TEMA 2: REGRESIÓN LINEAL

Los modelos de regresión tienen por objetivo generalizar los patrones de relación de dos o más variables **numéricas,** observados en una muestra, mediante la formulación de un modelo matemático que capte cuantitativamente los aspectos esenciales (esto es, no ruidosos) de esa relación entre las variables. Esta **generalización de la información recolectada mediante la experiencia**, el modelo de regresión, facilita entonces dos labores:

i) **Explicar** cómo los valores observados de una variable *Y* dependen de los valores de otra variable *X,* o de más variables, a través de una relación funcional del tipo $Y = f(X)$ (por este motivo, a la variable *Y* se la llama *variable dependiente*, y a la variable *X variable independiente o explicativa*); y

ii) **Predecir** el valor de la variable *Y* para valores no observados de la/s variable/s *X*.

Estas dos posibilidades tienen un gran valor de cara al análisis y la resolución de muy diversos problemas prácticos, en los que suelen aparecer determinadas variables de interés (científico, industrial, económico, político, social, medioambiental, y un largo etcétera) que, por varias razones, es necesario o conveniente expresar o explicar en función de otras variables.

El modelo de relación más sencillo entre un conjunto de variables es el **lineal**, esto es, aquel en el que la función *f* que conecta/relaciona la variable dependiente *Y* con las explicativas es de tipo lineal. Dentro de estos modelos de regresión lineales, el caso más básico es aquel en el que se tiene una única variable explicativa o independiente *X*, de modo que se busca generalizar las observaciones de una variable estadística bidimensional (X,Y) mediante un modelo cuantitativo

$$Y = f(X) = \beta_0 + \beta_1 X$$

que constituya una *buena* aproximación lineal de los datos. Estos modelos con solo dos variables se conocen como **simples**, por oposición a los **múltiples**, que aparecen al considerarse más de una variable explicativa.

Así pues, comenzaremos el estudio de los modelos de regresión atendiendo a estos modelos de **regresión lineal simple,** para luego estudiar su generalización al caso de la **regresión lineal múltiple**.

2.1 CONCEPTOS BÁSICOS

De cara al objetivo de estudiar y modelizar de manera funcional cómo se relacionan un par de variables estadísticas, es preciso recordar algunas nociones básicas, en particular las de **covarianza** y **correlación**, que ya avanzan en la dirección de describir la relación existente entre dos variables.

Así, recuérdese que si se tiene un conjunto de *n* observaciones bidimensionales de las variables *X* e *Y* de la forma $(x_i, y_i) \in \mathbb{R}^2$, $i = 1,...,n$, la **covarianza** (muestral) entre *X* e *Y* se define como

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

donde $\overline{x}$ e $\overline{y}$ representan las medias aritméticas o muestrales de las observaciones de cada variable. Nótese que (a diferencia de la varianza) la covarianza puede ser positiva o negativa.

Es importante entender la interpretación del signo que toma esta medida de covarianza. En concreto, un valor positivo de la covarianza indica que existe una **relación directa** entre ambas variables, esto es, cuando una variable se incrementa también tiende a incrementarse la otra;

de modo similar, un valor negativo de S_{XY} indica la existencia de una **relación inversa** entre *X* e *Y*. Dicho con otras palabras, la covarianza será positiva si las observaciones (x_i, y_i) tienden a agruparse alrededor de una recta creciente, y negativa si se tienden a agrupar en torno a una recta decreciente. Por este motivo, suele decirse que la covarianza (o mejor dicho su signo) describe el tipo de relación **lineal** entre un par de variables.

La Figura 2.1 más abajo permite entender intuitivamente esta asociación entre el signo de la covarianza y el tipo de relación lineal entre dos variables. Nótese que la disposición de cada observación (x_i, y_i) en alguno de los cuatro cuadrantes con origen en el punto $(\overline{x}, \overline{y})$ determina el signo de la contribución $(x_i - \overline{x})(y_i - \overline{y})$ de esa observación a la covarianza. Así, si las observaciones con contribución positiva son predominantes, entonces de algún modo ambas variables tienden a crecer conjuntamente, y se tendrá que $S_{XY} > 0$. Por el contrario, si las que predominan son las observaciones con contribución negativa, entonces existirá una relación inversa entre *X* e *Y*, y se tendrá que $S_{XY} < 0$.

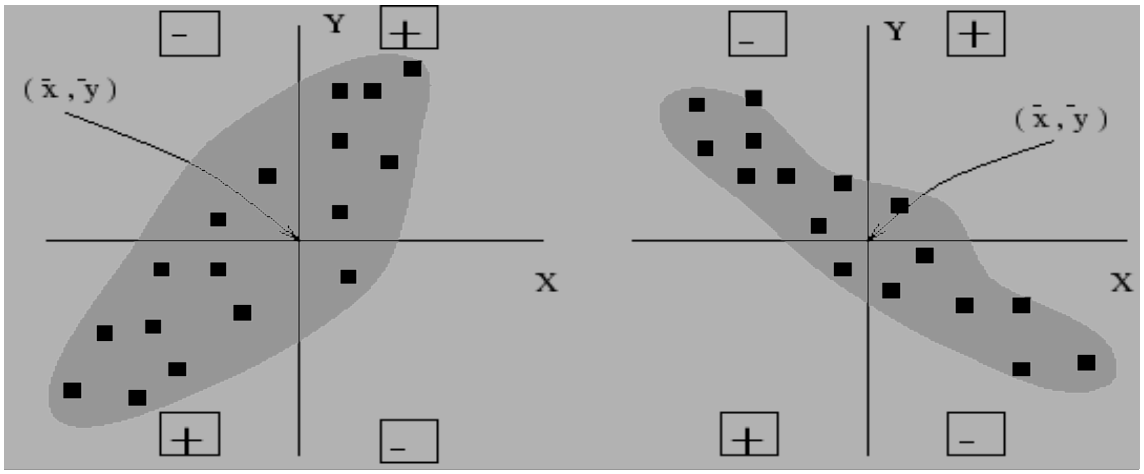

Figura 2.1: Diagrama de cuadrantes para interpretar el signo de la covarianza

Aunque, como acabamos de ver, el signo de la covarianza es informativo sobre el tipo de relación lineal existente entre las variables consideradas, su magnitud $|S_{XY}|$ no aporta realmente ninguna información sobre la intensidad de esa relación (esto es, cuánto de próximas están a esa recta en torno a la que se agrupan las observaciones), ya que esta magnitud es dependiente de las unidades con que se miden las variables *X* e *Y*. En particular, si se realiza un cambio de escala en estas variables, digamos que $X' = aX$ e $Y' = bY$, entonces se tiene que $S_{X'Y'} = abS_{XY}$. Esto invalida la covarianza como una medida de la intensidad de la relación (lineal) entre variables.

Para solucionar este problema y poder contar con una medida descriptiva de la intensidad de la relación lineal entre variables, se introduce la noción de correlación lineal, que como veremos no será dependiente de cambios de escala. Así pues, recuérdese que el **coeficiente de correlación lineal** (de Pearson) viene definido por

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

donde S_X y S_Y denotan respectivamente las (cuasi) desviaciones típicas muestrales de las observaciones de las variables *X* e *Y*. En tanto que $S_{X'} = |a| S_X$ y $S_{Y'} = |b| S_Y$, es directo observar que el coeficiente de correlación es invariante ante cambios de escala de las variables consideradas. Además, es claro que la correlación r_{XY} presenta el mismo signo que la covarianza

S_{XY} puesto que $S_X, S_Y \geq 0$. Como también es fácil comprobar que se cumple la desigualdad $S_{XY} \leq S_X S_Y$, se tiene que el coeficiente de correlación lineal está acotado entre -1 y +1, esto es, $-1 \leq r_{XY} \leq 1$. De hecho, los valores extremos +1 y -1 corresponden respectivamente a los casos en que todas las observaciones (x_i, y_i) están perfectamente alineadas de manera creciente o decreciente.

Un último apunte es que el coeficiente de correlación lineal de Pearson de las variables *X* e *Y* es idéntico a la covarianza obtenida al estandarizar las variables. Esto es, si para cada *i* = 1,…,*n* se denota ahora

$$x_i' = \frac{x_i - \overline{x}}{S_X} \text{ e } y_i' = \frac{y_i - \overline{y}}{S_Y}$$

entonces es directo comprobar que

$$r_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{S_X}\right)\left(\frac{y_i - \overline{y}}{S_Y}\right) = S_{X'Y'}.$$

Es importante recalcar que el coeficiente de correlación r_{XY} solo mide la intensidad (y la dirección) de la relación lineal entre las observaciones de las variables *X* e *Y*. Si la relación entre estas variables es no lineal, la correlación r_{XY} deja de ser una medida fiable de la intensidad de esa relación. En particular, aun existiendo una relación funcional perfecta entre las variables, por ejemplo si $Y = X^2$, se puede tener que $r_{XY} = 0$ (**Ejercicio**: construir un ejemplo que pruebe esta afirmación).

Aunque tenemos una medida *relativamente* fiable de la dirección y la intensidad de la relación lineal entre dos variables *X* e *Y*, dada por la correlación lineal, es preciso entender que sin embargo esta medida no permite realmente cuantificar en términos precisos esa relación. En otras palabras, la correlación por sí misma no proporciona un modelo matemático *f* que especifique funcionalmente la relación (aproximada) $Y = f(X)$, donde *f* representa una función de tipo lineal. Más allá del interés matemático de obtener aproximaciones lineales a un conjunto de puntos (x_i, y_i) del plano, el interés en dar este paso reside en la **gran variedad de aplicaciones prácticas** que permite la construcción de este tipo de modelos, en particular cuando la variable *independiente X* se reemplaza por un conjunto de variables X_1,…, X_k, que son usadas conjuntamente para **explicar** y **predecir** una variable (*dependiente*) de interés *Y*.

Ejemplos de aplicaciones prácticas de los modelos de regresión

1. **Ciencias agrarias:** En una cooperativa ganadera se tiene interés en la capacidad de predecir la producción de leche en el mes actual a partir de una serie de variables que pueden ser medidas fácilmente.
2. **Estudios sociales:** ¿Qué efecto ha tenido la aplicación o no de una determinada ley sobre el coste de la vida para una familia media en distintas zonas del país?
3. **Arqueología/Antropología:** Estimación de la antigüedad de restos arqueológicos, por ejemplo, calaveras egipcias, a partir de algunas de sus características fisiológicas.
4. **Toma de decisiones políticas:** De cara a elaborar medidas adecuadas para gestionar la inmigración doméstica, es preciso establecer un modelo que prediga los movimientos poblacionales que ocurrirán y que permita responder a la pregunta de por qué la gente deja un sitio para ir a otro, a ser posible en base a una colección de indicadores socioeconómicos de las distintas regiones.

5. **Medio ambiente:** ¿Qué tipo de actividades en las cuencas hidrográficas de los ríos están asociadas a mayores niveles de polución en esos ríos?
6. **Sector inmobiliario:** ¿Es posible predecir el precio de una vivienda a partir de una descripción de sus características (superficie, antigüedad, localización, calidad, etc.)? ¿Es posible usar estas predicciones (y otras variables socioeconómicas) como indicadores para estimar la renta disponible en un hogar?

Estos casos (reales, como veremos) son solo una muestra de la enorme utilidad práctica que encuentran los modelos de regresión. En todos estos casos, la clave de su utilidad reside en que la aplicación de un modelo de regresión a datos adecuadamente recogidos puede permitir establecer una relación matemática entre los valores de la variable de interés y los valores de las variables descriptoras, de modo que sea posible evaluar el efecto de estas variables sobre la variable de interés y realizar predicciones para combinaciones nuevas de valores de esas variables explicativas o independientes. Este tipo de modelos, que pueden programarse fácilmente, permiten entonces un estudio más sistemático de los problemas que nos interesan, así como la automatización de procesos relacionados con la estimación de cantidades relevantes en la explotación de grandes volúmenes de datos.

2.2 SUPUESTO DE LINEALIDAD Y ESTIMACIÓN DEL MODELO

Suponiendo entonces que se cuenta con un conjunto de *n* observaciones bidimensionales $(x_i, y_i) \in \mathbb{R}^2$, $i = 1,...,n$, de un par de variables *X* e *Y*, nos centraremos por ahora en la estimación de un modelo lineal que permita explicar/predecir los valores observados de la variable *Y* a partir de los valores observados de la variable *X*.

Es importante entender que esto conlleva implícitamente el introducir una primera hipótesis o supuesto sobre el tipo de relación funcional que se da entre las variables *X* e *Y*, a saber, que esta relación existe y es de tipo lineal.

(S1) Supuesto 1: Se supone que existe una relación estadística de tipo lineal entre las variables *X* e *Y*, en particular esta relación toma la forma

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

donde $\beta_0, \beta_1 \in \mathbb{R}$ son coeficientes o parámetros (usualmente desconocidos) que caracterizan el modelo lineal, y ε es un término de error aleatorio, que recoge las desviaciones del modelo lineal para ajustar los datos exactamente.

Nótese que se supone la existencia de una relación funcional entre las variables, y de hecho se establece que su forma ha de ser lineal. Pero cuantitativamente la relación en sí no es conocida, ya que depende de **parámetros desconocidos** (y habitualmente no observables). Además, esta relación de tipo lineal establece de algún modo la influencia o efecto de la variable *X* sobre *Y*, pero no recoge el sinfín de otros factores que pueden haber influenciado la aparición de los valores concretos que han sido observados en la variable *Y*.

Estas fluctuaciones de los valores observados de *Y* alrededor de la recta $\beta_0 + \beta_1 X$, debidas a otros factores no observados o no contemplados en el modelo, son recogidas en el **término de error ε**, que tendrá entonces un comportamiento aleatorio, esto es, **ε es una variable aleatoria** que representa las diferencias entre los valores observados de *Y* y los que se obtendrían a partir de los valores observados de *X* a través de la recta anterior.

Para la muestra observada $(x_1, y_1),...,(x_n, y_n)$, la hipótesis anterior puede escribirse en términos del sistema de ecuaciones

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

o, más resumidamente,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ con } i = 1,\ldots,n .$$

Nótese que **los parámetros** β_0, β_1 **son los mismos en cada ecuación** asociada a un par de datos (x_i, y_i), pero también que **el término de error** ε_i **es específico de cada ecuación**. De algún modo, se está suponiendo que cada par de observaciones (x_i, y_i) está relacionado a través de la misma expresión $y_i \approx f(x_i) = \beta_0 + \beta_1 x_i$, que aproxima el valor observado y_i a partir de x_i, constituyendo el término de error ε_i la diferencia entre la aproximación efectuada, $f(x_i) = \beta_0 + \beta_1 x_i$, y el valor real observado, y_i, esto es,

$$\varepsilon_i = y_i - f(x_i) = y_i - \beta_0 - \beta_1 x_i .$$

El hecho de aplicar una hipótesis o supuesto acerca de una relación entre dos variables sobre un conjunto de observaciones empíricas de esas dos variables tiene la utilidad de abrirnos un camino para contrastar nuestra suposición mediante la experiencia, observando cómo el modelo obtenido en base a ese supuesto se ajusta al comportamiento de esas variables en la realidad (o al menos en ese *trozo* de la realidad que es la muestra).

Pero, para ello, es antes necesario proporcionar una manera de obtener un modelo concreto, esto es, que especifique cuantitativamente la relación entre las variables, lo que nos conduce al problema de **cómo estimar los parámetros** β_0 y β_1, que precisamente se encargan de caracterizar el modelo desde el punto de vista cuantitativo.

A este respecto, es preciso observar que a priori existen **infinitos modelos** $Y = \beta_0 + \beta_1 X + \varepsilon$ que podrían ajustar los datos, en tanto que las posibles elecciones de β_0 y β_1 son incontables. En otras palabras, dada una colección $(x_i, y_i) \in \mathbb{R}^2$, $i = 1,\ldots,n$, de puntos del plano, en principio podríamos escoger los parámetros β_0, β_1 y los términos de error ε_i de infinitas maneras de modo que se cumpla el sistema de ecuaciones anterior. Es decir, de las infinitas rectas en el plano, ¿cuál de ellas he de elegir para que represente la relación aproximada entre las observaciones de ambas variables? Obviamente, no todas las rectas serán igual de *buenas* a la hora de aproximar esa relación.

Por ejemplo, algunas de ellas producirán términos de error ε_i muy grandes para todas las observaciones, otras solo para algunas observaciones, y aun otras podrían tener valores de error relativamente pequeños en todas o casi todas las observaciones. Claramente, siempre será preferible una recta de esta última clase a una de las de la primera clase. Así que **las magnitudes de los errores** ε_i **proporcionan de hecho pistas importantes de cara a establecer qué es una *buena* recta para mi modelo**.

Con esto en mente, una buena idea para resolver nuestro problema de estimación de los parámetros β_0 y β_1 es plantearlo como un **problema de optimización**, en el que se busquen los valores β_0, β_1 que minimicen la suma de las magnitudes $|\varepsilon_i|$ de los errores, esto es, el problema de optimización no lineal (pero, ¿linealizable?)

$$\min S(\beta_0,\beta_1)=\sum_{i=1,\dots,n}|\varepsilon_i|$$
sujeto a
$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,\dots,n$$
$$\beta_0,\beta_1\in\mathbb{R}$$

Este problema de optimización fue históricamente complicado de resolver por la dificultad de aplicar herramientas analíticas (i.e. derivación) sobre la función *S*, una suma de valores absolutos. En su lugar, era mucho más tratable el problema obtenido al considerar la minimización de la suma de los cuadrados ε_i^2 de los errores, dado por

$$\min S(\beta_0,\beta_1)=\sum_{i=1,\dots,n}\varepsilon_i^2$$
sujeto a
$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,\dots,n$$
$$\beta_0,\beta_1\in\mathbb{R}$$

Al ser un problema sin restricciones reales sobre las variables de decisión β_0,β_1 (recuérdese que los pares (x_i,y_i) han sido observados), este puede ser resuelto fácilmente utilizando técnicas analíticas (**Ejercicio**). En particular, al igualar a cero las derivadas parciales respecto a β_0 y β_1 de la función *S* se obtiene rápidamente que el mínimo global de la función $S(\beta_0,\beta_1)=\sum_{i=1,\dots,n}\varepsilon_i^2$ está dado por las estimaciones

$$\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}=\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{(x_i-\overline{x})\cdot\overline{x}}{\sum_{j=1,\dots,n}(x_j-\overline{x})^2}\right)y_i$$

$$\hat{\beta}_1=\frac{S_{XY}}{S_X^2}=\sum_{i=1}^{n}\frac{(x_i-\overline{x})(y_i-\overline{y})}{\sum_{j=1,\dots,n}(x_j-\overline{x})^2}=\sum_{i=1}^{n}\frac{(x_i-\overline{x})}{\sum_{j=1,\dots,n}(x_j-\overline{x})^2}y_i$$

Este método de estimación se conoce comúnmente como **estimación por mínimos cuadrados**, y por este motivo a $\hat{\beta}_0$ y $\hat{\beta}_1$ se les suele conocer también como los **estimadores de mínimos cuadrados** de los coeficientes de regresión β_0 y β_1. Nótese que en la práctica será más eficiente computar primero $\hat{\beta}_1$ (a partir de la covarianza y la varianza muestrales de *X*) y luego obtener $\hat{\beta}_0$ a partir de $\hat{\beta}_1$ usando las medias muestrales de *X* e *Y*.

Introduzcamos ahora algunas definiciones habituales en el marco de la regresión, así como algunas propiedades de los estimadores de mínimos cuadrados.

Definición 1. Llamamos ***valores predichos*** $\hat{y}_i$ a los producidos mediante el modelo estimado $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$ para los valores observados x_i, esto es,

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i \text{ para todo } i=1,\dots,n.$$

Y se conocen como **residuos** e_i a las diferencias entre los valores observados y_i y los valores predichos, esto es,

$$e_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1 x_i,\ i=1,\dots,n.$$

Proposición 1. Se cumplen las siguientes propiedades de los estimadores y residuos de mínimos cuadrados:

1) $\sum_{i=1,\dots,n} e_i = 0$

2) $\sum_{i=1,\dots,n} \hat{y}_i = \sum_{i=1,\dots,n} y_i = n\overline{y}$

3) La recta $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ pasa por el punto $(\overline{x}, \overline{y})$.

4) $\sum_{i=1,\dots,n} x_i e_i = 0$

5) $\sum_{i=1,\dots,n} \hat{y}_i e_i = 0$

6) Los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$, así como los valores predichos $\hat{y}_i, i = 1,\dots,n$, se pueden expresar como una combinación lineal de los valores observados y_i.

Demostración. Ejercicios. Nótese que el método de estimación de $\hat{\beta}_1$ es indiferente en las proposiciones 1 – 3. En cambio, las proposiciones 4 – 6 solo son ciertas cuando $\hat{\beta}_1$ es el estimador de mínimos cuadrados.

2.3 SUPUESTO PROBABILÍSTICO. TEOREMA DE GAUSS-MARKOV

Nótese que, hasta este momento, solo se ha introducido una suposición, el supuesto S1 sobre la relación lineal entre las variables *X* e *Y*, mediada por los parámetros inobservables β_0 y β_1 y distorsionada por una fuente de error aleatorio ε. Este supuesto, de carácter bastante general, nos ha permitido desarrollar un método de estimación de estos parámetros relativamente sensato, la estimación por mínimos cuadrados.

De cara a valorar la bondad de este método de estimación, necesitamos, como en cualquier otro contexto de estimación paramétrica, estudiar el **sesgo** y la **variabilidad** de los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$. Es decir, necesitamos conocer si en promedio nuestros estimadores aproximan correctamente los parámetros que pretenden estimar, y con qué precisión lo hacen. En otras palabras, nos interesa obtener expresiones para las esperanzas y varianzas $E[\hat{\beta}_0], E[\hat{\beta}_1], V[\hat{\beta}_0], V[\hat{\beta}_1]$, que nos permitan conocer el sesgo de los estimadores y cómo su precisión varía en función de factores como el tamaño muestral o la presencia de datos atípicos.

Sin embargo, no es posible obtener estas esperanzas y varianzas a partir del supuesto S1 exclusivamente. Esto es así ya que, en particular, la hipótesis S1 no introduce supuestos probabilísticos sobre las variables *X* e *Y* más allá de informarnos de que la variable *Y* está afectada por un error aleatorio ε y, por lo tanto, ha de ser también una variable aleatoria. En estas condiciones, sin conocer los momentos (medias, varianzas poblacionales) de las variables *X* y ε, no es posible obtener los momentos poblacionales de los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$. En este sentido, nótese que, por ejemplo, la esperanza de las observaciones $y_i = Y|X{=}x_i$ está determinada por esos momentos, ya que

$$E[y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 E[x_i] + E[\varepsilon_i],$$

por lo que no es posible determinar $E[y_i]$ sin conocer $E[x_i]$ y $E[\varepsilon_i]$. De igual modo, no será posible obtener los momentos de $\hat{\beta}_0$ y $\hat{\beta}_1$ sin introducir nuevos supuestos sobre la variable *X* y sobre el error ε.

En relación a la variable *X*, el supuesto más habitual es que esta variable puede ser, y de hecho es, **controlada** por el experimentador que observa los pares muestrales (x_i, y_i), $i = 1,\ldots,n$. Esto quiere decir que el experimentador *fija* los valores x_i de *X* para los que quiere observar el correspondiente valor y_i de *Y*. De algún modo, bajo este supuesto, la variable *X,* o mejor dicho cada valor x_i, representa las condiciones bajo las cuales es observada la variable *Y*, y se entiende que dichas condiciones son controlables por el experimentador. En este contexto, **la variable *X* no es una variable aleatoria, sino una variable controlable o determinística**, y por lo tanto los valores x_i concretos que toma no están sujetos a aleatoriedad, esto es, $E[x_i] = x_i$ y $V[x_i] = 0$.

Obviamente, otros factores aparte de la variable *X* pueden influir en los valores observados de *Y*, pero bajo S1 se supone que estos factores no tienen un efecto sistemático en *Y*, por lo que no se controlan y no se contemplan explícitamente en el modelo, sino que se entiende que se acumulan en el término de error ε. Este error ε sí es entonces una variable aleatoria, ya que en este contexto recoge las diversas contribuciones sobre *Y* de un conjunto de factores no controlados. En particular, cada realización ε_i de este proceso de error constituye una desviación, que puede ser positiva o negativa, sobre el valor de *Y* que se observaría si solo tuviese influencia el valor x_i dispuesto por el experimentador (bajo S1, sería $\beta_0 + \beta_1 x_i$). Por este motivo, no es descabellado suponer que, en promedio, **al repetir el experimento numerosas veces bajo las mismas condiciones controladas x_i, esas desviaciones tenderán a compensarse entre sí**, de modo que el error medio sería $E[\varepsilon_i] = 0$. Esto se puede extender al resto de observaciones $i = 1,\ldots,n$.

Igualmente, en una primera aproximación puede ser razonable asumir que la variabilidad de estos errores ε_i es la misma para cada observación *i*, esto es, $V[\varepsilon_i] = \sigma^2$ para todo $i = 1,\ldots,n$. En otras palabras, en principio se puede suponer que **la magnitud o dispersión de estos errores es en promedio la misma en cada condición x_i en que se lleva a cabo el experimento** u observación de *Y*.

Y finalmente, también puede ser en principio razonable suponer que el error que aparece en una observación, digamos la *i*-ésima, no tiene influencia en el error asociado a otra observación, por ejemplo la *j*-ésima. En otras palabras, **se supone que los errores son independientes entre sí**, lo que en particular conlleva que también **están incorrelados entre sí**, $\text{Cor}(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$.

Así pues, se pueden resumir estos supuestos de tipo probabilístico sobre la variable *X* y el término de error ε como sigue.

(S2) Supuesto 2: Se supone que la variable *X* está controlada por el experimentador, por lo que no tiene carácter aleatorio. Además, se supone que todos los términos de error aleatorio ε_i son independientes entre sí y proceden de una misma distribución con media 0 y varianza constante σ^2, esto es, se supone que

i) $E[\varepsilon_i] = 0$ y $V[\varepsilon_i] = \sigma^2$ para todo $i = 1,\ldots,n$, y
ii) $\text{Cor}(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$.

Al igual que con el supuesto S1, este nuevo supuesto S2 debe ser validado en la práctica. En capítulos siguientes veremos que es posible llevar a cabo esta validación mediante el análisis de los residuos e_i, que en cierto modo constituyen aproximaciones de los errores ε_i. Esta comprobación es importante ya que, como veremos en un momento, bajo S2 los estimadores

de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen propiedades muy deseables, que podrían no darse en caso de que este supuesto no se cumpla en la práctica.

Asumiendo S2, los **momentos poblacionales de las observaciones y_i** pueden ser obtenidos sin dificultad. En particular, para todo $i, j = 1,\ldots,n$ se tiene que

$$E[y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 x_i + E[\varepsilon_i] = \beta_0 + \beta_1 x_i$$

$$V[y_i] = V[\beta_0 + \beta_1 x_i + \varepsilon_i] = V[\varepsilon_i] = \sigma^2$$

$$Cov(y_i, y_j) = \begin{cases} \sigma^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

De igual modo, es ya posible obtener los **momentos poblacionales** de $\hat{\beta}_0$ y $\hat{\beta}_1$:

Proposición 2: Suponiendo S1 y S2, se tienen los siguientes resultados:

1) $E[\hat{\beta}_1] = \beta_1$

2) $E[\hat{\beta}_0] = \beta_0$

3) $V[\hat{\beta}_1] = \dfrac{\sigma^2}{\sum\limits_{i=1,\ldots,n} (x_i - \overline{x})^2}$

4) $V[\hat{\beta}_0] = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum\limits_{i=1,\ldots,n} (x_i - \overline{x})^2} \right)$

Demostración.

1) $E[\hat{\beta}_1] = E\left[\dfrac{\sum_i (x_i - \overline{x}) y_i}{\sum_i (x_i - \overline{x})^2} \right] = \dfrac{\sum_i (x_i - \overline{x}) E[y_i]}{\sum_i (x_i - \overline{x})^2} = \dfrac{\sum_i (x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{\sum_i (x_i - \overline{x})^2} = \beta_0 \dfrac{\sum_i (x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} + \beta_1 \dfrac{\sum_i (x_i - \overline{x}) x_i}{\sum_i (x_i - \overline{x})^2} = \beta_1$

2) $E[\hat{\beta}_0] = E[\overline{y} - \hat{\beta}_1 \overline{x}] = E[\overline{y}] - \overline{x}\, E[\hat{\beta}_1] = E[\frac{1}{n} \sum_i (\beta_0 + \beta_1 x_i + \varepsilon_i)] - \overline{x} \beta_1 = \frac{1}{n} \sum_i (\beta_0 + \beta_1 x_i) - \overline{x} \beta_1 = \beta_0$

3) $V[\hat{\beta}_1] = V\left[\dfrac{\sum_i (x_i - \overline{x}) y_i}{\sum_i (x_i - \overline{x})^2} \right] = \dfrac{\sum_i (x_i - \overline{x})^2 V[y_i]}{\left(\sum_i (x_i - \overline{x})^2 \right)^2} + \dfrac{\sum_i \sum_{j \neq i} (x_i - \overline{x})(x_j - \overline{x}) Cov[y_i, y_j]}{\left(\sum_i (x_i - \overline{x})^2 \right)^2} = \dfrac{\sum_i (x_i - \overline{x})^2 \sigma^2}{\left(\sum_i (x_i - \overline{x})^2 \right)^2} = \dfrac{\sigma^2}{\sum_i (x_i - \overline{x})^2}$

4) $V[\hat{\beta}_0] = V[\overline{y} - \hat{\beta}_1 \overline{x}] = V[\overline{y}] + V[\hat{\beta}_1 \overline{x}] - 2Cov[\overline{y}, \hat{\beta}_1 \overline{x}] = \dfrac{\sigma^2}{n} + \overline{x}^2 V[\hat{\beta}_1] - 2\overline{x} Cov[\overline{y}, \hat{\beta}_1] =$

$$= \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2} \right)$$

Ejercicio: Demostrar que $Cov[\overline{y}, \hat{\beta}_1] = 0$.

Ahora ya estamos en condiciones de establecer que los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ son los *mejores* estimadores lineales de los parámetros β_0 y β_1.

Teorema de Gauss-Markov: Dado el modelo $Y = \beta_0 + \beta_1 X + \varepsilon$ (S1) y suponiendo que $E[\varepsilon_i] = 0$, $V[\varepsilon_i] = \sigma^2$ y que los errores ε_i, ε_j son incorrelados entre sí para todo $i \neq j$ (S2), se tiene que los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados y tienen mínima varianza dentro de la clase de estimadores insesgados lineales (i.e. obtenidos como combinación lineal de los y_i), en otras palabras, son *BLUE* (Best Linear Unbiased Estimators).

Demostración. Se verá de modo más general en el contexto de la regresión lineal múltiple.

Es importante señalar que este Teorema de Gauss-Markov responde afirmativamente a la pregunta que nos hacíamos antes acerca de si el método de estimación por mínimos cuadrados (desarrollado suponiendo solo S1) proporciona *buenos* estimadores de los parámetros de regresión β_0 y β_1. Y en efecto así es cuando se asume también el supuesto S2, ya que en esas condiciones los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ reúnen **dos propiedades que los hacen ciertamente idóneos**:

i) En primer lugar, son **insesgados**, esto es, en promedio se asemejan al parámetro que estiman;
ii) Y en segundo lugar, y refiriéndonos a la clase de estimadores insesgados obtenidos como combinación lineal de las observaciones y_i, tienen **mínima varianza**, esto es, son los estimadores más precisos que se pueden encontrar cuando nos restringimos a esa clase de estimadores lineales insesgados[1].

Entonces, ¿podemos desarrollar solo a partir de los supuestos S1 y S2 una **teoría inferencial** que nos permita el uso de herramientas prácticas estadísticas como intervalos de confianza o contrastes de hipótesis? En el fondo, aunque tengamos unos estimadores con buenas propiedades, vamos a necesitar este tipo de herramientas inferenciales para poder dar respuesta a cuestiones prácticas tales como:

- ¿La muestra disponible permite garantizar (con cierto nivel de confianza) que la variable *X* tiene un efecto lineal significativo en nuestra respuesta *Y*?
- ¿Cuál es la incertidumbre asociada a la estimación de los parámetros del modelo de regresión?
- ¿Cómo construir un estimador de σ^2, la varianza poblacional (supuesta constante) del error ε?
- ¿El modelo que proporcionan los estimadores de mínimos cuadrados para unos datos determinados es adecuado?

La respuesta, como veremos a continuación, es que los supuestos S1 y S2 no se bastan para desarrollar herramientas inferenciales en este contexto de regresión. Por ello, necesitaremos introducir un nuevo supuesto de tipo distribucional, que especifique modelos de probabilidad concretos (no solo los momentos de la distribución) sobre los que aplicar técnicas de estadística inferencial.

2.4 SUPUESTO DISTRIBUCIONAL Y ESTIMACIÓN DE σ^2

Una cuestión clave de cara a la aplicación práctica de los modelos de regresión reside en la cuestión de su ***significatividad***. Para entender esto mejor, pongámonos en la situación en que

[1] La restricción de linealidad no es especialmente relevante, pero sí tiene cierta importancia la de no sesgo. En particular, se han desarrollado en la literatura especializada diversos métodos de estimación que proporcionan estimadores sesgados, pero con varianza menor que la de los estimadores de mínimos cuadrados.

hemos observado una muestra $(x_1, y_1),\ldots,(x_n, y_n)$ y hemos ajustado un modelo de regresión concreto del tipo $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, por ejemplo se ha obtenido $\hat{Y} = 18 + 0.5X$. Aunque sabemos que estas estimaciones $\hat{\beta}_0 = 18$ y $\hat{\beta}_1 = 0.5$ son teóricamente *buenas*, en la práctica no disponemos aun de una manera de medir realmente esta bondad, ya que, por ejemplo, no conocemos cómo estimar la varianza σ^2 del error, de la que depende la dispersión de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$. Y sin ello no podemos valorar la incertidumbre asociada a que, si hubiéramos observado una muestra distinta a la que hemos usado para estimar la recta anterior, nuestro modelo podría ser muy diferente, por ejemplo $\hat{Y} = 18 - 0.5X$ o $\hat{Y} = 18(-0X)$. Nótese que hay diferencias prácticamente relevantes entre estos modelos, ya que en el primero la variable *X* tiene una contribución positiva sobre *Y*, mientras que en el segundo es negativa y en el tercero inexistente.

La idea de significatividad estadística precisamente hace referencia a este tipo de incertidumbre, asociada a la **variabilidad muestral** que subyace a cualquier proceso de recogida/generación de datos en el que la **aleatoriedad** juegue un papel relevante (por ejemplo, debido a la aleatoriedad en la selección de la muestra, o a la introducción de errores aleatorios no controlables en nuestras observaciones). En el caso de la regresión, la aleatoriedad se introduce a través del término de error ε, que introduce ruido en nuestras observaciones de un proceso supuestamente lineal, y ello conlleva la aparición de **incertidumbre acerca de la validez del modelo concreto que se ha obtenido**: si hubiésemos captado otro ruido, quizás nuestro modelo podría presentar diferencias relevantes.

Afortunadamente, la **estadística inferencial** proporciona valiosas herramientas para tratar con este tipo de incertidumbre. Así, por ejemplo, si en la situación anterior pudiésemos decir que con un 95% de confianza el parámetro β_1 se encuentra en el intervalo [0.25,0.6], o que se puede rechazar la hipótesis nula $H_0 : \beta_1 = 0$ con un p-valor p = 0.001, entonces estaríamos en condiciones de afirmar con bastante seguridad (esto es, con poca incertidumbre) que realmente el parámetro β_1 es positivo. Más técnicamente, diríamos que este parámetro es significativamente mayor que 0, o que el modelo de regresión obtenido es significativo.

Así pues, necesitamos poder aplicar este tipo de herramientas de **estadística inferencial en el contexto de la regresión lineal**, para **valorar y tratar la incertidumbre** asociada a la obtención de modelos matemáticos de relación entre variables a partir de información finita y ruidosa.

Pero para ello, sin embargo, será preciso introducir un **nuevo supuesto** que establezca un modelo probabilístico particular, una familia concreta de distribuciones de probabilidad, que represente con más detalle esa **variabilidad asociada al error *ε***. Conocer solo los momentos de esa distribución, como nos ha permitido el supuesto S2, no es suficiente, ya que estos por sí solos no nos permiten obtener los percentiles distribucionales del tipo $z_{\alpha/2}$ o $t_{n;\alpha/2}$ que necesitamos para establecer intervalos de confianza $100 \cdot (1 - \alpha)\%$, o regiones críticas para contrastes de hipótesis con un nivel de significación α conocido. Para obtener este tipo de percentiles necesitamos conocer las distribuciones de muestreo de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, de cara a ser capaces de producir variables pivotales que sustenten el uso de las técnicas inferenciales. Este es el objetivo del nuevo supuesto que se introduce a continuación.

(S3) Supuesto 3: Los errores ε_i se distribuyen normalmente, esto es, $\varepsilon_i \sim N(0, \sigma^2)$ para todo i = 1,…,n.

En conjunción con los supuestos anteriores, S1 y sobre todo S2, este supuesto conlleva una serie de consecuencias inmediatas (**Ejercicio**):

i) Se puede entender el término de error ε como una variable normal multivariante $\varepsilon \sim N(\mathbf{0}, \Sigma)$, con $\mathbf{0} = (0,\ldots,0)$ y $\Sigma = \sigma^2 \cdot Id_{nxn}$.

ii) Además, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ para todo $i = 1,\ldots,n$, y también son independientes.

iii) Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ son también variables aleatorias normales.

A partir de estos resultados es entonces directo comprobar que, por ejemplo, el estadístico *Z* dado por

$$Z = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sqrt{V[\hat{\beta}_1]}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum (x_i - \overline{x})^2}} \sim N(0,1)$$

es una variable pivotal para el parámetro β_1, ya que aunque el estadístico *Z* es una función del parámetro β_1, la distribución de *Z* no depende de este parámetro. Si conociéramos el valor de la varianza σ^2, sería entonces directo usar esta variable pivotal *Z* para construir intervalos de confianza o estadísticos de contraste para el parámetro β_1, usando los correspondientes percentiles $z_{\alpha/2}$ de la distribución normal estándar. Un resultado similar se puede obtener sin dificultad para β_0.

En la práctica habitual, sin embargo, la varianza σ^2 no será conocida, y habrá que estimarla a partir de los datos.

Estimación de σ^2 a través de la Suma de Cuadrados de los Residuos

La manera más habitual de estimar σ^2 es a través del estadístico conocido como ***suma de cuadrados de los residuos***, que se denota por SCR y viene dado por

$$SCR = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 .$$

No es complejo demostrar que, bajo los supuestos S2 y S3, se tiene que

$$\frac{SCR}{\sigma^2} \sim \chi^2_{n-2},$$

y entonces es directo obtener que

$$E\left[\frac{SCR}{n-2}\right] = \sigma^2 .$$

Esto nos proporciona un estimador insesgado de σ^2, dado por la ***media de cuadrados de los residuos***, denotada por MCR y dada por

$$\hat{\sigma}^2 = MCR = \frac{SCR}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} .$$

Es importante darse cuenta de que la bondad de $\hat{\sigma}^2 = MCR$ como estimador de la varianza σ^2 depende de los supuestos S2 y S3, y en particular también de S1 ya que se computa a partir de los residuos del modelo de regresión. También es importante el cumplimiento de S3, ya que de lo contrario la distribución de la suma de cuadrados de los residuos no se distribuiría como una χ^2 y no se tendría por tanto un estimador insesgado de σ^2.

Así pues, en caso de que estos supuestos no se cumplan en la práctica, la utilidad de $\hat{\sigma}^2 = MCR$ como estimador se puede ver comprometida seriamente. A cambio de esta desventaja, el

estimador $\hat{\sigma}^2 = MCR$ es fácil de obtener directamente a partir de cualquier conjunto de observaciones $(x_1, y_1),\ldots,(x_n, y_n)$.

2.5 CONTRASTES DE HIPÓTESIS Y SIGNIFICATIVIDAD DE LA REGRESIÓN

Una vez contamos con la posibilidad de estimar σ^2 a partir de la muestra $(x_1, y_1),\ldots,(x_n, y_n)$, es directo aplicar esta estimación a la construcción de variables pivotales con las que desarrollar herramientas inferenciales como contrastes de hipótesis e intervalos de confianza. Nos centraremos en esta sección en el primer tipo de herramienta, los contrastes de hipótesis que permiten valorar la significatividad del modelo de regresión obtenido a partir de la muestra observada.

Contrastes *t* sobre el valor de los parámetros β_0 y β_1

Así pues, en tanto que

$$\hat{\beta}_{1,} \sim N(\beta_1, \sigma^2 / \sum (x_i - \overline{x})^2) \Leftrightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum (x_i - \overline{x})^2}} \sim N(0,1)$$

$$\frac{SCR}{\sigma^2} = \frac{(n-2)MCR}{\sigma^2} \sim \chi^2_{n-2}$$

se tiene que

$$t_0 = \frac{\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum (x_i - \overline{x})^2}}}{\sqrt{\dfrac{\dfrac{(n-2)MCR}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MCR / \sum (x_i - \overline{x})^2}} \sim t_{n-2}$$

ya que en las condiciones anteriores t_0 es el cociente entre una variable normal estándar y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad. Claramente, t_0 es una variable pivotal (depende de β_1 pero su distribución es independiente de este parámetro), y puede ser fácilmente usada para contrastar hipótesis del tipo

$$H_0 : \beta_1 = \beta_{10}$$
$$H_1 : \beta_1 \neq \beta_{10}$$

con $\beta_{10} \in \mathbb{R}$ el valor que se desea contrastar para el parámetro. Esto es así ya que, bajo H_0, t_0 se distribuye según una distribución *t* con $n - 2$ grados de libertad. De hecho, nótese que, bajo H_0, t_0 se puede expresar en la forma

$$t_0 = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{se(\hat{\beta}_1)}$$

pues si H_0 es cierta se tiene que $E[\hat{\beta}_1] = \beta_1 = \beta_{10}$, y donde $se(\hat{\beta}_1)$, que se lee como *error estándar* (*standard error*) *de* $\hat{\beta}_1$, denota una estimación de la desviación típica de $\hat{\beta}_1$.

Este contraste procede entonces mediante los siguientes pasos:

i) Se calcula $t_0 = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MCR / \sum (x_i - \overline{x})^2}}$.

ii) Se calcula el percentil $\alpha/2$ superior de la distribución t_{n-2}, esto es, el valor $t_{n-2;\alpha/2}$.

iii) Si $|t_0| > t_{n-2;\alpha/2}$ se rechaza H_0 al nivel de significación $1 - \alpha$.

El p-valor asociado a este contraste se obtiene como $p = P(|t_{n-2}|>|t_0|)$.

De modo similar pueden contrastarse hipótesis relativas a β_0, que de manera genérica serán denotadas por

$$H_0 : \beta_0 = \beta_{00}$$
$$H_1 : \beta_0 \neq \beta_{00}$$

usando para ello el estadístico

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MCR\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)}} \sim t_{n-2}$$

y siguiendo los pasos anteriores.

Significatividad de la regresión y descomposición en suma de cuadrados

Un caso particular de gran importancia de los contrastes *t* anteriores es el dado cuando $\beta_{10} = 0$, esto es, cuando se contrastan las hipótesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Este contraste se relaciona con **la significatividad del modelo de regresión lineal**, en tanto que NO rechazar $H_0 : \beta_1 = 0$ implica que no hay evidencia para afirmar que existe una relación lineal entre las variables *X* e *Y* . Esto puede deberse tanto a la ineficacia de *X* como predictor/variable explicativa de *Y* (por ejemplo, cuando los puntos (x_i, y_i) se aproximan mediante una recta horizontal), como a que la relación entre estas dos variables es no lineal (por ejemplo cuando $Y = X^2$). Por el contrario, si es posible rechazar $H_0 : \beta_1 = 0$ esto implica que *X* es ***útil*** como predictor lineal de *Y*, aunque esto no garantiza necesariamente que el modelo lineal sea un *buen* modelo.

En el caso de la regresión simple, con una sola variable independiente *X*, este tipo de contraste sobre la significatividad del modelo lineal puede realizarse de dos maneras:

1) Mediante un contraste *t* sobre la hipótesis $H_0 : \beta_1 = 0$, usando $t_0 = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}$.

2) Mediante la técnica del análisis de la varianza (ANOVA, del inglés *Analisys Of Variance*).

El ANOVA se basa en particionar la variabilidad total de la variable respuesta *Y* mediante la **descomposición en sumas de cuadrados**, que enunciamos en la siguiente proposición.

Proposición 3: Se conoce como **Suma de cuadrados total** (SCT) a la suma de los cuadrados de las desviaciones de las observaciones y_i respecto de su media muestral $\overline{y}$, esto es,

$$SCT = \sum_{i=1,\dots,n} (y_i - \overline{y})^2 .$$

Además, se conoce como **Suma de cuadrados del modelo** (SCM, también conocida como Suma de cuadrados de regresión) a la suma de los cuadrados de las variaciones de los valores predichos $\hat{y}_i$ respecto de la media $\overline{y}$, es decir

$$SCM = \sum_{i=1,\dots,n} (\hat{y}_i - \overline{y})^2 .$$

Ya se ha introducido anteriormente la **Suma de cuadrados de los residuos** SCR, dada por

$$SCR = \sum_{i=1,\dots,n} (y_i - \hat{y}_i)^2 = \sum_{i=1,\dots,n} e_i^2 .$$

Entonces, se cumple la **igualdad *SCT* = *SCM* + *SCR***, esto es,

$$\sum_{i=1,\dots,n} (y_i - \overline{y})^2 = \sum_{i=1,\dots,n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1,\dots,n} (y_i - \hat{y}_i)^2 .$$

Demostración: $\sum_{i=1,\dots,n} (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) =$

$= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$ ya que $\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum \hat{y}_i (y_i - \hat{y}_i) - \overline{y} \sum (y_i - \hat{y}_i) = \sum \hat{y}_i e_i - \overline{y} \sum e_i$ y se tiene que $\sum \hat{y}_i e_i = 0$ y $\sum e_i = 0$ (resultados ya vistos como ejercicio).

Cada una de estas sumas de cuadrados tiene una interpretación relevante:

- **SCT**, la suma de cuadrados total, cuantifica la **variabilidad total** de las observaciones de la variable *Y* respecto a la media de estas observaciones. En ausencia de un modelo lineal como el dado por S1, el modelo base para *Y* sería el dado por $Y = \mu_Y + \varepsilon$, que se estimaría mediante mínimos cuadrados como $\hat{Y} = \overline{y}$. Los residuos de este modelo serían las diferencias $y_i - \overline{y}$, y la suma de sus cuadrados SCT representa de algún modo la variabilidad de *Y* que no puede ser explicada simplemente por su media.
- **SCM**, la suma de cuadrados del modelo, se interpreta como la parte de esa variabilidad total SCT no explicada por la media que **SI es explicada por el modelo de regresión**. Esto es así en tanto que SCM recoge las diferencias $\hat{y}_i - \overline{y}$ entre los valores predichos $\hat{y}_i$ por el modelo de regresión $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ y los valores predichos $\overline{y}$ por el anterior modelo base $\hat{Y} = \overline{y}$.
- **SCR**, la suma de cuadrados de los residuos, recoge la parte de la variabilidad total SCT que **NO es explicada por el modelo de regresión**, ya que se obtiene a partir de los residuos $e_i = y_i - \hat{y}_i$, esto es, la diferencia entre los valores realmente observados y_i y los ajustados por el modelo de regresión $\hat{y}_i$.

De este modo, la descomposición en suma de cuadrados anterior puede interpretarse diciendo que **la variabilidad total de las observaciones de la variable *Y* se divide en dos partes**:

- Una parte de **variabilidad explicada** a través de la variable *X* en el modelo de regresión (en tanto que las medias de cada y_i varían con x_i);
- Y otra parte de **variabilidad no explicada** por las variaciones de la variable *X* (y que se acumula en los residuos).

Contraste *F* de significatividad del modelo de regresión. Método ANOVA.

Claramente, cuanto mayor sea la parte de variabilidad explicada SCM en relación a la no explicada SCR, mayor será la significatividad del modelo de regresión. En este sentido, nótese

que cuando todas las observaciones están perfectamente alineadas se tiene que $\hat{y}_i = y_i$ para todo *i*, y por lo tanto ha de ser SCR = 0 y SCM = SCT. Y en el otro extremo, cuando $\hat{y}_i = \overline{y}$ para todo *i* (es decir, cuando la recta estimada es horizontal, $\hat{\beta}_1 = 0$) se cumple que SCM = 0 y SCR = SCT. Esto proporciona la intuición de que la comparación entre SCR y SCM es relevante de cara a juzgar la significatividad del modelo.

Por otro lado, aplicando el teorema de Cochran visto en el tema anterior se puede demostrar que, bajo la hipótesis $H_0 : \beta_1 = 0$ se cumple que

$$\frac{SCR}{\sigma^2} \sim \chi^2_{n-2} \quad \text{y} \quad \frac{SCM}{\sigma^2} \sim \chi^2_1$$

y ambas variables son independientes. Entonces, cuando $H_0 : \beta_1 = 0$ se tiene que

$$E\left[MCR = \frac{SCR}{n-2}\right] = \sigma^2 \quad \text{y} \quad E\left[MCM = \frac{SCM}{1}\right] = \sigma^2$$

y además podemos conocer la distribución del estadístico

$$F_0 = \frac{SCM/1}{SCR/(n-2)} = \frac{MCM}{MCR} \sim F_{1,n-2}\,.$$

Esto es, bajo $H_0 : \beta_1 = 0$ el estadístico F_0 es el cociente entre dos distribuciones chi-cuadrado independientes divididas por sus correspondientes grados de libertad, por lo que ese cociente se distribuye según una *F* con 1 grado de libertad en el numerador y *n* – 2 grados de libertad en el denominador.

Además, en tanto que bajo $H_0 : \beta_1 = 0$ se tiene que E[MCR] = E[MCM] = σ^2, cuando esta hipótesis H_0 es cierta se ha de esperar un valor de F_0 próximo a 1, mientras que si H_0 es falsa se esperaría que F_0 tome un valor bastante mayor que 1. De este modo, para contrastar las hipótesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

el método ANOVA procede mediante los siguientes pasos:

1) Se obtiene $F_0 = \frac{MCM}{MCR}$.

2) Se obtiene el percentil α superior de la distribución $F_{1,n-2}$, esto es, $F_{1,n-2;\alpha}$.

3) Si $F_0 > F_{1,n-2;\alpha}$ se rechaza $H_0 : \beta_1 = 0$ a nivel de significación $1 - \alpha$.

La información necesaria para realizar este contraste *F* típico del método ANOVA se suele organizar mediante una tabla como la siguiente, que proporcionan prácticamente todos los paquetes de software estadístico:

FUENTE DE VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F_0
REGRESIÓN	SCM	1	MCM	MCM/MCR
RESIDUOS	SCR	*n* - 2	MCR	
TOTAL	SCT	*n* - 1		

Figura 2.2: Tabla ANOVA del modelo de regresión simple.

Equivalencia de contrastes *t* y *F*

Así pues, hemos visto dos maneras de contrastar la hipótesis $H_0 : \beta_1 = 0$ de significatividad del modelo de regresión, una basada en un contraste *t* para el estimador $\hat{\beta}_1$ dado por el estadístico de contraste

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

y otra basada en el contraste *F* del método ANOVA, obtenido a través de la descomposición de la variabilidad total de los y_i en sumas de cuadrados, y dado por el estadístico de contraste

$$F_0 = \frac{MCM}{MCR}.$$

La pregunta obvia es entonces, ¿**son equivalentes estos contrastes *t* y *F***? La respuesta es SI. Para verlo, recordemos que

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MCR / \sum (x_i - \overline{x})^2}}$$

y si lo elevamos al cuadrado queda

$$t_0^2 = \frac{\hat{\beta}_1^2 \cdot \sum (x_i - \overline{x})^2}{MCR} = \frac{\hat{\beta}_1 \cdot \sum (x_i - \overline{x})(y_i - \overline{y})}{MCR} = \frac{MCM}{MCR} = F_0,$$

ya que $\hat{\beta}_1 \cdot \sum (x_i - \overline{x})(y_i - \overline{y})$ = *MCM* (**Ejercicio**).

En general, de hecho, se tiene que si $T \sim t_n$, entonces $T^2 \sim F_{1,n}$. Se puede comprobar además que $(t_{n;\alpha/2})^2 = F_{1,n;\alpha}$.

Por tanto, **ambos contrastes, el de la *t* y el de la *F*, son totalmente equivalentes**, aunque el contraste de la *t* es más flexible en tanto que permite contrastar otras hipótesis, como por ejemplo $H_0 : \beta_1 \geq 0$ o $H_0 : \beta_1 = \beta_{10} \neq 0$. La importancia del contraste *F* del ANOVA se verá con más claridad en el contexto del análisis de regresión múltiple, cuando se tiene más de una variable independiente, puesto que en ese caso los contrastes *t* y *F* ya no son equivalentes, y el único modo de contrastar la significatividad *global* del modelo de regresión será a través de un contraste *F* similar al anterior.

Un test alternativo es posible utilizando la correlación r_{XY} y el correspondiente parámetro poblacional

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V[X] \cdot V[Y]}}.$$

Si $\rho_{XY} \neq 0$ entonces *X* e *Y* están relacionados linealmente (puesto que su correlación lineal *poblacional* es no nula). Un contraste apropiado para las hipótesis

$$H_0 : \rho_{XY} = 0$$
$$H_1 : \rho_{XY} \neq 0$$

viene dado por el estadístico

$$t_1 = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1 - r_{XY}^2}} \sim t_{n-2}$$

de modo que se rechaza $H_0 : \rho_{XY} = 0$ cuando $|t_1| > t_{n-2;\alpha/2}$.

Nótese que rechazar $H_0 : \rho_{XY} = 0$ implica que hay evidencia para afirmar la existencia de una relación lineal entre *X* e *Y*, del mismo modo que al rechazar $H_0 : \beta_1 = 0$. Así que, de nuevo, la pregunta obligada es, ¿es este test equivalente a los anteriores? Y la respuesta de nuevo es SI (**Ejercicio**).

2.6 CONSIDERACIONES PRÁCTICAS

A pesar de que el uso de los modelos de regresión está muy extendido, son también muy frecuentes los malos usos o las aplicaciones incorrectas. Comentamos a continuación algunos **ejemplos de estas malas prácticas**:

- Los modelos de regresión han de ser entendidos como una **ecuación de interpolación en el rango observado de la variable independiente**. Hay que ser extremadamente prudente al usar el modelo fuera del rango observado.
- **La disposición de los valores x_i juega un papel importante en el ajuste por mínimos cuadrados**. En particular, la pendiente $\hat{\beta}_1$ de la recta de regresión depende fuertemente de la existencia de valores extremos en los x_i, ya que como hemos visto

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x}) y_i}{\sum_i (x_i - \overline{x})^2}.$$

  Estas situaciones a menudo requieren de algún tipo de acción correctiva, como investigar más en profundidad (o incluso eliminar) las observaciones inusuales, la utilización de un procedimiento de estimación robusta[2] o considerar la inclusión de otras variables explicativas. En general, hay que ser conscientes de que algunas observaciones (o un pequeño clúster) pueden tener una gran influencia, hasta el punto de controlar en la práctica algunas propiedades clave del modelo.
- **Los datos atípicos de la variable respuesta *Y* pueden distorsionar gravemente el ajuste por mínimos cuadrados**. Si estos puntos son realmente atípicos, la estimación de β_0 puede ser incorrecta y el estimador $\hat{\sigma}^2 = MCR$ puede inflar sustancialmente la estimación de σ^2.
- Es importante entender que solo porque un análisis de regresión indique una fuerte relación entre las observaciones de dos variables no se ha de concluir que una relación causal real entre esas variables. **Causalidad implica correlación, pero el recíproco no es necesariamente cierto**. Por ejemplo, dos conjuntos de datos monótonos casi siempre presentan correlación.
- En algunas aplicaciones, el valor de *X* con el que se ha de predecir *Y* es desconocido. Por ejemplo, al predecir el consumo eléctrico (*Y*) en función de la máxima temperatura diaria (*X*), es necesario un método previo de predicción para *X*. Por tanto, la predicción de *Y* es condicional a la predicción de *X*, y la calidad del modelo y su utilidad práctica dependen también de la precisión en la estimación de *X*, y no solo de la validez del modelo que relaciona *Y* con *X*.

2.7 REGRESIÓN LINEAL MÚLTIPLE

Vamos a estudiar ahora el caso general de modelo de regresión lineal en que existe **más de una variable independiente** con la que explicar/predecir la variable respuesta de interés *Y*. Esto es,

[2] Esto es, un método de ajuste alternativo al de mínimos cuadrado que conceda menos peso a estas observaciones inusuales, de modo que el modelo obtenido sea más *robusto* ante su existencia.

ahora se asumirá que se dispone de una colección $X_1,\ldots,X_k$, $k \geq 1$ de variables explicativas y, de manera similar a lo realizado en el modelo de regresión lineal simple, se buscará establecer un modelo

$$Y = f(X_1,\ldots,X_k) + \varepsilon$$

que relacione estas variables con la respuesta Y generalizando un conjunto de observaciones empíricas en que se han registrado todas estas variables para una muestra de individuos de la población de interés. Esto es lo que se conoce como un **modelo de regresión múltiple**.

Así pues, vamos a estudiar el ajuste y el análisis de estos modelos. Para ello, se generalizarán varias de las nociones y resultados del modelo de regresión lineal simple al contexto de la regresión lineal múltiple. Esto conllevará, por comodidad y generalidad, traducir y extender esas nociones y resultados en un lenguaje matricial, en lugar de basado en sumatorios como en el caso de la regresión simple.

En la práctica, los modelos de regresión múltiple permiten considerar conjuntos de factores que influyen en la respuesta, lo que favorece la obtención de modelos más realistas y útiles y amplía enormemente el abanico de posibilidades de cara a la aplicación real de estos modelos de regresión.

2.7.1 INTRODUCCIÓN. NOTACIÓN MATRICIAL

El modelo de regresión lineal múltiple parte del supuesto de que la función f que relaciona la variable respuesta Y con los regresores o variables independientes $X_1,\ldots,X_k$ asume una forma lineal, esto es,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

donde los parámetros $\beta_0, \beta_1, \ldots, \beta_k \in \mathbb{R}$ son constantes que caracterizan el modelo y se conocen como ***coeficientes de regresión***, y ε representa un término de error o ruido aleatorio. Nótese que, a excepción de la constante β_0, cada parámetro β_j está asociado a una variable explicativa X_j, $j = 1,\ldots,k$, recogiendo el **efecto *parcial*** de esa variable sobre la respuesta Y. Por tanto, el número de parámetros p es igual al número de variables más 1, esto es, $p = k + 1$.

Claramente, este supuesto generaliza el supuesto S1 introducido al comienzo de este tema, permitiendo considerar la influencia de más de un factor o variable explicativa sobre la respuesta Y. Por tanto, para valores de las variables explicativas en el rango de los datos observados, se asume que la ecuación lineal anterior proporciona una aproximación aceptable de la *verdadera* relación entre Y y los regresores $X_1,\ldots,X_k$. En otras palabras, se supone que Y es aproximadamente una función lineal de los regresores $X_1,\ldots,X_k$, y que el término de error ε, que mide la discrepancia entre esta aproximación y las observaciones, no contiene información sistemática para determinar Y que no haya sido ya capturada en los regresores.

Los datos para el ajuste de este modelo consisten en observaciones conjuntas $(x_{11},\ldots,x_{1k};y_1),\ldots,(x_{n1},\ldots,x_{nk};y_n)$ de la respuesta Y y las variables explicativas $X_1,\ldots,X_k$ en una muestra de n individuos de la población en estudio. Es decir, la observación del individuo i toma la forma $(x_{i1},\ldots,x_{ik};y_i)$, $i = 1,\ldots,n$, y al asumir que se cumple el supuesto S1 anterior se obtiene por tanto un sistema de ecuaciones

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i,\ i = 1,\ldots,n,$$

en el que, como en el caso de la regresión simple, el valor de los coeficientes $\beta_0, \beta_1, \ldots, \beta_k$ es el mismo en todas las ecuaciones, y por tanto son los términos de error ε_i los que aportan la

holgura (o grados de libertad) que permiten el tratamiento algebraico del problema (ya que los datos se consideran dados).

De cara a facilitar la presentación de este tratamiento algebraico, es conveniente traducir el anterior sistema de ecuaciones a notación matricial. Para ello, los parámetros, observaciones y errores anteriores se organizan de manera vectorial y matricial en la forma

$$\mathbf{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad , \quad \mathbf{X}=\begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{21} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{kn} \end{pmatrix}_{n\times p} \quad , \quad \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{p\times 1} \quad , \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

y es entonces directo comprobar que en términos de esta notación el sistema de ecuaciones anterior se puede expresar, de manera mucho más compacta, como

$$\mathbf{y}=\mathbf{X}\beta+\varepsilon\,.$$

Nótese que la matriz **X** contiene una columna de 1's en tanto que el parámetro β_0 no está asociado a ninguna variable. En lo que sigue se supondrá siempre que:

i) $n > p = k+1$, esto es, el número de observaciones es mayor que el número de parámetros;

ii) $rg(\mathbf{X}) = p$, esto es, la matriz **X** tiene rango máximo (por columnas). En otras palabras, las columnas de datos $\mathbf{x}_j=(x_{1j},....,x_{nj})^t$ asociadas a cada variable X_j, $j = 1,\ldots,k$, son linealmente independientes entre sí y en relación a la columna de 1's.

2.7.2 ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN

En tanto que el vector de coeficientes $\beta=(\beta_0,\beta_1,...,\beta_k)^t$ es a priori desconocido y no observable en la práctica habitual, será necesario introducir un método de estimación que permita obtener *buenas* aproximaciones de los parámetros β_j a partir de las observaciones $(x_{i1},...,x_{ik};y_i)$, con $i = 1,\ldots,n$ y $j = 0,\ldots,k$. Al igual que en el caso de la regresión simple, para esto se usará la estimación por mínimos cuadrados, consistente en obtener el valor $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k)\in\mathbb{R}^p$ del vector β que minimiza la suma de cuadrados de los errores, esto es, se trata de minimizar la función

$$S(\beta)=S(\beta_0,\beta_1,...,\beta_k)=\sum_{i=1}^{n}\varepsilon_i^2=\varepsilon^t\varepsilon=(\mathbf{y}-\mathbf{X}\beta)^t(\mathbf{y}-\mathbf{X}\beta)$$

para lo que se procede analíticamente tomando derivadas parciales respecto a cada parámetro β_j e igualando a cero, esto es,

$$\frac{\partial S}{\partial \beta_j}(\hat{\beta})=0\,,\ j=0,\ldots,k\,,$$

o equivalentemente, utilizando derivación matricial,

$$\frac{\partial S}{\partial \beta}(\hat{\beta})=\mathbf{0}$$

con $\mathbf{0}=(0,...,0)^t_{1\times p}$. Es sencillo comprobar que

$$\frac{\partial S}{\partial \beta}(\hat{\beta}) = -2\mathbf{X}^t\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\hat{\beta}\,,$$

por lo que al igualar a **0** se obtiene un sistema de $p = k + 1$ ecuaciones con p incógnitas dado por

$$\mathbf{X}^t\mathbf{X}\hat{\beta} = \mathbf{X}^t\mathbf{y}\,,$$

las cuales se conocen como ***ecuaciones normales***. Este sistema de ecuaciones normales admite entonces una única solución, dada por

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}\,.$$

Nótese que:

i) $\mathbf{X}^t\mathbf{X}$ es una matriz cuadrada pxp y simétrica (ya que $(\mathbf{X}^t\mathbf{X})^t = \mathbf{X}^t\mathbf{X}$), y

ii) Para que exista esta solución $\hat{\beta}$ del sistema de ecuaciones normales es necesario que exista la inversa $(\mathbf{X}^t\mathbf{X})^{-1}$, lo cual se cumple siempre que la matriz **X** tenga rango máximo por columnas, esto es, cuando $rg(\mathbf{X}) = p$.

A partir de aquí es directo generalizar las nociones de ***valores predichos*** y ***residuos*** del modelo de regresión, introducidas en el tema anterior. Así, los **valores predichos** por el modelo de regresión vienen dados por el vector

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

y los **residuos** se obtienen como la diferencia entre el vector de valores observados y el vector de valores predichos, esto es,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}\ .$$

Nótese que los vectores de valores predichos y de residuos pueden ser obtenidos como una transformación lineal de los valores observados **y**, ya que, al ser $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, se tiene que

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{H}\mathbf{y}$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

donde la matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ se denomina ***matriz sombrero*** o ***gorro*** (del inglés *hat matrix*) e **I** denota la matriz identidad nxn. Es sencillo comprobar que la matriz **H** es una matriz cuadrada nxn, simétrica e idempotente, ya que $(\mathbf{X}^t\mathbf{X})^{-1}$ es simétrica (puesto que $\mathbf{X}^t\mathbf{X}$ lo es) y se tiene que

$$\mathbf{H}\cdot\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{H}\,.$$

Ejercicio: Hallar una expresión para los elementos de las matrices $\mathbf{X}^t\mathbf{X}$ y $(\mathbf{X}^t\mathbf{X})^{-1}$.

<u>Interpretación geométrica de la estimación por mínimos cuadrados</u>

A veces ayuda tener una interpretación geométrica intuitiva de la estimación por mínimos cuadrados. Esta interpretación geométrica procede como sigue.

Obsérvese que el vector de observaciones $\mathbf{y}^t = (y_1, y_2, ..., y_n)$ define un vector en un espacio muestral n-dimensional, del origen al punto A, como se muestra en la Figura 2.3 de más abajo (en esta figura el espacio muestral es tridimensional). Por otro lado, la matriz **X** consiste en p vectores columna nx1, comenzando por el vector $\mathbf{1} = (1,\ldots,1)^t$ y continuando con los vectores columna $\mathbf{x}_j = (x_{1j},\ldots, x_{nj})^t$, también nx1, con las observaciones de cada variable X_j, $j = 1,\ldots,k$. Estos

p vectores definen un subespacio *p*-dimensional, llamado *espacio de estimación*. El espacio de estimación para *p* = 2 se muestra en la Figura 2.3 como un plano horizontal. En particular, se puede representar cualquier punto de este subespacio mediante una combinación lineal de los vectores **1**, $\mathbf{x}_1,\ldots,\mathbf{x}_k$. Así, cualquier punto del espacio de estimación tiene la forma **X**β, con β un vector *p*x1 cualquiera. Sea el punto B de la Figura 1 el determinado por el vector **X**β. La distancia de B a A elevada al cuadrado es

$$S(\beta)=(\mathbf{y}-\mathbf{X}\beta)^t(\mathbf{y}-\mathbf{X}\beta)$$

Así, para minimizar la distancia al cuadrado del punto A, definido por el vector de observación **y,** al espacio de estimación **X**β, se requiere determinar el punto del espacio de estimación que está más cercano a A. Esta distancia al cuadrado S(β) será mínima cuando el punto **X**β del espacio de estimación sea la proyección ortogonal de A sobre el espacio de estimación.

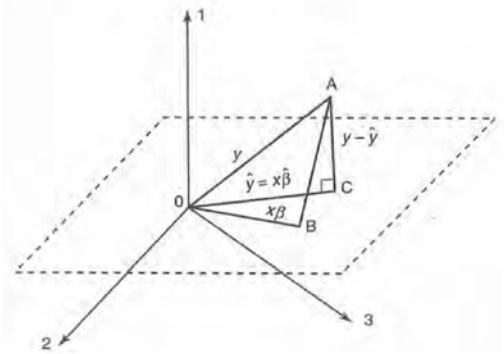

Figura 2.3: Interpretación geométrica de la estimación por mínimos cuadrados.

Ese punto proyectado en el espacio de estimación es el punto C de la Figura 2.3. Este punto se define con el vector $\hat{\mathbf{y}}=\mathbf{X}\hat{\beta}$. Por consiguiente, como el vector $\mathbf{y}-\hat{\mathbf{y}}=\mathbf{y}-\mathbf{X}\hat{\beta}$ es perpendicular al espacio de estimación, ha de cumplirse que

$$\mathbf{X}^t(\mathbf{y}-\mathbf{X}\hat{\beta})=0$$

es decir,

$$\mathbf{X}^t\mathbf{X}\hat{\beta}=\mathbf{X}^t\mathbf{y}$$

que son, de nuevo, las ecuaciones normales de mínimos cuadrados. En este sentido, nótese que la matriz **H** es la matriz de la proyección ortogonal sobre el espacio de estimación **X**β, ya que transforma el vector **y** en el vector proyectado $\hat{\mathbf{y}}=\mathbf{H}\mathbf{y}$.

2.7.3 PROPIEDADES DE LOS ESTIMADORES DE MÍNIMOS CUADRADOS

Las propiedades vistas en el caso de la regresión simple para los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ se pueden generalizar fácilmente al caso del vector de estimaciones $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k)$. Ahora, en lugar de estudiar y hallar los momentos poblacionales de cada estimador $\hat{\beta}_j$ por separado, *j* = 0,…,*k*, esto se hará de manera conjunta para todos los estimadores simultáneamente usando notación vectorial.

Antes de proceder a obtener estos momentos del estimador de mínimos cuadrados $\hat{\beta}$ es preciso hacer dos apuntes:

i) Al igual que en el tema anterior, no es posible obtener estos momentos solamente en base al supuesto S1, que establece la forma lineal del modelo $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. Es preciso imponer algún supuesto sobre los momentos del error *ε*, para luego poder trasladarlos a las observaciones **y** y al estimador $\hat{\beta}$. Este supuesto será el supuesto S2 ya introducido en el tema anterior, ahora sin modificaciones, aunque expresado en forma matricial:

$$E[\varepsilon] = \mathbf{0} \quad V[\varepsilon] = E[\varepsilon\varepsilon^t] = \sigma^2\mathbf{I}.$$

ii) Nótese que la varianza poblacional de un vector aleatorio de tamaño *p* es una matriz *p*x*p*, conocida como matriz de varianzas-covarianzas. Esta matriz contiene en la diagonal la varianza de cada variable aleatoria en el vector, y fuera de la diagonal contiene las covarianzas para cada par de variables aleatorias en el vector. Por tanto, la matriz de varianzas-covarianzas de un vector aleatorio es siempre simétrica, y también se puede demostrar que ha de ser semidefinida positiva.

Proposición 4: Esperanza y varianza de $\hat{\beta}$. Se tienen los siguientes resultados:

i) $E[\hat{\beta}] = \beta$

ii) $V[\hat{\beta}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$

Demostración:

i) $E[\hat{\beta}] = E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}] = E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\beta+\varepsilon)] = E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\beta + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\varepsilon)] = E[\beta] + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tE[\varepsilon] = \beta$

ii) $V[\hat{\beta}] = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^t] = E[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} - \beta)((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} - \beta)^t] =$

$= E[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\beta+\varepsilon) - \beta)((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\beta+\varepsilon) - \beta)^t] = E[(\beta + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\varepsilon - \beta)(\beta + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\varepsilon - \beta)^t] =$

$= E[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\varepsilon)((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\varepsilon)^t] = E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\varepsilon\varepsilon^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tE[\varepsilon\varepsilon^t]\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$

Así pues, por el primer apartado de la proposición anterior, se tiene que el estimador de mínimos cuadrados $\hat{\beta}$ es insesgado, y se cumple que $E[\hat{\beta}_j] = \beta_j$ para todo *j*. Además, si denotamos por $\mathbf{C} = (\mathbf{X}^t\mathbf{X})^{-1}$, se tiene que $V[\hat{\beta}_{j-1}] = \sigma^2 C_{jj}$, *j* = 1,…,*p* (recuérdese que *p* = *k*+1), y $Cov[\hat{\beta}_{i-1}, \hat{\beta}_{j-1}] = \sigma^2 C_{ij}$.

Teorema de Gauss-Markov: El estimador de mínimos cuadrados $\hat{\beta}$ es BLUE, esto es, es el *mejor* estimador lineal insesgado de β.

Demostración: Claramente $\hat{\beta}$ es un estimador lineal en las observaciones **y** (recuérdese que $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$), y acabamos de ver que también es insesgado. De cara a probar que es el *mejor* estimador de esta clase, tendríamos que probar que es el que tiene mínima varianza. Sin embargo, la varianza de $\hat{\beta}$, como acabamos de ver, no es un escalar sino una matriz, que no tiene mucho sentido minimizar. En lugar de esto, lo que se probará es que $\hat{\beta}$ minimiza la varianza de cualquier combinación lineal $v^t\hat{\beta}$ de los coeficientes estimados, donde *v* es cualquier vector de dimensión *p*.

Así pues, en primer lugar nótese que $V[v^t\hat{\beta}] = v^tV[\hat{\beta}]v = v^t\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}v = \sigma^2v^t(\mathbf{X}^t\mathbf{X})^{-1}v$ es un escalar. Sea entonces $\tilde{\beta}$ otro estimador lineal insesgado de β. Lo que se trata de demostrar es que $V[v^t\tilde{\beta}] \geq \sigma^2v^t(\mathbf{X}^t\mathbf{X})^{-1}v$, y que existe al menos un *v* tal que $V[v^t\tilde{\beta}] > \sigma^2v^t(\mathbf{X}^t\mathbf{X})^{-1}v$.

En segundo lugar, nótese que, dado que $\tilde{\beta}$ es un estimador lineal, puede ser escrito en la forma

$$\tilde{\beta} = [(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B}]\mathbf{y}$$

siendo **B** una matriz *pxn*, que ajusta el estimador de mínimos cuadrados para formar el estimador alternativo $\tilde{\beta}$. Como este estimador ha de ser insesgado, se tiene que

$$\beta = E[\tilde{\beta}] = E[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})\mathbf{y}] = ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})E[\mathbf{y}] = ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})\mathbf{X}\beta = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\beta + \mathbf{B}\mathbf{X}\beta = \beta + \mathbf{B}\mathbf{X}\beta$$

por lo que se tiene que cumplir que **BX** = **0**. Por otro lado, se tiene que la varianza de $\tilde{\beta}$ es

$$V[\tilde{\beta}] = V[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})\mathbf{y}] = ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})V[\mathbf{y}]((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})^t = \sigma^2((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})^t =$$

$$= \sigma^2((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{B})(\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} + \mathbf{B}^t) = \sigma^2((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{B}^t + \mathbf{B}\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}^t) =$$

$$\sigma^2((\mathbf{X}^t\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}^t) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} + \sigma^2\mathbf{B}\mathbf{B}^t = V[\hat{\beta}] + \sigma^2\mathbf{B}\mathbf{B}^t,$$

ya que **BX** = **0** y por tanto también $\mathbf{X}^t\mathbf{B}^t$ = **0**. En consecuencia, se tiene que

$$V[v^t\tilde{\beta}] = v^t(V[\hat{\beta}] + \sigma^2\mathbf{B}\mathbf{B}^t)v = v^tV[\hat{\beta}]v + v^t\sigma^2\mathbf{B}\mathbf{B}^tv = V[v^t\hat{\beta}] + \sigma^2v^t\mathbf{B}\mathbf{B}^tv.$$

Sea $w = \mathbf{B}^tv$. Entonces

$$v^t\mathbf{B}\mathbf{B}^tv = w^tw = \sum_{i=1\ldots p} w_i^2 > 0$$

siempre que $v \neq \mathbf{0}$, a menos que **B** = **0**. Por tanto, el estimador de mínimos cuadrados $\hat{\beta}$ es el *mejor* estimador lineal insesgado, como se quería demostrar.

2.7.4 ESTIMACIÓN DE σ^2 Y CONTRASTES *t* PARA COEFICIENTES INDIVIDUALES

Una vez se introduce el supuesto S3 ($\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$), como en el caso de la regresión simple, es posible estimar la varianza constante σ^2 del término de error ε *a* través de la Suma de Cuadrados de los Residuos (SCR).

Recordemos entonces que la Suma de Cuadrados de los Residuos (SCR) viene dada por

$$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^t\mathbf{e},$$

y que el estimador de σ^2 correspondiente se obtiene dividiendo SCR por el número de grados de libertad asociado a esta suma, que es el número de observaciones menos el número de parámetros estimados en el modelo, esto es, *n – p*:

$$\hat{\sigma}^2 = MCR = \frac{SCR}{n-p}.$$

Este estimador es insesgado, ya que se cumple que

$$\frac{SCR}{\sigma^2} \sim \chi^2_{n-p}, \text{ y por tanto } E[MCR = \frac{SCR}{n-p}] = \sigma^2.$$

Recordemos también que la Suma de Cuadrados del Modelo (SCM) viene dada por la expresión

$$SCM = \sum_{i=1,\ldots,n} (\hat{y}_i - \bar{y})^2$$

a partir de la cual se obtiene la Media de Cuadrados del Modelo (MCM) dividiendo SCM por sus grados de libertad, que se obtienen como el número de parámetros del modelo (*p*) menos 1 o, equivalentemente, como el número de variables explicativas del modelo, $k = p - 1$. Es posible demostrar que la esperanza de la MCM viene dada por

$$E\left[MCM = \frac{SCM}{p-1}\right] = \sigma^2 + \frac{1}{p-1}\beta_R^t \mathbf{X}_C^t \mathbf{X}_C \beta_R$$

donde $\beta_R^t = (\beta_1, ..., \beta_p)$ es el vector de parámetros asociados a las variables explicativas (esto es, sin el término independiente β_0) y se denota

$$\mathbf{X}_C = \begin{bmatrix} x_{11} - \overline{x}_1 & \cdots & x_{1p} - \overline{x}_p \\ \vdots & & \vdots \\ x_{n1} - \overline{x}_1 & \cdots & x_{np} - \overline{x}_p \end{bmatrix}$$

Además, bajo la hipótesis de que todos los parámetros asociados a variables explicativas son nulos, esto es, bajo $H_0 : \beta_1 = ... = \beta_p = 0$ o, expresado vectorialmente, $H_0 : \beta_R = \mathbf{0}$, se puede usar el teorema de Cochran para demostrar que la SCM sigue una distribución

$$\frac{SCM}{\sigma^2} \sim \chi^2_{p-1}$$

y es independiente de la SCR.

Contrastes *t* para coeficientes individuales

Una vez que se ha obtenido un estimador adecuado para σ^2, dado por $\hat{\sigma}^2 = MCR$, es entonces posible estimar la varianza del estimador de mínimos cuadrados $\hat{\beta}$. Así, como se tiene que

$$V[\hat{\beta}] = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1},$$

el estimador natural de esta varianza será

$$\widehat{V[\hat{\beta}]} = \hat{\sigma}^2 (\mathbf{X}^t \mathbf{X})^{-1} = MCR \cdot (\mathbf{X}^t \mathbf{X})^{-1}.$$

Obsérvese que, en particular, esto nos proporciona un estimador de la varianza de los estimadores individuales $\hat{\beta}_j$, $j = 0,...,k$, dado por

$$\widehat{V[\hat{\beta}_j]} = \hat{\sigma}^2 C_{j+1 j+1} = MCR \cdot C_{j+1 j+1},$$

donde, recordemos, C_{jj} denota el elemento diagonal *j*-ésimo de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$. Este resultado nos permite replicar el desarrollo de herramientas inferenciales para evaluar la significatividad del modelo que se estudió en el caso de la regresión simple.

Al ajustar un modelo de regresión múltiple, en el que un conjunto usualmente amplio de potenciales variables explicativas está disponible, averiguar qué variables regresoras, de entre todas las disponibles, son útiles de cara a explicar la variable respuesta *Y* es una tarea importante. Las hipótesis para contrastar si un regresor individual X_j es significativo son

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Cuando H_0 no es rechazada, suele ser aconsejable eliminar el regresor X_j del modelo. El estadístico de contraste para estas hipótesis es

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\widehat{V[\hat{\beta}_j]}}} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{j+1j+1}}} \sim t_{n-p} \ .$$

Así pues, la hipótesis nula H_0 se rechaza cuando $|t_0| > t_{n-p,\alpha/2}$.

Es importante ver que este contraste es un test parcial o marginal, en el sentido de que $\hat{\beta}_j$ depende de todos los demás regresores X_i ($i \neq j$) que se consideran en el modelo. En otras palabras, este es un contraste de la significatividad de X_j dadas el resto de variables regresoras que se encuentran en el modelo.

2.7.5 CONTRASTE GENERAL DE HIPÓTESIS LINEALES

Cualquier hipótesis que relacione linealmente los parámetros del modelo de regresión puede ser contrastada usando un marco teórico unificado. Este marco teórico se basa en el **principio de la suma de cuadrados extra**, que estudia la diferencia entre las sumas de cuadrados residuales de un **modelo completo** (MC) y un **modelo reducido** (MR), contenido en el anterior, de cara a contrastar si se obtiene una mejora al pasar del MR al MC. La intuición de algún modo es que el MR se obtiene a partir del MC descartando un subconjunto de sus variables, por lo que el MC será un modelo más adecuado a menos que las variables extra que incorpora no sean predictores útiles de la respuesta.

De manera general, se contrastarán las hipótesis

H_0: El modelo reducido es adecuado

H_1: El modelo completo es adecuado

utilizando las correspondientes sumas de cuadrados residuales SCR de ambos modelos,

$$SCR(MC) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e^t e \quad \text{y} \quad SCR(MR) = \sum_{i=1}^{n} (y_i - \hat{y}_i^*)^2 = e^{*t} e^*$$

donde $\hat{\mathbf{y}} = \mathbf{X}_{MC}\hat{\beta}_{MC}$ son los valores predichos por el modelo completo, caracterizado por la matriz de datos $\mathbf{X}_{MC}$ y por el vector de parámetros estimados $\hat{\beta}_{MC}$, y donde $\hat{\mathbf{y}}^* = \mathbf{X}_{MR}\hat{\beta}_{MR}$ denota a su vez los valores predichos por el modelo reducido, dado por la matriz $\mathbf{X}_{MR}$ y el vector de estimaciones $\hat{\beta}_{MR}$. La clave para realizar este contraste reside en el hecho de que el estadístico

$$F_0 = \frac{[SCR(MR) - SCR(MC)]/(p-r)}{SCR(MC)/(n-p)} \sim F_{p-r,n-p}$$

donde p es el número de parámetros del modelo completo y r el del modelo reducido, por lo que $p > r$. Así pues, H_0 se rechaza cuando $F_0 \geq F_{p-r,n-p,\alpha}$.

Nótese que $SCR(MC) \leq SCR(MR)$, ya que la incorporación de nuevas variables no puede conllevar un aumento de la suma de cuadrados residual, por lo que entonces la diferencia SCR(MR) – SCR(MC) representa el incremento en la suma de cuadrados residual cuando se descarta (en el MR) la información extra introducida en el MC respecto al MR.

Veamos a continuación algunos casos particulares de aplicación de este principio.

Contrastando si todos los coeficientes de regresión son cero

Un caso particular de gran importancia del anterior principio se obtiene cuando se contrasta la hipótesis de que *todas* las variables predictoras $X_1,\ldots,X_k$ en consideración no tienen poder

explicativo sobre la variable respuesta *Y*, y por tanto todos sus coeficientes $\beta_1,...,\beta_k$ deberían ser cero. Así pues, en este caso se tiene que

$$MR: Y = \beta_0 + \varepsilon$$
$$MC: Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$$

por lo que $r = 1$ y $p = k + 1$ son respectivamente el número de parámetros del modelo reducido y del modelo completo, y esta formulación del MR y el MC equivale al contraste de las hipótesis

$$H_0 : \beta_1 = ... = \beta_k = 0$$
$$H_1 : \exists \beta_j \neq 0 \, (j = 1,...,k)$$

Nótese que en este caso se tiene que

$$SCR(MC) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SCR$$

con $n - k + 1 = n - p$ grados de libertad. Además, como el estimador de mínimos cuadrados de β_0 en el MR es $\hat{\beta}_0^{MR} = \overline{y}$, y entonces $\hat{y}_i^* = \hat{\beta}_0^{MR} = \overline{y}$, la suma de cuadrados residual del MR viene dada por

$$SCR(MR) = \sum_{i=1}^{n} (y_i - \hat{y}_i^*)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 = SCT \, .$$

Por tanto, el estadístico de contraste F_0 toma en este caso la forma

$$F_0 = \frac{[SCR(MR) - SCR(MC)] / (p-1)}{SCR(MC) / (n-p)} = \frac{[SCT - SCR] / k}{SCR / (n-p)} = \frac{SCM / k}{SCR / (n-p)} = \frac{MCM}{MCR} \sim F_{k,n-p}$$

donde

$$SCM = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

es la suma de cuadrados del modelo de regresión o explicada, y MCM = SCM/*k* es la correspondiente media de cuadrados del modelo, y se ha usado el hecho de que la validez de la descomposición en suma de cuadrados SCT = SCM + SCR se mantiene en el caso de la regresión múltiple. Así pues, la hipótesis nula H_0 se rechazará cuando $F_0 \geq F_{k,n-p,\alpha}$.

Es importante notar que este contraste generaliza el contraste *F* introducido en el caso de la regresión simple. De hecho, el test recién introducido puede entenderse como un contraste ANOVA, con su correspondiente tabla dada como sigue:

FUENTE DE VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	F_0
REGRESIÓN	SCM	k	MCM = SCM/k	MCM/MCR
RESIDUOS	SCR	$n - p = n - k - 1$	MCR = SCR / $(n - p)$	
TOTAL	SCT	n - 1		

Figura 2.4: Tabla ANOVA del modelo de regresión múltiple.

Como en el caso de la regresión simple, este test contrasta la ***significatividad global del modelo***, esto es, si las variables introducidas en el modelo permiten *conjuntamente* una mayor capacidad explicativa que el modelo reducido sin ninguna variable. En tanto en el caso de la regresión simple estudiado en el tema anterior solo se contaba con una única variable explicativa,

contrastar la significatividad del modelo equivalía a contrastar si el parámetro β_1 asociado a esa única variable regresora es igual a 0 o no. En el caso de la regresión múltiple, el modelo será globalmente significativo si alguna de las variables introducidas tiene un parámetro asociado distinto de 0, por lo que contrastar la significatividad del modelo equivale en este caso a contrastar si alguno de los coeficientes de regresión $\beta_1,...,\beta_k$ es distinto de 0.

A pesar de estas semejanzas, es también importante resaltar que, a diferencia del contraste *F* estudiado para la regresión simple, este contraste de significatividad global en regresión múltiple **no** es equivalente a los contrastes *t* para coeficientes individuales. En particular, habrá *k* contrastes *t* individuales, algunos de los cuales pueden ser significativos (es decir, rechazarán H_0) y otros no serlo, pero solo hay un único contraste *F* global. Sin embargo, si al menos uno de los *k* contrastes *t* es significativo, entonces el contraste *F* global será también significativo. Además, el contraste *t* para un parámetro β_j (*j* = 1,…,*k*) en el modelo completo sí que es equivalente al contraste *F* obtenido al contrastar el modelo completo con un modelo reducido en el que solo se haya eliminado de MC la variable X_j, como veremos a continuación.

Contrastando si un subconjunto de coeficientes de regresión son cero

Supongamos que, de los *k* regresores disponibles $X_1,\ldots,X_k$, se tiene certeza de la capacidad explicativa de los primeros $k-l$ regresores $X_1,\ldots,X_{k-l}$ ($k > l$), y se desea contrastar la posible mejora que se obtendría al considerar alguno de los restantes regresores $X_{k-l+1},\ldots,X_k$. En esta situación, el vector de parámetros $\boldsymbol{\beta}$ se particiona como $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$, con $\boldsymbol{\beta}_1=(\beta_0,...,\beta_{k-l})$ y $\boldsymbol{\beta}_2=(\beta_{k-l+1},...,\beta_k)$, al igual que la matriz $\mathbf{X}=(\mathbf{X}_1\ \mathbf{X}_2)$, y se plantean los modelos

$$MR: Y=\beta_0+\beta_1X_1+....+\beta_{k-l}X_{k-l}+\varepsilon\equiv Y=\mathbf{X}_1\boldsymbol{\beta}_1+\varepsilon \text{ (}r=k-l+1=p-l\text{ parámetros)}$$

$$MC: Y=\beta_0+\beta_1X_1+....+\beta_kX_k+\varepsilon\equiv Y=\mathbf{X}\boldsymbol{\beta}+\varepsilon=\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\varepsilon \text{ (}p\text{ parámetros)}$$

Esto se corresponde con las hipótesis

$$H_0:\boldsymbol{\beta}_2=0$$
$$H_1:\boldsymbol{\beta}_2\neq 0$$

dado que se parte del modelo reducido $Y=\mathbf{X}_1\boldsymbol{\beta}_1+\varepsilon$ y se le añaden las variables asociadas a los parámetros en $\boldsymbol{\beta}_2$ para formar el modelo completo. Este contraste se lleva a cabo mediante el estadístico

$$F_0=\frac{[SCR(MR)-SCR(MC)]/l}{SCR(MC)/(n-p)}=\frac{[(\hat{\beta}^t\mathbf{X}^t-\hat{\boldsymbol{\beta}}_1^t\mathbf{X}_1^t)\mathbf{y}]/l}{[\mathbf{y}^t\mathbf{y}-\hat{\beta}^t\mathbf{X}^t\mathbf{y}]/(n-p)}=\frac{[(\hat{\beta}^t\mathbf{X}^t-\hat{\boldsymbol{\beta}}_1^t\mathbf{X}_1^t)\mathbf{y}]/l}{MCR}\sim F_{l,n-p}$$

donde $\hat{\beta}$ es la estimación de los parámetros del modelo completo, $\hat{\beta}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, y $\hat{\boldsymbol{\beta}}_1$ es la estimación de los parámetros del modelo reducido, $\hat{\boldsymbol{\beta}}_1=(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{y}$, por lo que se tendría que $SCR(MR)=\mathbf{y}^t\mathbf{y}-\hat{\boldsymbol{\beta}}_1^t\mathbf{X}_1^t\mathbf{y}$, y de ahí el resultado para F_0.

Es habitual usar en este contexto la notación $SCR(MR)=SCR(\boldsymbol{\beta}_1)$ y $SCR(MC)=SCR(\boldsymbol{\beta})$, y denotar $SCR(\boldsymbol{\beta}_2\mid\boldsymbol{\beta}_1)=SCR(MR)-SCR(MC)=SCR(\boldsymbol{\beta}_1)-SCR(\boldsymbol{\beta})$ para enfatizar que se mide el descenso en la suma de cuadrados residual al añadir los regresores asociados a los parámetros en $\boldsymbol{\beta}_2$ al modelo reducido, que solo considera los regresores asociados a los parámetros en $\boldsymbol{\beta}_1$. Es decir, se mide la suma de cuadrados residual *extra* con que cuenta el MR respecto al MC.

Ejemplo 1: Considerar el modelo completo $MC: Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$, y los modelos reducidos $MR_1: Y=\beta_0+\beta_1X_1+\varepsilon$ y $MR_2: Y=\beta_0+\beta_2X_2+\varepsilon$. Entonces, hay que observar que las

diferencias $SCR(\beta_2 \mid \beta_0, \beta_1)$ y $SCR(\beta_1 \mid \beta_0, \beta_2)$ permiten contrastar la significatividad de los descensos en la suma de cuadrados residual al pasar de MR_1 o MR_2 al MC, esto es, añadiendo justo el regresor que falta.

Nótese además que en este caso se tiene que $l = 1$ (y $k = 2$, luego $p = 3$), por lo que entonces $F_0 \sim F_{1,n-p} \sim (t_{n-p})^2$, es decir, los contrastes F correspondientes a considerar el paso de MR_1 o MR_2 al MC son de hecho equivalentes a los respectivos contrastes t para β_1 y β_2.

Ejemplo 2: Considérese el modelo cuadrático $MC: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \varepsilon.$ Entonces, hallando $SCR(\beta_1, \beta_2 \mid \beta_0)$ se podría contrastar la contribución de los términos de primer orden, mientras que calculando $SCR(\beta_{12}, \beta_{11}, \beta_{22} \mid \beta_0, \beta_1, \beta_2)$ se podría contrastar la contribución de los términos de segundo orden al modelo con solo los términos de primer orden.

Hipótesis lineales generales

De modo general, mediante el principio de la suma de cuadrados extra se pueden contrastar cualesquiera hipótesis de la forma $H_0: \mathbf{T}\beta = \mathbf{c}$, donde **c** es un vector mx1 de constantes y **T** es una matriz mxp tal que $rg(T)$ es máximo por filas. Nótese que lo anterior equivale a un sistema de m ecuaciones expresadas en términos de los parámetros en β .

El modelo completo se expresa en forma general como $MC: Y = \mathbf{X}\beta + \varepsilon,$ y el modelo reducido se obtiene a partir de MC usando la igualdad $\mathbf{T}\beta = \mathbf{c}$ para expresar m de los coeficientes del modelo MC en términos de los $r = p - m$ coeficientes restantes. Esto conduce a un modelo reducido $MR: Y = \mathbf{Z}\gamma + \varepsilon$, donde **Z** es una matriz nxm y γ es un vector de parámetros de dimensión r x 1. La estimación de γ viene entonces dada por $\hat{\gamma} = (\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{y}$. Obsérvese que $SCR(MC) = \mathbf{y}^t\mathbf{y} - \hat{\beta}^t\mathbf{X}^t\mathbf{y}$ y $SCR(MR) = \mathbf{y}^t\mathbf{y} - \hat{\gamma}^t\mathbf{Z}^t\mathbf{y}$, por lo que el estadístico de contraste F_0 toma la forma

$$F_0 = \frac{[SCR(MR) - SCR(MC)]/(p-r)}{SCR(MC)/(n-p)} = \frac{[(\hat{\beta}^t\mathbf{X}^t - \hat{\gamma}^t\mathbf{Z}^t)\mathbf{y}]/m}{[\mathbf{y}^t\mathbf{y} - \hat{\beta}^t\mathbf{X}^t\mathbf{y}]/(n-p)} = \frac{(\mathbf{T}\hat{\beta} - \mathbf{c})^t[\mathbf{T}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{T}^t]^{-1}(\mathbf{T}\hat{\beta} - \mathbf{c})/m}{MCR} \sim F_{m.n-p}$$

y se rechaza $H_0: \mathbf{T}\beta = \mathbf{c}$ cuando $F_0 > F_{m,n-p,\alpha}$. Nótese que el numerador de F_0 mide la distancia al cuadrado entre $\mathbf{T}\hat{\beta}$ y **c** estandarizada por la matriz de covarianzas de $\mathbf{T}\hat{\beta}$.

Ejemplo 3: **Contrastando la igualdad de coeficientes de regresión**

Sea el modelo completo $MC: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ ($p = 4$), y supongamos que se desea contrastar la hipótesis $H_0: \beta_1 = \beta_3$. Esta hipótesis puede escribirse en la forma $H_0: \mathbf{T}\beta = \mathbf{c}$ con $\mathbf{T} = [0,1,0,-1]_{1x4}$ y **c** = 0 (luego $m = 1$). Al sustituir $\beta_1 = \beta_3$ en MC se obtiene el modelo reducido dado por $MR: Y = \beta_0 + \beta_1(X_1 + X_3) + \beta_2 X_2 + \varepsilon = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \varepsilon$, con $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 = \beta_3$, $\gamma_2 = \beta_2$, $Z_1 = X_1 + X_3$ y $Z_2 = X_2$. Por tanto, el MR tiene $p - m = 4 - 1 = 3$ parámetros.

Ejemplo 4: Suponiendo el mismo modelo completo anterior, ¿qué hipótesis se contrastan si **c** = $(0,0)^t$ y $\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$? La respuesta es que entonces $H_0: \beta_1 = \beta_3, \beta_2 = 0$, con el modelo reducido $MR: Y = \gamma_0 + \gamma_1 Z_1 + \varepsilon$, en el que $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 = \beta_3$, $\beta_2 = 0$ y $Z_1 = X_1 + X_3$.

Ejemplo 5: Bajo el mismo modelo completo, considérese ahora la hipótesis $H_0: \beta_1 - \beta_3 = 2$. Entonces, ha de ser $\mathbf{T} = [0,1,0,-1]$ y **c** = 2, conduciendo al modelo reducido

$$MR: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + (\beta_1 - 2)X_3 + \varepsilon \Leftrightarrow Y + 2X_3 = \beta_0 + \beta_1 (X_1 + X_3) + \beta_2 X_2 + \varepsilon \Leftrightarrow$$

$$\Leftrightarrow Y' = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \varepsilon \text{ , con } Y' = Y + 2X_3 .$$

2.7.6 INTERVALOS Y REGIONES DE CONFIANZA

El análisis de intervalos de confianza sobre coeficientes de regresión individuales y sobre valores predichos/respuestas medias juega en el contexto de la regresión múltiple el mismo papel que en la regresión simple, esto es, el de extender las técnicas de estimación puntual de modo que reflejen mejor la incertidumbre asociada al proceso de estimación. Además, veremos cómo, en el caso de la regresión múltiple, es posible obtener regiones de confianza que delimiten *simultáneamente* los valores probables de *todos* los parámetros.

Intervalos de confianza para coeficientes de regresión individuales

Asumiendo que $\varepsilon_i \sim N(0, \sigma^2)$, i = 1,…,n, ya sabemos que entonces se tiene que las observaciones y_i y el vector de estimaciones $\hat{\beta}$ se distribuyen normalmente. Más concretamente, se tiene que el vector de estimadores se distribuye conjuntamente como $\hat{\beta} \sim N(\beta, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1})$, de donde es directo obtener la distribución marginal de cada estimador individual, $\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{j+1j+1})$, con C_{j+1j+1} el elemento diagonal j+1-ésimo de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$, j = 0,…,k. Por tanto, se obtiene fácilmente que

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{j+1j+1}}} \sim t_{n-p} \Leftrightarrow P\left(\hat{\beta}_j - t_{n-p,\alpha/2} \sqrt{\hat{\sigma}^2 C_{j+1j+1}} \le \beta_j \le \hat{\beta}_j + t_{n-p,\alpha/2} \sqrt{\hat{\sigma}^2 C_{j+1j+1}} \right) = 1 - \alpha .$$

Regiones e intervalos de confianza simultáneos para coeficientes de regresión

Es importante entender que los intervalos de confianza anteriores para coeficientes de regresión individuales son intervalos del tipo habitual “uno cada muestra”, en el sentido de que el nivel de confianza 1 – α especifica la probabilidad de que el intervalo contenga el valor real del parámetro estimado cuando se toman muestras aleatorias independientes y sucesivas y se construye la estimación por intervalo para cada una de esas muestras.

Sin embargo, es habitual que haya que construir *todos* los intervalos para los parámetros β_j (j = 0,…,k) a partir de *una única muestra*, es decir, habrá que hallar un conjunto de intervalos para los que el nivel de confianza 1 – α se aplique *simultáneamente*. Así, un conjunto de intervalos de confianza que conjuntamente contengan los valores reales de todos los parámetros con confianza 1 – α se conocen como **intervalos de confianza conjuntos o simultáneos**.

Ejemplo: Considérese el modelo de regresión simple $Y = \beta_0 + \beta_1 X + \varepsilon$, en el que se desea estimar β_0 y β_1 . Supongamos que disponemos de intervalos de confianza 1 – α = 0.95 para cada parámetro. Si estos intervalos fuesen independientes, la probabilidad de que contengan simultáneamente el valor de ambos parámetros es $0.95^2 = 0.9025$. Por tanto, no se tiene un nivel de confianza 1 – α respecto de ambos parámetros simultáneamente. De hecho, si los intervalos se construyen a partir de la misma muestra realmente no serán independientes, lo que introduce una complicación añadida para determinar el nivel de confianza real del conjunto de intervalos.

Atendiendo estas consideraciones, una posibilidad para obtener el nivel de confianza conjunto deseado, es optar por definir una **región de confianza conjunta** para los parámetros β a partir de la expresión

$$\frac{(\hat{\beta}-\beta)^t \mathbf{X}^t\mathbf{X}(\hat{\beta}-\beta)}{p\cdot MCR} \sim F_{p,n-p} \Leftrightarrow P\left(\frac{(\hat{\beta}-\beta)^t \mathbf{X}^t\mathbf{X}(\hat{\beta}-\beta)}{p\cdot MCR} \le F_{p,n-p,\alpha}\right) = 1-\alpha,$$

que describe una región de forma elíptica en el espacio de parámetros $\beta \in \mathbb{R}^p$ y centrada en el punto $\hat{\beta}$.

Otra aproximación, que conduce a la obtención de intervalos de confianza simultáneos, consiste en tratar de controlar el nivel de confianza conjunto de los intervalos, que tomarán la forma general $\hat{\beta}_j \pm \Delta se(\hat{\beta}_j)$, j = 0,…,k, eligiendo la constante Δ de manera que se obtenga un nivel de confianza 1 – α simultáneo. Existen diversos procedimientos para obtener esta constante Δ en función del nivel de confianza deseado. Uno de estos procedimientos es el obtenido a través de la desigualdad de Bonferroni: se toma $\Delta = t_{n-p,\alpha/2p}$, de manera que, para todo j = 0,…,k se obtienen intervalos $\hat{\beta}_j \pm t_{n-p,\alpha/2p} se(\hat{\beta}_j)$, que entonces cumplen que

$$P\left(\hat{\beta}_j - t_{n-p,\alpha/2p} se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{n-p,\alpha/2p} se(\hat{\beta}_j), \forall j = 0,\ldots,k\right) \ge 1-\alpha.$$

En general, el método de la elipse de confianza es más eficiente que la construcción de intervalos conjuntos, en el sentido de que, para un nivel de confianza dado, esta elipse suele ser más *pequeña* o estar contenida en el producto cartesiano de los intervalos de confianza. Sin embargo, los intervalos de confianza simultáneos son más fáciles de calcular y de interpretar cuando k es relativamente alto ($k \ge 3$).

Extrapolación oculta

Al predecir nuevas observaciones y/o respuestas medias para una configuración de las variables independientes $x_0 = (1, x_1^0, \ldots, x_k^0)^t$, es importante considerar si este punto x_0 se encuentra o no en la región del espacio de datos $\mathbb{R}^k$ que contiene las observaciones originales contenidas en la matriz **X** y con las que se ha estimado el modelo. Esto es relevante en tanto que determina si el modelo de regresión se utiliza como un mecanismo de interpolación o, por el contrario, de extrapolación.

En general, el modelo de regresión múltiple suele funcionar bien como interpolador, pero puede producir desviaciones graves al extrapolar fuera de la región que contiene a los datos observados. En la regresión múltiple es fácil extrapolar de manera inadvertida, ya que los rangos observados de cada regresor $X_1,\ldots,X_k$ no constituyen necesariamente una buena descripción de la región de datos observados, como se ilustra en la Figura 2 a continuación.

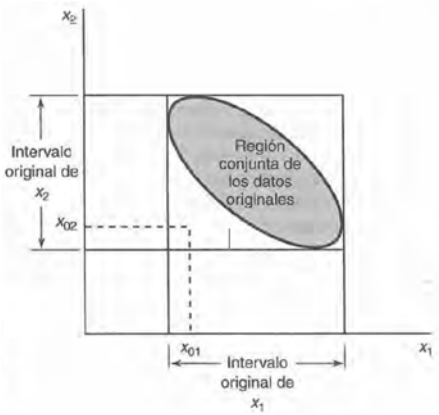

Figura 2.5: El producto cartesiano de los rangos observados puede no aproximar correctamente la región (típicamente elíptica) que contiene los datos observados.

Así pues, se define la **envolvente de regresores** como el mínimo conjunto convexo que contiene las n observaciones $x_i = (x_{i1}, ..., x_{ik})$, $i = 1,\ldots,n$. Cuando el punto x_0 está en la envolvente, el modelo se usa como interpolador, mientras que si está fuera se usa como extrapolador.

Los elementos diagonales de la matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ son útiles para detectar la extrapolación oculta, ya que el punto x_j, con $j = 1,\ldots,n$ tal que

$$\max_{i=1,\ldots,n} h_{ii} = h_{jj} \overset{notación}{=} h_{\max}$$

generalmente se encuentra en la frontera de la envolvente convexa. Se puede entonces demostrar que el conjunto $\left\{x \in \mathbb{R}^k \mid x^t(\mathbf{X}^t\mathbf{X})^{-1}x \le h_{\max}\right\}$ define un elipsoide que contiene a la envolvente de regresores. Así, si se denota $h_0 = x_0^t(\mathbf{X}^t\mathbf{X})^{-1}x_0$, se tiene que cuando $h_0 > h_{\max}$ el modelo se usa para extrapolar, mientras que si $h_0 < h_{\max}$ entonces x_0 se encuentra cerca o en el interior de la envolvente y el modelo se puede usar con mayor seguridad como interpolador.

2.7.7 COEFICIENTE DE DETERMINACIÓN.

El coeficiente de determinación R^2 se define exactamente igual que en el caso de la regresión simple, esto es, como el cociente entre la suma de cuadrados del modelo y la suma de cuadrados total,

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}$$

Además, la interpretación de esta ratio R^2 en el contexto de la regresión múltiple es también la misma, esto es, representa la proporción de variabilidad de la respuesta explicada por el modelo de regresión. Sin embargo, su utilidad como medida de bondad de ajuste de un modelo de regresión múltiple admite algunos matices, que pasamos a comentar a continuación.

Como hemos visto al estudiar el principio de la suma de cuadrados extra (Sección 2.7.5), al incorporar variables a un modelo de regresión la suma de cuadrados residual SCR tiende a decrecer (no puede crecer), o equivalentemente, la suma de cuadrados del modelo SCM tiende a crecer (nunca disminuye). Esto quiere decir que, aun cuando las nuevas variables no aportan realmente una gran capacidad explicativa de la respuesta respecto al modelo ya existente, el R^2 tiende a crecer al añadir variables al modelo, ya que crece la SCM mientras la SCT permanece constante. Pero este crecimiento del R^2 no implica necesariamente una mejora del modelo si las nuevas variables aportan poco respecto a las ya presentes. De hecho, esto puede crear la percepción engañosa de que la introducción de más y más variables se traduce en una mejora sucesiva del modelo.

Recordemos en este sentido que un *buen* modelo es aquél que captura los aspectos esenciales de la relación entre los regresores y la respuesta, logrando explicar una proporción relevante de la variabilidad de la respuesta *usando el menor número de regresores posible*. Así pues, en tanto que no permite realmente discriminar si su aumento se debe o no a la incorporación de una variable realmente significativa, suele desaconsejarse la utilización del R^2 para comparar la bondad de modelos con distinto número de variables.

Como alternativa, se propone el ***coeficiente de determinación ajustado***, que se define como

$$R^2_{adj} = 1 - \frac{SCR/(n-p)}{SCT/(n-1)} = 1 - \frac{MCR}{S_Y^2}$$

Nótese que este R^2_{adj} solo crece cuando decrece la MCR. En particular, el descenso de la SCR al introducir una nueva variable no necesariamente implica un crecimiento del R^2_{adj} si no es suficientemente amplio como para compensar el decrecimiento del factor $n - p$ (ya que p crece en una unidad al introducir una nueva variable).

Este coeficiente de determinación ajustado permite entonces una herramienta con la que controlar el **sobreajuste** (*overfitting*), un problema habitual de los modelos generados a partir de datos que consiste en que, una vez que se supera un determinado umbral en el número de parámetros, el modelo tiende a *aprender* o generalizar no solo la relación entre respuesta y variables independientes, sino también el ruido presente en los datos. En el caso de la regresión múltiple, las potenciales relaciones espúreas entre algunos regresores y el ruido contenido en los residuos de la respuesta (tras eliminar el efecto de los regresores útiles) pueden conllevar que la introducción de un número cada vez mayor de variables produzca la apariencia de una mejora en la capacidad explicativa del modelo, aunque simplemente se esté ajustando con cada vez mayor precisión ese ruido o error experimental que no es esencial a la relación entre respuesta y regresores. Otra manera de controlar este problema del sobreajuste se basa en la utilización de un conjunto de datos de *test* diferente a los datos de *entrenamiento* con que se ajusta el modelo para *validar* con mayor rigor su capacidad explicativa.