencia veremos qué pasa cuando intentamos escribir las condiciones de ortogonalidad a mano.

3 El problema aritmético: demasiadas ecuaciones

La exigencia de que $\widehat{\boldsymbol{e}}$ sea ortogonal a TODO el subespacio implica que debe ser ortogonal a **cada uno de sus generadores**.

Tendremos k condiciones de ortogonalidad simultáneas:

$$\begin{aligned}
\langle \widehat{\boldsymbol{e}}, \mathbf{1} \rangle_s = 0 \quad &\Longrightarrow \tfrac{1}{n} \textstyle\sum \hat{e}_i = 0 \\
\langle \widehat{\boldsymbol{e}}, \boldsymbol{x}_2 \rangle_s = 0 \quad &\Longrightarrow \tfrac{1}{n} \textstyle\sum x_{i2} \hat{e}_i = 0 \\
\langle \widehat{\boldsymbol{e}}, \boldsymbol{x}_3 \rangle_s = 0 \quad &\Longrightarrow \tfrac{1}{n} \textstyle\sum x_{i3} \hat{e}_i = 0 \\
&\dots \\
\langle \widehat{\boldsymbol{e}}, \boldsymbol{x}_k \rangle_s = 0 \quad &\Longrightarrow \tfrac{1}{n} \textstyle\sum x_{ik} \hat{e}_i = 0
\end{aligned}$$

(donde x_{ij} denota la observación i-ésima del regresor j: fila i, columna j).

Recordando que $\hat{e}_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik})$, solo queda encontrar los k parámetros $\hat{\beta}_j$ que satisfacen estas k condiciones.

Pero escribir y resolver este sistema con sumatorios es agotador. Necesitamos **comprimir** la notación.

El problema aritmético: demasiadas ecuaciones

En la lección 7 pudimos resolver el sistema de la regresión simple a mano porque solo teníamos dos ecuaciones ($\langle \widehat{\boldsymbol{e}}, \mathbf{1} \rangle_s = 0$ y $\langle \widehat{\boldsymbol{e}}, \boldsymbol{x} \rangle_s = 0$) y dos incógnitas ($\hat{\beta}_1, \hat{\beta}_2$). Despejamos en una, sustituimos en la otra, y llegamos al resultado.

Ahora tenemos k regresores (la constante y las $k-1$ variables). Para que el residuo sea perpendicular a todo el subespacio, tiene que ser perpendicular a cada uno de sus ejes generadores. Esto nos arroja un sistema de k ecuaciones con k incógnitas ($\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$).

Si intentamos sustituir $\hat{e}_i$ en cada ecuación y desarrollar los sumatorios, nos encontraremos con expresiones interminables. Despejar a mano un sistema de, por ejemplo, 5 regresores usando el símbolo $\sum$ es una pesadilla de subíndices cruzados.

No hay ningún concepto nuevo que descubrir aquí, solo un problema de notación: nuestra forma de escribir las cosas no está a la altura de la dimensión del problema. Para solucionarlo, vamos a introducir una herramienta matemática diseñada específicamente para agrupar vectores y hacer múltiples productos escalares de una sola vez: las **matrices**. Pero lo haremos usando una notación especial muy intuitiva.

4 La matriz de datos y el operador selector

En lugar de tener k vectores sueltos, los agrupamos como *columnas* en una matriz de datos $\mathbf{X}$. Las matrices se escriben en palo seco negrita ($\mathbf{X}$).

$$\mathbf{X} \;=\; [\mathbf{1};\; \boldsymbol{x}_2;\; \dots\; ; \boldsymbol{x}_k]$$

Para "navegar" por la matriz usaremos un *operador selector* (la barra vertical $|$):

- Actuando por la *derecha*, selecciona la *columna* j:

$$\mathbf{X}_{|j} \quad \text{es el } j\text{-ésimo regresor.}$$

- Actuando por la *izquierda*, selecciona la *fila* i:

$$_{i|}\mathbf{X} \quad \text{son los datos del } i\text{-ésimo individuo.}$$

El selector sobre un ***vector****:* siempre extrae su componente i-ésima:

$$\boldsymbol{v}_{|i} \;=\; {}_{i|}\boldsymbol{v} \;=\; v_i$$

La matriz de datos y el operador selector

Para evitar hundirnos en un mar de sumatorios, vamos a agrupar todos nuestros regresores. Una matriz no es más que una tabla, una lista de vectores-columna del mismo tamaño puestos uno al lado del otro. Si agrupamos nuestra constante y todos nuestros regresores, construimos la matriz de diseño o matriz de datos $\mathbf{X}$. Las matrices las denotaremos siempre con letras mayúsculas de palo seco en negrita, para distinguirlas visualmente de los vectores.

Para operar con esta matriz de una forma que sea perfectamente coherente con todo lo que ya sabemos de vectores, vamos a introducir una notación especial basada en un *operador selector* (la barra vertical), que actúa como un índice inteligente. Esta notación no es la habitual en los manuales clásicos anglosajones (pero es la que se utiliza en el *Curso de Álgebra Lineal* que nos sirvió de referencia en las primeras lecciones del curso).

La mecánica es muy visual:

- Si el selector actúa sobre la matriz $\mathbf{X}$ por la *derecha* ($\mathbf{X}_{|j}$), la "corta" verticalmente. Nos devuelve la columna j-ésima (es decir, el vector que contiene todas las observaciones del regresor j).
- Si el selector actúa por la *izquierda* ($_{i|}\mathbf{X}$), la "corta" horizontalmente. Nos devuelve la fila i-ésima (es decir, el conjunto de características del individuo i).

¿Qué pasa si aplicamos este selector a un vector, que solo tiene una dimensión? La regla es muy simple: el selector extrae la componente i-ésima del vector, y *da igual por qué lado se aplique.* Tanto $\boldsymbol{v}_{|i}$ como ${}_{i|}\boldsymbol{v}$ significan lo mismo: el número v_i.

Adviértase que esta barra vertical no tiene nada que ver con la barra de condicionamiento que empleamos en la lección 9 (como en $E[Y \mid X]$): allí separaba una variable aleatoria de aquello sobre lo que se condiciona; aquí es, sencillamente, un subíndice que indica qué fila o columna extraer de una matriz (o qué componente de un vector). Son dos usos tipográficamente parecidos pero conceptualmente ajenos; el contexto —esperanza condicional en un caso, matriz o vector en el otro— no deja lugar a confusión real.

Conviene cerrar esta sección con una precisión que no cambia nada de lo dicho, pero que conviene señalar cuanto antes, porque reaparecerá al hablar de variables dummy: lo verdaderamente necesario para que el ajuste sea válido no es que alguna columna de $\mathbf{X}$ sea literalmente el vector constante $\mathbf{1}$, sino que $\mathbf{1}$ pertenezca al subespacio generado por las columnas de $\mathbf{X}$ (es decir, que exista alguna combinación lineal de los regresores que reconstruya $\mathbf{1}$). En este curso, hasta ahora, ambas cosas coinciden porque hemos incluido $\mathbf{1}$ como columna explícita; pero no siempre es así: más adelante veremos un modelo con dos variables indicadoras (por ejemplo, sexo) que, sumadas, dan $\mathbf{1}$, sin que ninguna de las dos columnas sea, por separado, constante.

Para profundizar:

- Bujosa, M. *Curso de Álgebra Lineal*, lecciones 1 a 3. Definición del operador selector y sus reglas. libro online.

5 Vector por Matriz y las Condiciones de Ortogonalidad

Definiremos la operación *vector* × *matriz* ($\boldsymbol{v}\mathbf{X}$) de forma que el resultado sea un nuevo vector, cuya componente j-ésima sea el *producto escalar* entre $\boldsymbol{v}$ y la columna j de la matriz:

$$(\boldsymbol{v}\mathbf{X})_{|j} \;=\; \boldsymbol{v}\cdot\mathbf{X}_{|j}\,; \quad \text{es decir} \quad \langle \boldsymbol{v},\, \mathbf{X}_{|j}\rangle_e$$

Aplicando esto a nuestro problema, la enorme lista de k productos escalares $\widehat{\boldsymbol{e}}\cdot\mathbf{X}_{|j} = 0$ (euclídeos: como $\widehat{\boldsymbol{e}}\cdot\boldsymbol{x}_j = n\langle\widehat{\boldsymbol{e}},\boldsymbol{x}_j\rangle_s$, valen 0 igualmente) se comprime en una única y elegante ecuación:

$$\boxed{\widehat{\boldsymbol{e}}\mathbf{X} \;=\; \mathbf{0}}$$

(donde $\mathbf{0}$ es el vector formado por k ceros), pues

$$\widehat{\boldsymbol{e}}\mathbf{X} \;=\; \widehat{\boldsymbol{e}}\,[\mathbf{1};\, \boldsymbol{x}_2;\, \ldots\, ;\boldsymbol{x}_k] \;=\; (\widehat{\boldsymbol{e}}\cdot\mathbf{1},\, \widehat{\boldsymbol{e}}\cdot\boldsymbol{x}_2, \ldots\, ,\widehat{\boldsymbol{e}}\cdot\boldsymbol{x}_k) \;=\; (0,\, 0, \ldots,\, 0).$$

Simplificación de la notación: Como $(\boldsymbol{v}\mathbf{X})_{|j} = \boldsymbol{v}\cdot(\mathbf{X}_{|j})$, podemos *prescindir de los paréntesis y el punto* y escribir sencillamente $\boldsymbol{v}\mathbf{X}_{|j}$ para denotar ambas operaciones. Lo re-interpretamos como: "el selector sobre la matriz tiene precedencia sobre el producto".

Vector por Matriz y las Condiciones de Ortogonalidad

Si recordamos la transparencia 3, nuestro problema aritmético consistía en que teníamos k ecuaciones de la forma $\langle\widehat{\boldsymbol{e}},\, \boldsymbol{x}_j\rangle_e = \widehat{\boldsymbol{e}}\cdot\boldsymbol{x}_j = 0$. Necesitamos una operación que "empaquete" todos esos productos escalares de una sola vez.[1]

Esa operación es precisamente el producto de un vector por una matriz. Lo definimos geométricamente: $\boldsymbol{v}\mathbf{X}$ da como resultado un vector, y su componente j-ésima se calcula haciendo el producto escalar del vector $\boldsymbol{v}$ contra la columna j-ésima de la matriz $\mathbf{X}$.

Gracias a esta operación, podemos agrupar la extenuante lista de sumatorios en la expresión más limpia de toda la econometría: $\widehat{\boldsymbol{e}}\mathbf{X} = \mathbf{0}$. Esta ecuación nos dice, de un solo golpe de vista, que el vector de residuos es ortogonal a cada una de las columnas (regresores) de la matriz de diseño.

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice E.1-E.2: la notación matricial de las ecuaciones normales en regresión múltiple.
- Bujosa, M. *Curso de Álgebra Lineal*, cap./lección 2. Combinaciones lineales. libro online.

[1] Fíjese que en esta y en las siguientes transparencias de notación matricial trabajamos con el producto euclídeo (también conocido como producto punto) $\langle \_,\, \_\rangle_e = \_\cdot\_$; recuerde que esto no cambia ninguna condición de ortogonalidad.

6 Matriz por Vector (El Ajuste) y la Doble Lectura

Por otro lado, listamos los parámetros a estimar en un vector: $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$.

Definiremos la operación *matriz* $\times$ *vector* ($\mathbf{X}\boldsymbol{a}$) como la *combinación lineal de las columnas* de la matriz $\mathbf{X}$, usando los elementos del vector $\boldsymbol{a}$ como ponderadores:

$$\mathbf{X}\widehat{\boldsymbol{\beta}} \;=\; \hat{\beta}_1 \mathbf{X}_{|1} + \hat{\beta}_2 \mathbf{X}_{|2} + \cdots + \hat{\beta}_k \mathbf{X}_{|k}$$

Primera lectura

Esta combinación lineal de las columnas es la definición que dimos de nuestro vector de ajuste $\widehat{\boldsymbol{y}}$ en la transparencia 1:

$$\boxed{\widehat{\boldsymbol{y}} \;=\; \mathbf{X}\widehat{\boldsymbol{\beta}}}$$

Así, la descomposición $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}$ se expresa en una sola línea del siguiente modo:

$$\boxed{\boldsymbol{y} \;=\; \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}} \qquad \Longleftrightarrow \qquad \widehat{\boldsymbol{e}} \;=\; \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

Segunda lectura

Y ¿cuál es el valor ajustado para el individuo i?

Recordando que el operador por la izquierda selecciona la i-ésima componente del ajuste (y la precedencia del selector):

$$_{i|}\widehat{\boldsymbol{y}} \;=\; {}_{i|}\mathbf{X}\widehat{\boldsymbol{\beta}} \;=\; {}_{i|}\mathbf{X} \cdot \widehat{\boldsymbol{\beta}}$$

El ajuste para un individuo es el producto punto de sus datos (la fila $_{i|}\mathbf{X}$*) por los parámetros ajustados* $\widehat{\boldsymbol{\beta}}$*.*

Matriz por Vector (El Ajuste) y la Doble Lectura

Ya hemos visto qué pasa cuando multiplicamos por la izquierda (vector por matriz). Ahora vamos a definir qué ocurre cuando multiplicamos por la derecha (matriz por vector).

Agrupamos todos nuestros coeficientes $\hat{\beta}_j$ en un único vector de parámetros $\widehat{\boldsymbol{\beta}}$.

Definiremos $\mathbf{X}\widehat{\boldsymbol{\beta}}$ de la manera más natural para nuestro marco de referencia: como una *combinación lineal* de las columnas. Es decir, tomamos cada columna de la matriz $\mathbf{X}$, la multiplicamos por su correspondiente coeficiente en el vector $\widehat{\boldsymbol{\beta}}$, y sumamos todo.

Si nos fijamos bien, esta suma ponderada de las columnas de $\mathbf{X}$ es literalmente la ecuación que define nuestro ajuste $\widehat{\boldsymbol{y}}$. El vector $\widehat{\boldsymbol{y}}$ pertenece al subespacio generado por las columnas de $\mathbf{X}$, y $\widehat{\boldsymbol{\beta}}$ nos da las "coordenadas" de $\widehat{\boldsymbol{y}}$ en ese subespacio (nos dice cuánto avanzar o retroceder en la dirección de cada regresor para llegar al punto $\widehat{\boldsymbol{y}}$).

A partir de ahora, cuando veamos la expresión $\mathbf{X}\widehat{\boldsymbol{\beta}}$, debemos leerla mentalmente como "la combinación lineal de los regresores" que queda más cerca del vector $\boldsymbol{y}$ (del regresando).

Merece la pena detenerse en esta pieza, porque es, literalmente, la ecuación que nos va a acompañar el resto del curso. Recordemos que, desde la lección 6, veníamos escribiendo $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}$ (forma que la lección 7 hizo explícita en su primera transparencia). Con la notación matricial recién definida, esa misma ecuación se escribe, sin ningún ingrediente nuevo, como $\boldsymbol{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}$, de donde también $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$. Esta última expresión es la que sustituiremos dentro de la condición de ortogonalidad más adelante en esta misma lección, al derivar las ecuaciones normales.

Ahora sabemos que $\mathbf{X}\widehat{\boldsymbol{\beta}}$ es el vector con todos los valores ajustados $\widehat{\boldsymbol{y}}$. Para seleccionar el ajuste correspondiente a un individuo i aplicaremos el operador selector por la izquierda: gracias a las reglas de nuestro operador selector[2], ${}_{i|}\mathbf{X}\widehat{\boldsymbol{\beta}}$ se evalúa como el producto escalar de la fila i-ésima de la matriz $\mathbf{X}$ por el vector $\widehat{\boldsymbol{\beta}}$, es decir ${}_{i|}\mathbf{X} \cdot \widehat{\boldsymbol{\beta}}$. Conviene comprobar, aunque sea una sola vez, que esto no es un convenio sino un hecho: la componente i de la combinación lineal de columnas es

$$ {}_{i|}\Big(\hat{\beta}_1 \mathbf{X}_{|1} + \cdots + \hat{\beta}_k \mathbf{X}_{|k}\Big) \;=\; \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \;=\; {}_{i|}\mathbf{X} \cdot \widehat{\boldsymbol{\beta}}, $$

porque tomar la componente i de una suma de vectores es sumar las componentes i de cada uno, y la componente i de la columna j es precisamente el dato x_{ij} de la fila i. Por eso el convenio de quitar los paréntesis no es ambiguo. Esto arroja una **doble lectura** del producto *matriz por vector*: en su conjunto es una combinación lineal de columnas pero, "por dentro" (componente a componente) es un listado de productos punto de cada fila por el vector.

Esta dualidad no es una novedad aislada de la regresión múltiple: ya la vimos, en su versión más simple, en la lección 4 (transparencia "La media: dos caras de la misma moneda"). Allí $\mu_{\boldsymbol{y}}$ era, a la vez, el producto escalar $\langle \boldsymbol{y}, \mathbf{1} \rangle_s$ (la lectura "por fila", con un único regresor $\mathbf{1}$) y el coeficiente de la combinación lineal $\mu_{\boldsymbol{y}}\mathbf{1}$ que reconstruía la proyección $\overline{\boldsymbol{y}}$ (la lectura "por columna"). Lo que hacemos aquí es, sencillamente, generalizar esa misma dualidad —anunciada ya, de pasada, en el guiño final de la lección 5— a k regresores simultáneos.

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice E.1-E.2: la notación matricial de las ecuaciones normales en regresión múltiple.
- Bujosa, M. *Curso de Álgebra Lineal*, cap./lección 2. Combinaciones lineales. libro online.

7 Transposición

La **transposición** de una matriz ($\mathbf{X}^\top$) convierte sus filas en columnas (y viceversa). Su definición rigurosa usando el selector es:

$$ {}_{i|}\mathbf{X} \;=\; (\mathbf{X}^\top)_{|i} \quad \Big(\text{o equivalentemente} \quad \mathbf{X}_{|i} \;=\; {}_{i|}(\mathbf{X}^\top)\Big). $$

Por tanto, $\boldsymbol{v}\mathbf{X}$ y $\mathbf{X}^\top\boldsymbol{v}$ son **el mismo vector**:

$$ (\boldsymbol{v}\mathbf{X})_{|j} \;=\; \boldsymbol{v} \cdot \mathbf{X}_{|j} \;=\; \mathbf{X}_{|j} \cdot \boldsymbol{v} \;=\; {}_{j|}(\mathbf{X}^\top) \cdot \boldsymbol{v} \;=\; {}_{j|}(\mathbf{X}^\top\boldsymbol{v}) $$

[2] La misma regla de precedencia sobre el producto fijada para $\boldsymbol{v}\boldsymbol{X}$ por la derecha se aplica, simétricamente, cuando el selector actúa por la izquierda.

Así, la condición de ortogonalidad $\widehat{\boldsymbol{e}}\mathbf{X} = \mathbf{0}$ puede escribirse de forma equivalente como:

$$\boxed{\mathbf{X}^\top \widehat{\boldsymbol{e}} \;=\; \mathbf{0}.}$$

Transposición

Para poder conectar las operaciones por la derecha y por la izquierda, introducimos la **transposición de matrices**.[3] La definición es que la fila i de $\mathbf{X}$ es la columna i de su transpuesta $\mathbf{X}^\top$. En lenguaje de selectores: ${}_{i|}\mathbf{X} = (\mathbf{X}^\top)_{|i}$.

Si cruzamos esta definición con lo que sabemos de los productos, llegamos a una identidad vital: multiplicar el vector $\boldsymbol{v}$ por la matriz $\mathbf{X}$ da el mismo resultado que multiplicar la matriz transpuesta $\mathbf{X}^\top$ por el vector $\boldsymbol{v}$. ¡Es el mismo vector!

Veámoslo: $(\mathbf{X}^\top)\boldsymbol{v}$ es la combinación lineal de las columnas de $\mathbf{X}^\top$ (las filas de $\mathbf{X}$), y también es el vector cuyas componentes son los productos punto de $\boldsymbol{v}$ con cada columna de $\mathbf{X}$ (cada fila de $\mathbf{X}^\top$); pero esa era la definición de $\boldsymbol{v}\mathbf{X}$. Por tanto $(\mathbf{X}^\top)\boldsymbol{v} = \boldsymbol{v}\mathbf{X}$.

Esto nos permite reescribir nuestra gran ecuación de ortogonalidad. Sabíamos que $\widehat{\boldsymbol{e}}\mathbf{X} = \mathbf{0}$. Gracias a esta dualidad, podemos escribirla con toda legitimidad poniendo la matriz (transpuesta) al principio: $\mathbf{X}^\top\widehat{\boldsymbol{e}} = \mathbf{0}$. Esta es la llave que abre la puerta a las Ecuaciones Normales.

Veamos un ejemplo numérico muy simple:[4]

$$\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{bmatrix} 2 & 8 \\ -1 & -4 \end{bmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}; \qquad \begin{bmatrix} 2 & -1 \\ 8 & -4 \end{bmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

De propina: ahora tenemos una nueva interpretación de la operación *vector por matriz*. Dado que $(\mathbf{X}^\top)\boldsymbol{v} = \boldsymbol{v}\mathbf{X}$ y dado que $(\mathbf{X}^\top)\boldsymbol{v}$ es la combinación lineal de las columnas de $\mathbf{X}^\top$ (que son las filas de $\mathbf{X}$), tenemos que $\boldsymbol{v}\mathbf{X}$ es la combinación lineal de las filas: la misma combinación lineal de los mismos vectores. Resumiendo, al igual que *matriz por vector* es una combinación lineal de columnas, *vector por matriz* es una combinación lineal de filas.

Para llegar a las ecuaciones normales —la forma compacta de las k condiciones de ortogonalidad de la transparencia 3— nos faltan dos ingredientes que enunciamos sin demostrar. El primero es el *producto de matrices*, $\mathbf{X}^\top\mathbf{X}$, del que solo diremos unas palabras en los apuntes de la transparencia siguiente. El segundo es que todas las operaciones de hoy (selector, matriz por vector, vector por matriz, transposición y producto de matrices) son *lineales*, y de esa linealidad usaremos exactamente dos reglas:

$$\mathbf{X}^\top(\boldsymbol{a} - \boldsymbol{b}) = \mathbf{X}^\top\boldsymbol{a} - \mathbf{X}^\top\boldsymbol{b} \qquad \text{y} \qquad \mathbf{X}^\top(\mathbf{X}\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})\widehat{\boldsymbol{\beta}}.$$

La primera dice que $\mathbf{X}^\top$ "distribuye" sobre la resta; la segunda, que da igual dónde pongamos los paréntesis.[5]

[3] A diferencia de otros textos donde se transponen vectores (como si un vector pudiera estar de pie o tumbado), aquí somos fieles a la geometría: los vectores son listas de números (listas de coordenadas), punto. Lo único que se transpone (se reordena) son las tablas, las matrices.

[4] El ejemplo respeta el convenio generalizado en los manuales de escribir los vectores que multiplican por la izquierda en horizontal (combinación de filas de la matriz) y los que multiplican por la derecha en vertical (combinación de las columnas).

[5] Para ver en detalle el producto de matrices y la demostración de linealidad para cada una de las operaciones puede consultar el manual Un Curso de Álgebra Lineal.

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice E.1-E.2: la notación matricial de las ecuaciones normales en regresión múltiple.
- Bujosa, M. *Curso de Álgebra Lineal*, lecciones 1 a 3. libro online.

8 Ecuaciones normales

Sustituyendo $\quad \widehat{\boldsymbol{e}} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \quad$ en $\quad \mathbf{X}^\top \widehat{\boldsymbol{e}} = \mathbf{0}$:

$$\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0} \implies \boxed{\mathbf{X}^\top\mathbf{X}\,\widehat{\boldsymbol{\beta}} \;=\; \mathbf{X}^\top\boldsymbol{y}.}$$

(Hemos usado que $\mathbf{X}^\top$ distribuye sobre la resta.) Aquí aparece (por primera vez en el curso) un producto de matrices: $\mathbf{X}^\top\mathbf{X}$; baste saber que es otra matriz.

La expresión del recuadro es el sistema de *ecuaciones normales*.

Cuando los k regresores son linealmente independientes (ninguno es combinación lineal de los demás), este sistema tiene una única solución, cuya expresión matricial es:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}.$$

No se preocupe por la expresión, los ordenadores saben calcularla. Lo importante es que dicha solución garantiza que $\mathbf{X}^\top\widehat{\boldsymbol{e}} = \mathbf{0}$*; es decir, que al usar* $\widehat{\boldsymbol{\beta}}$ *para calcular* $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$*, obtenemos la combinación lineal de los regresores más cercana a* $\boldsymbol{y}$*.*

Ecuaciones normales

La transparencia anterior nos dejó las k condiciones de ortogonalidad comprimidas en $\mathbf{X}^\top\widehat{\boldsymbol{e}} = \mathbf{0}$. Resolver este sistema para obtener $\widehat{\boldsymbol{\beta}}$ es puramente algebraico: sustituimos la definición del residuo, $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, y distribuimos la matriz traspuesta:

$$\mathbf{X}^\top\boldsymbol{y} - \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0} \implies \mathbf{X}^\top\mathbf{X}\,\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\boldsymbol{y}.$$

A este sistema de k ecuaciones lineales con k incógnitas se le llama, tradicionalmente, *ecuaciones normales* (el nombre viene de que expresan condiciones de ortogonalidad, es decir, de perpendicularidad o "normalidad" geométrica, no de la distribución normal de probabilidad).

Aquí aparece, por primera vez en el curso, un *producto matriz-matriz*: $\mathbf{X}^\top\mathbf{X}$, una matriz cuadrada de $k \times k$. No vamos a desarrollar ninguna teoría general sobre productos de matrices; nos basta con saber que su elemento (j, j') es un producto escalar del espacio euclídeo: el producto entre la columna j-ésima y la columna j'-ésima de $\mathbf{X}$ —es decir, entre regresores—. Por ejemplo, el elemento $(1,1)$ (fila y columna constantes $\mathbf{1}$) es $\langle \mathbf{1}, \mathbf{1}\rangle_e = n$; el elemento $(1, j)$ es $\langle \mathbf{1}, \boldsymbol{x}_j\rangle_e = n\mu_{\boldsymbol{x}_j}$ (recordando la lección 4); y el elemento (j, j') es $\langle \boldsymbol{x}_j, \boldsymbol{x}_{j'}\rangle_e$. Es decir: toda la matriz $\mathbf{X}^\top\mathbf{X}$ está construida, literalmente, con los mismos productos escalares que ya sabíamos calcular desde la lección 2; la novedad es solo organizarlos en una tabla.

Si los k regresores $\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ son linealmente independientes (si ninguno es combinación lineal de los demás), puede probarse —sin que nosotros lo hagamos aquí, pues excede el alcance

del curso— que la matriz $\mathbf{X}^\top\mathbf{X}$ es *invertible* —es decir, que existe otra matriz, denotada $(\mathbf{X}^\top\mathbf{X})^{-1}$, que *deshace* la multiplicación por $\mathbf{X}^\top\mathbf{X}$; es el análogo matricial de dividir entre un número distinto de cero—, y que el sistema de ecuaciones normales tiene una única solución (en la jerga del álgebra lineal se dice que el sistema es *compatible* *determinado*):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}.$$

Esta es la fórmula que aparece en la práctica totalidad de los manuales de econometría, y es también, de hecho, la que Gretl (y cualquier otro software estadístico) utiliza internamente para calcular una regresión múltiple. Aquí la mencionamos por completitud —conviene saber que existe, y saber reconocerla si la ve en un libro—, pero *no* vamos a usarla como herramienta de razonamiento en este curso: no vamos a calcular inversas de matrices a mano, ni vamos a apoyarnos en propiedades de la inversa para demostrar nada. El camino conceptual que nos interesa —y que reaparecerá en la lección 11 al demostrar propiedades del estimador— sigue siendo, siempre, la condición de ortogonalidad $\mathbf{X}^\top\widehat{\boldsymbol{e}} = \mathbf{0}$, no la fórmula con inversa.

Para cerrar, vale la pena señalar que la existencia de una única solución para las ecuaciones normales es lo único que falla cuando los regresores son linealmente dependientes: en tal caso $\mathbf{X}^\top\mathbf{X}$ no es invertible (*colinealidad exacta*) y el sistema de ecuaciones normales pasa a tener infinitas soluciones (es un sistema *compatible* pero *indeterminado*); en consecuencia hay infinitos vectores $\widehat{\boldsymbol{\beta}}$ que arrojan el mismo ajuste $\widehat{\boldsymbol{y}}$.[6]

La siguiente transparencia se ocupa de la situación contraria: cuando los regresores son perpendiculares entre sí.

Para profundizar:

- Wooldridge, J. M. (2020), Apéndice E: derivación matricial completa de las ecuaciones normales y condiciones de existencia de la inversa.

9 Familia ortogonal

Recordemos (lección 3, Cauchy–Schwarz) la fórmula de proyección sobre *un* generador: $\alpha = \langle \boldsymbol{a}, \boldsymbol{b}\rangle / \langle \boldsymbol{a}, \boldsymbol{a}\rangle$. Y (lección 7) la "vía alternativa" con *dos* generadores ortogonales.

Si $\{\mathbf{1}, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k\}$ son *mutuamente ortogonales* (nótese que $\mathbf{1} \perp \boldsymbol{z}_j$ significa $\mu_{\boldsymbol{z}_j} = 0$), cada coeficiente se calcula *por separado*, sin resolver ningún sistema:

$$\hat{\beta}_1 = \frac{\langle \boldsymbol{y}, \mathbf{1}\rangle_s}{\langle \mathbf{1}, \mathbf{1}\rangle_s} = \mu_{\boldsymbol{y}}, \qquad \hat{\beta}_j = \frac{\langle \boldsymbol{y}, \boldsymbol{z}_j\rangle_s}{\langle \boldsymbol{z}_j, \boldsymbol{z}_j\rangle_s} = \frac{\sigma_{\boldsymbol{y}\boldsymbol{z}_j}}{\sigma^2_{\boldsymbol{z}_j}} \quad (j = 2, \ldots, k).$$

(aplicando, una vez más, el truco del vector de media nula de la lección 6: como $\boldsymbol{z}_j$ tiene media cero, su producto escalar con $\boldsymbol{y}$ coincide con la covarianza entre ambos).

Este extraordinario caso es la excepción: con datos económicos reales, los regresores prácticamente nunca son ortogonales entre sí. El caso general exige resolver las ecuaciones normales de la transparencia anterior.

[6] En la lección 14 veremos un problema distinto: la *colinealidad de grado*, una situación en la que el sistema de ecuaciones normales tiene solución única, pero extremadamente sensible a pequeños cambios en los datos (algo que acarrea un grave problema en la interpretación de resultados). Ocurre cuando algunos regresores están casi alineados.

Familia ortogonal

Esta transparencia cobra una promesa muy antigua, sembrada en la lección 3 al demostrar Cauchy–Schwarz: allí construimos $\boldsymbol{h} = \boldsymbol{b} - \alpha\boldsymbol{a}$, con $\alpha = \langle \boldsymbol{a}, \boldsymbol{b}\rangle / \langle \boldsymbol{a}, \boldsymbol{a}\rangle$, y dijimos que ese $\alpha\boldsymbol{a}$ era la "sombra" de $\boldsymbol{b}$ sobre la dirección de $\boldsymbol{a}$ —la *proyección ortogonal*. En la lección 7 usamos esa misma fórmula, con *dos* regresores ortogonales ($\mathbf{1}$ y $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$), para obtener $\hat{\beta}_2$ sin resolver ningún sistema: la llamamos, entonces, "vía alternativa".

Generalicemos esa vía alternativa a k generadores. Supongamos que, además de $\mathbf{1}$, disponemos de $k-1$ vectores $\boldsymbol{z}_2, \ldots, \boldsymbol{z}_k$ que son *mutuamente ortogonales entre sí* y, además, ortogonales a $\mathbf{1}$ (es decir, todos ellos de media cero). En ese caso —y *solo* en ese caso—, la proyección de $\boldsymbol{y}$ sobre $\mathcal{L}(\mathbf{1}, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k)$ se descompone en k proyecciones *independientes*, cada una calculada con la misma fórmula de Cauchy–Schwarz de la lección 3, aplicada por separado a cada generador:

$$\hat{\beta}_1 = \frac{\langle \boldsymbol{y}, \mathbf{1}\rangle_s}{\langle \mathbf{1}, \mathbf{1}\rangle_s} = \mu_{\boldsymbol{y}}, \qquad \hat{\beta}_j = \frac{\langle \boldsymbol{y}, \boldsymbol{z}_j\rangle_s}{\langle \boldsymbol{z}_j, \boldsymbol{z}_j\rangle_s} = \frac{\sigma_{\boldsymbol{y}\boldsymbol{z}_j}}{\sigma^2_{\boldsymbol{z}_j}} \quad (j = 2, \ldots, k),$$

(aplicando, una vez más, el truco del vector de media nula de la lección 6: como $\boldsymbol{z}_j$ tiene media cero, su producto escalar con cualquier vector coincide con la covarianza entre ambos. Así, $\langle \boldsymbol{y}, \boldsymbol{z}_j\rangle_s = \sigma_{\boldsymbol{y}\boldsymbol{z}_j}$ y, por el mismo truco, $\langle \boldsymbol{z}_j, \boldsymbol{z}_j\rangle_s = \sigma^2_{\boldsymbol{z}_j}$, que es la covarianza consigo mismo).

Por tanto, en este caso tan especial, no hace falta resolver ningún sistema de ecuaciones normales: cada coeficiente se calcula de forma completamente aislada, exactamente como ocurría con dos generadores ortogonales que vimos en la lección 7.

Conviene cerrar esta transparencia con una advertencia importante, que prepara el terreno para la siguiente. Este escenario —regresores mutuamente ortogonales— es cómodo computacionalmente, pero es la *excepción*, no la norma. Con datos económicos reales, los regresores casi siempre están correlacionados entre sí en alguna medida (por ejemplo, la antigüedad y la experiencia de un trabajador) y no suelen tener media cero. Cuando los regresores no son mutuamente ortogonales, no hay atajo: hay que resolver el sistema completo de ecuaciones normales de la transparencia anterior.

Para profundizar:

- Strang, G. (2016), cap. 4, sección 4.4: ortogonalización de Gram-Schmidt (mención, sin desarrollar en este curso).

10 R^2 en regresión múltiple

Como $\langle \widehat{\boldsymbol{e}}, \mathbf{1}\rangle_s = 0$ sigue siendo una de las k condiciones: $\mu_{\boldsymbol{y}} = \mu_{\widehat{\boldsymbol{y}}}$ (lección 7) y STC = SEC+SRC (lección 8) *siguen valiendo igual*, sin importar k.

Pero (pregunta 5 de la práctica de regresión simple), en general: $R^2 \neq \rho^2_{\boldsymbol{x}_2\boldsymbol{y}} + \cdots + \rho^2_{\boldsymbol{x}_k\boldsymbol{y}}$.

¿Por qué? En la lección 8, $R^2 = \rho^2_{\boldsymbol{x}\boldsymbol{y}}$ porque los vectores de $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$ con media cero forman una recta: todos alineados con $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$.

Con $k > 2$ regresores ese subespacio tiene *dimensión* $k-1$: $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$ ya no está alineado, en general, con ningún $\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j}\mathbf{1}$.

Excepción (transparencia anterior): regresores en desviaciones mutuamente ortogonales $\implies$ $R^2 = \rho^2_{\boldsymbol{x}_2\boldsymbol{y}} + \cdots + \rho^2_{\boldsymbol{x}_k\boldsymbol{y}}$ (Pitágoras generalizado).

R^2 en regresión múltiple

Podemos ahora responder, con precisión geométrica, a la pregunta que quedó abierta en la práctica de regresión simple (su pregunta 5): ¿por qué el R^2 de un modelo con dos (o más) regresores no es la suma de los R^2 de los modelos simples correspondientes?

Primero, lo que *sí* se conserva sin cambio alguno. La condición $\langle \widehat{\boldsymbol{e}}, \mathbf{1}\rangle_s = 0$ sigue siendo una de nuestras k condiciones de ortogonalidad, igual que en las lecciones 6 y 7 (allí era una de dos; aquí, una de k). De ella se deduce, sin ningún cambio en el argumento, que $\mu_{\boldsymbol{y}} = \mu_{\widehat{\boldsymbol{y}}}$ (lección 7). Y como $\widehat{\boldsymbol{y}} \perp \widehat{\boldsymbol{e}}$ sigue siendo cierto (el residuo es ortogonal a *todo* el subespacio, en particular al propio ajuste, que pertenece a él), el mismo triángulo rectángulo de la lección 8 —hipotenusa $\boldsymbol{y} - \overline{\boldsymbol{y}}$, catetos $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$ y $\widehat{\boldsymbol{e}}$— sigue siendo válido, sea cual sea k. Por tanto, STC = SEC + SRC y $R^2 = \text{SEC}/\text{STC} = \cos^2\theta$ (con θ el ángulo entre $\boldsymbol{y} - \overline{\boldsymbol{y}}$ y $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$) se mantienen, sin necesidad de ninguna demostración nueva: son consecuencia, otra vez, de las mismas dos piezas (ortogonalidad y Pitágoras), que no dependen del número de regresores.

Lo que *deja de valer*, en cambio, es la identidad exclusiva de la regresión simple, $R^2 = \rho^2_{\boldsymbol{x}\boldsymbol{y}}$. Recordemos por qué era cierta allí: en la lección 8 demostramos que, en $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$, *todo* vector en desviaciones cae en la misma recta $\mathcal{L}(\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1})$ —porque ese subespacio tiene dimensión 1—. Eso obligaba a $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$ a estar alineado con $\boldsymbol{x} - \mu_{\boldsymbol{x}}\mathbf{1}$, y de ahí salía $\rho_{\boldsymbol{y}\widehat{\boldsymbol{y}}} = |\rho_{\boldsymbol{x}\boldsymbol{y}}|$.

Con $k > 2$ regresores, el subespacio de vectores de $\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$ que están en desviaciones tiene, en general, dimensión $k-1$ en lugar de dimensión 1. El vector $\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}}$ es, ahora, una combinación de los $k-1$ vectores en desviaciones $\boldsymbol{x}_2 - \mu_{\boldsymbol{x}_2}\mathbf{1}, \ldots, \boldsymbol{x}_k - \mu_{\boldsymbol{x}_k}\mathbf{1}$, y ya no tiene por qué estar alineado con ninguno de ellos en particular. Por eso R^2, aunque sigue siendo $\cos^2\theta$, ya no coincide, en general, con el cuadrado de ninguna correlación simple $\rho_{\boldsymbol{x}_j\boldsymbol{y}}$ —y tampoco con la suma de esos cuadrados.

Hay, sin embargo, una excepción donde R^2 es la suma de las correlaciones al cuadrado, y es precisamente el caso especial de la transparencia anterior. Conviene tender bien el puente con aquella transparencia, porque allí los generadores $\boldsymbol{z}_j$ ya tenían media cero y aquí los regresores $\boldsymbol{x}_j$ no la tienen. El puente es el mismo que usamos en la "vía alternativa" de la lección 7. Sustituir cada $\boldsymbol{x}_j$ por su versión centrada $\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j}\mathbf{1}$ no cambia el subespacio $\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, porque cada vector centrado es combinación lineal de $\mathbf{1}$ y del original, y viceversa. Por tanto no cambia el ajuste $\widehat{\boldsymbol{y}}$ ni los coeficientes $\hat{\beta}_j$ con $j \geq 2$; solo cambia el de $\mathbf{1}$, que pasa a ser $\mu_{\boldsymbol{y}}$. Así que los $\boldsymbol{z}_j$ de la transparencia anterior son estos vectores centrados, y la parte de $\widehat{\boldsymbol{y}}$ en desviaciones es

$$\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}} = \hat{\beta}_2(\boldsymbol{x}_2 - \mu_{\boldsymbol{x}_2}\mathbf{1}) + \cdots + \hat{\beta}_k(\boldsymbol{x}_k - \mu_{\boldsymbol{x}_k}\mathbf{1}).$$

Con esto a la vista: cuando los regresores en desviaciones $\boldsymbol{x}_2 - \mu_{\boldsymbol{x}_2}\mathbf{1}, \ldots, \boldsymbol{x}_k - \mu_{\boldsymbol{x}_k}\mathbf{1}$ son *mutuamente ortogonales*, el Pitágoras generalizado[7] nos da

$$\sigma^2_{\boldsymbol{y}} = \sigma^2_{\widehat{\boldsymbol{y}}} + \sigma^2_{\widehat{\boldsymbol{e}}}, \qquad \sigma^2_{\widehat{\boldsymbol{y}}} = \|\hat{\beta}_2(\boldsymbol{x}_2 - \mu_{\boldsymbol{x}_2}\mathbf{1})\|^2_s + \cdots + \|\hat{\beta}_k(\boldsymbol{x}_k - \mu_{\boldsymbol{x}_k}\mathbf{1})\|^2_s.$$

[7] El teorema de Pitágoras (lección 3) se enunció para *dos* vectores ortogonales, pero se extiende, sin ninguna dificultad adicional, a $k+1$ vectores mutuamente ortogonales: si $\mathbf{1}, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k, \widehat{\boldsymbol{e}}$ son mutuamente ortogonales, e $\boldsymbol{y} = \hat{\beta}_1\mathbf{1} + \hat{\beta}_2\boldsymbol{v}_2 + \cdots + \hat{\beta}_k\boldsymbol{v}_k + \widehat{\boldsymbol{e}}$, entonces: $\|\boldsymbol{y}\|^2_s = \mu^2_{\boldsymbol{y}} + \|\hat{\beta}_2\boldsymbol{v}_2\|^2_s + \cdots + \|\hat{\beta}_k\boldsymbol{v}_k\|^2_s + \|\widehat{\boldsymbol{e}}\|^2_s$.

Recordemos ahora que R^2, por definición (lección 8), es precisamente el cociente $\sigma^2_{\widehat{\boldsymbol{y}}}/\sigma^2_{\boldsymbol{y}}$. Dividiendo entre $\sigma^2_{\boldsymbol{y}}$ la segunda identidad de arriba —término a término—, obtenemos

$$R^2 = \frac{\sigma^2_{\widehat{\boldsymbol{y}}}}{\sigma^2_{\boldsymbol{y}}} = \frac{\|\hat{\beta}_2(\boldsymbol{x}_2 - \mu_{\boldsymbol{x}_2}\mathbf{1})\|^2_s}{\sigma^2_{\boldsymbol{y}}} + \cdots + \frac{\|\hat{\beta}_k(\boldsymbol{x}_k - \mu_{\boldsymbol{x}_k}\mathbf{1})\|^2_s}{\sigma^2_{\boldsymbol{y}}}.$$

Es decir: en este caso especial, R^2 queda escrito como una suma de $k-1$ cocientes, uno por cada dirección ortogonal del subespacio centrado. Falta solo un paso: identificar qué es cada uno de esos cocientes.

Veamos esto con detalle para un j cualquiera entre 2 y k. Por la propia definición de varianza como norma al cuadrado de un vector en desviaciones (lección 5), aplicada aquí al vector $\hat{\beta}_j\boldsymbol{x}_j$ en lugar de a $\boldsymbol{x}_j$ directamente —y usando que el vector en desviaciones de $\hat{\beta}_j\boldsymbol{x}_j$ es precisamente $\hat{\beta}_j(\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j}\mathbf{1})$, pues multiplicar por la constante $\hat{\beta}_j$ no hace sino reescalar el vector en desviaciones de $\boldsymbol{x}_j$ (lección 5, propiedad de escala)—, tenemos

$$\|\hat{\beta}_j(\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j}\mathbf{1})\|^2_s = \hat{\beta}_j^2\,\|\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j}\mathbf{1}\|^2_s = \hat{\beta}_j^2\,\sigma^2_{\boldsymbol{x}_j}.$$

Sustituyendo la fórmula de $\hat{\beta}_j$ que acabamos de obtener en "Familia ortogonal" ($\hat{\beta}_j = \sigma_{\boldsymbol{x}_j\boldsymbol{y}}/\sigma^2_{\boldsymbol{x}_j}$):

$$\hat{\beta}_j^2\,\sigma^2_{\boldsymbol{x}_j} = \left(\frac{\sigma_{\boldsymbol{x}_j\boldsymbol{y}}}{\sigma^2_{\boldsymbol{x}_j}}\right)^2 \sigma^2_{\boldsymbol{x}_j} = \frac{\sigma^2_{\boldsymbol{x}_j\boldsymbol{y}}}{\sigma^2_{\boldsymbol{x}_j}}.$$

Y dividiendo entre $\sigma^2_{\boldsymbol{y}}$:

$$\frac{\|\hat{\beta}_j(\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j}\mathbf{1})\|^2_s}{\sigma^2_{\boldsymbol{y}}} = \frac{\sigma^2_{\boldsymbol{x}_j\boldsymbol{y}}}{\sigma^2_{\boldsymbol{x}_j}\,\sigma^2_{\boldsymbol{y}}} = \left(\frac{\sigma_{\boldsymbol{x}_j\boldsymbol{y}}}{\sigma_{\boldsymbol{x}_j}\sigma_{\boldsymbol{y}}}\right)^2 = \rho^2_{\boldsymbol{x}_j\boldsymbol{y}},$$

que es la definición de correlación al cuadrado (lección 5). No hay ningún ingrediente nuevo: solo la fórmula de $\hat{\beta}_j$ obtenida hace un momento en esta misma sesión y la definición de correlación de la lección 5, aplicadas —por separado, gracias a la ortogonalidad mutua— a cada dirección $j = 2, \ldots, k$.

Sustituyendo este resultado, término a término, en la suma de cocientes que habíamos obtenido para R^2, concluimos que, en este caso muy particular, y solo en él,

$$R^2 = \rho^2_{\boldsymbol{x}_2\boldsymbol{y}} + \cdots + \rho^2_{\boldsymbol{x}_k\boldsymbol{y}}.$$

Pero, como ya señalamos, la ortogonalidad mutua entre regresores económicos reales es la excepción, no la norma —lo comprobaremos en la práctica que sigue a esta lección, con `rooms` y `nox`.

Para profundizar:

- Wooldridge, J. M. (2020), cap. 3, sección 3.2: por qué el R^2 de la regresión múltiple no se descompone, en general, en las correlaciones simples con cada regresor.

11 El modelo poblacional también se generaliza

Recordemos (lección 9): $Y = \beta_1\, 1 + \beta_2 X + U$, con tres supuestos (linealidad, $\mathrm{Var}(X) \neq 0$ y homocedasticidad) —más la exogeneidad estricta $E[U \mid X] = 0$, que allí salía *gratis* por la definición de U—.

Lo generalizamos a $k-1$ regresores no constantes $X_2, \ldots, X_k$:

$$Y = \beta_1\, 1 + \beta_2 X_2 + \cdots + \beta_k X_k + U.$$

Los mismos tres supuestos, condicionando ahora sobre todos los regresores a la vez:

1. $E[Y \mid X_2, \ldots, X_k] = \beta_1\, 1 + \beta_2 X_2 + \cdots + \beta_k X_k$
2. $1, X_2, \ldots, X_k$ *linealmente independientes* — generaliza $\mathrm{Var}(X) \neq 0$.
3. $Var[U \mid X_2, \ldots, X_k] = \sigma^2\, 1$

Y el mismo corolario gratuito: $E[U \mid X_2, \ldots, X_k] = 0$.

Y el método de los momentos se extiende tal cual: sustituir momentos poblacionales por muestrales reproduce las ecuaciones normales $\mathbf{X}^\top \mathbf{X} \widehat{\boldsymbol{\beta}} = \mathbf{X}^\top \boldsymbol{y}$.

Condicionamos sobre una lista de variables más larga. Pero la estructura no cambia.

El modelo poblacional también se generaliza

Toda esta sesión ha vivido en $\mathbb{R}^n$: hemos generalizado el plano de la regresión simple a un subespacio de k dimensiones, sin necesidad de hablar una sola vez de variables aleatorias. Pero la lección 9 ("El puente") nos había dado, además de las fórmulas muestrales, un modelo *poblacional* —$Y = \beta_1\, 1 + \beta_2 X + U$, con sus tres supuestos— del que aquellas fórmulas eran, según vimos, estimadores por el método de los momentos. Conviene, antes de cerrar la sesión, comprobar que ese modelo poblacional se generaliza igual, para no dejar la impresión de que el puente tendido en la lección 9 solo servía para un único regresor.

La generalización no exige ninguna idea nueva. Con $k-1$ regresores no constantes $X_2, \ldots, X_k$ (además de la constante 1), el modelo poblacional es

$$Y = \beta_1\, 1 + \beta_2 X_2 + \cdots + \beta_k X_k + U,$$

y los tres supuestos de la lección 9 se trasladan sin más cambio que alargar la lista de variables sobre las que condicionamos: linealidad de $E[Y \mid X_2, \ldots, X_k]$; independencia lineal de $1, X_2, \ldots, X_k$ —el supuesto que ahora nos interesa—; y homocedasticidad $Var[U \mid X_2, \ldots, X_k] = \sigma^2\, 1$. Y con ellos viaja también, intacto, el corolario gratuito de la lección 9: como U sigue definiéndose como lo que queda al proyectar Y sobre *todas* las funciones de los regresores, la exogeneidad estricta $E[U \mid X_2, \ldots, X_k] = 0$ se cumple por construcción, sin necesidad de imponerla.

En la lección 9 el supuesto 2 era, con un único regresor, una condición modesta: $\mathrm{Var}(X) \neq 0$, con la precisión, ya hecha allí, de que esto significa "X no es igual, con probabilidad uno, a ningún valor fijo". Con $k-1 \geq 2$ regresores, la condición deja de reducirse a la varianza de una única variable: dos regresores pueden tener, cada uno, varianza no nula por separado y, aun así, ser linealmente dependientes entre sí —algo distinto, y más fuerte, que la mera correlación de la transparencia

"Familia ortogonal"—. Ejemplo sencillo, sin ninguna salvedad necesaria: la temperatura de un mismo día medida en grados Celsius y en grados Fahrenheit son dos variables no constantes por separado, pero exactamente dependientes una de otra ($F = 32 + 1{,}8\,C$): es justo lo que en esta sesión llamamos "colinealidad exacta" al comentar $\mathbf{X}^\top\mathbf{X}$.

Por último, el método de los momentos —el "salto" de la lección 9— se extiende igual: sustituir cada momento poblacional desconocido por su análogo muestral reproduce el sistema de ecuaciones normales $\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\boldsymbol{y}$ obtenido hoy por vía puramente geométrica.

Para profundizar:

- Wooldridge, J. M. (2020), cap. 3, sección 3.3: supuestos del modelo de regresión múltiple (MLR.1 a MLR.5 en su numeración; recuerde el apéndice "Tres supuestos aquí, cinco en Wooldridge" de la lección 9), en particular el supuesto de no colinealidad perfecta, que es nuestro supuesto 2.

12 Recapitulación y guiño a las sesiones siguientes

Regresión simple	Regresión múltiple
$\boldsymbol{y} = \hat\beta_1\mathbf{1} + \hat\beta_2\boldsymbol{x} + \widehat{\boldsymbol{e}}$	$\boldsymbol{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}$
$\mathcal{L}(\mathbf{1}, \boldsymbol{x})$, un plano (dim. 2)	$\mathcal{L}(\mathbf{1}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, (dimensión k)
2 condiciones de ortogonalidad	k condiciones, comprimidas en $\mathbf{X}^\top\widehat{\boldsymbol{e}} = \mathbf{0}$
$\hat\beta_1, \hat\beta_2$ (escalares)	$\widehat{\boldsymbol{\beta}}$ (vector), resolviendo $\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\boldsymbol{y}$
$R^2 = \rho^2_{\boldsymbol{xy}}$ (caso particular)	$R^2 = \rho^2_{\boldsymbol{y}\widehat{\boldsymbol{y}}}$ (siempre); $= \sum_j \rho^2_{\boldsymbol{x}_j\boldsymbol{y}}$ solo si ortogonales

Como caso límite, obsérvese que $k = 1$ (ningún regresor no constante, solo la columna de unos $\mathbf{1}$) recupera exactamente la proyección de las lecciones 4 y 5.[8]

Con el modelo poblacional generalizado a k regresores, aún queda pendiente la misma pregunta: ¿es $\widehat{\boldsymbol{\beta}}$ un buen estimador del verdadero $\boldsymbol{\beta}$ poblacional? En la lección 11 retomaremos la cuestión.

Respecto a la interpretación de parámetros: cada $\hat\beta_j$ mide un *efecto parcial* —"manteniendo constantes los demás regresores" (*ceteris paribus*)—. En la práctica que sigue a esta lección lo veremos con datos reales. Además hablaremos del R^2 ajustado.

Recapitulación y guiño a las sesiones siguientes

El recorrido de esta sesión ha sido, en su estructura, una repetición deliberada de las lecciones 6 y 7, ahora con k regresores en lugar de dos.[9] Ninguna idea matemática nueva ha sido necesaria: la proyección ortogonal, las condiciones de ortogonalidad del residuo, Pitágoras. Lo único genuinamente nuevo ha sido el *lenguaje* —la notación matricial— que nos permite escribir de forma compacta un problema que, con sumatorios, requeriría k ecuaciones escritas una a una.

[8] En ese caso, la única condición de ortogonalidad $\mathbf{X}^\top\widehat{\boldsymbol{e}} = \mathbf{0}$ se reduce a $\langle\widehat{\boldsymbol{e}}, \mathbf{1}\rangle_s = 0$, y las ecuaciones normales colapsan en la única ecuación escalar $\hat\beta_1 = \mu_{\boldsymbol{y}}$: la media aritmética. La regresión múltiple no sustituye aquella construcción; la contiene como caso particular.

[9] Si lo piensa, también generaliza las lecciones 4 y 5: $\mathcal{L}(\mathbf{1})$ una recta (dimensión 1), una condición de ortogonalidad, un parámetro ($\mu_{\boldsymbol{y}}$) y, dado que el vector de medias tiene varianza cero, resulta el caso trivial $R^2 = 0$ que no fue comentado ahí.

La tabla resume el paralelismo completo entre ambos casos. Merece la pena detenerse en la última fila, que es la que responde a la pregunta que quedó abierta en la práctica de regresión simple: en regresión simple, R^2 *siempre* coincide con $\rho^2_{\boldsymbol{xy}}$, porque el subespacio de vectores en desviaciones que están en el plano sobre el que se proyecta queda reducido a una recta; en regresión múltiple, R^2 sigue siendo $\cos^2\theta$ —la definición geométrica no cambia—, pero deja de coincidir, en general, con ninguna suma del cuadrado de correlaciones simples, salvo en el caso especial (y poco frecuente con datos reales) de regresores mutuamente ortogonales.

Aún queda una pregunta por responder. La transparencia anterior mostró que el modelo poblacional de la lección 9 —con sus tres supuestos— se generaliza a k regresores sin ningún ingrediente conceptual nuevo. Con ese modelo ya generalizado, podemos formular con precisión la pregunta que quedaba pendiente: ¿es $\widehat{\boldsymbol{\beta}}$ un buen estimador del verdadero $\boldsymbol{\beta}$ poblacional?

La advertencia sobre la interpretación de los coeficientes merece subrayarse, porque es la primera vez en el curso que disponemos del aparato necesario para plantearla con precisión. Hasta ahora, con un único regresor, "el efecto de $\boldsymbol{x}$ sobre $\boldsymbol{y}$" y "el coeficiente $\hat{\beta}_2$" eran, sin más matices, la misma cosa. Con varios regresores, $\hat{\beta}_j$ mide algo más sutil: el *efecto parcial*, es decir, cuánto cambia el ajuste $\widehat{\boldsymbol{y}}$ ante un cambio unitario en $\boldsymbol{x}_j$, *manteniendo fijos los demás regresores del modelo* —de ahí la expresión latina *ceteris paribus* ("permaneciendo lo demás igual")—.

Cerramos con dos promesas explícitas para las próximas sesiones. La primera, de corto plazo: las prácticas de laboratorio que siguen a esta lección aplicarán todo lo de hoy con datos reales, calcularán el R^2 ajustado —que compensa el hecho de que el R^2 ordinario nunca disminuye al añadir regresores (advertencia ya sembrada en la lección 8)— y compararán coeficientes parciales frente a los coeficientes de las regresiones simples correspondientes. La segunda, de plazo algo mayor: en la lección 11 volveremos sobre la insesgadez y varianza de $\widehat{\boldsymbol{\beta}}$ con todo rigor, pero lo demostraremos solo para el caso más sencillo del modelo de regresión lineal *simple* (constante + regresor no constante) sin necesidad de recurrir a ninguna matriz.

Para profundizar:

- Wooldridge, J. M. (2020), cap. 3, sección 3.4: interpretación *ceteris paribus* de los coeficientes de la regresión múltiple, con ejemplos económicos.