ÍNDICE

T2. ANÁLISIS DE REGRESIÓN LINEAL. ESTIMACIÓN

2.1 MODELO DE REGRESIÓN

2.1.1 Relación entre dos variables aleatorias

En este tema planteamos una regresión lineal entre dos variables aleatorias X e Y, ambas con distribución poblacional desconocida. El modelo es una recta que define la relación entre ambas variables mediante una pendiente que indica el efecto que tiene una variación de una unidad de X sobre Y.

- Cada variable tendrá una media poblacional desconocida, que serán características propias de las distribuciones poblacionales de X e Y.

- La pendiente de la recta que relaciona X con Y también será una característica desconocida de la distribución poblacional conjunta. El problema que pretende resolver este tema es estimar dicha pendiente a partir de los datos muestrales de ambas variables.

  - Si existe solo una variable X para explicar la variable Y, tenemos una **Regresión Lineal Simpe**, cuya expresión es:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

  - En caso de existir varias variables explicativas (X_1, X_2, X_3,…X_k) para explicar la variable Y, tenemos una **Regresión Lineal Múltiple** (Regresión Lineal Multivariante), cuya expresión es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

- La relación entre Y y las k variables explicativas será aproximada, no exacta. Tenemos que determinar cómo dar cabida al resto de factores no explícitos pero que afectan también a la variabe Y. Y además hay que determinar cuál es la forma funcional que relaciona a cada una de estas k variables con Y.

2.1.2 Modelización con dos o varias variables

Normalmente nos encontramos que la variable objeto de estudio Y, está relacionada no solo con una variable X, sino con varias variables X_1, X_2, X_3,…X_k, y entonces nuestro objetivo será explicar cómo varía Y ante cambios en alguna o algunas de las k variables explicativas.

La lista de variables k, con toda seguridad, no será una relación exhaustiva de las variables que explican el comportamiento de Y, de manera que la relación entre la Y y las k variables no será exacta o determinada, sino solo aproximada. Puesto que la relación solo puede ser aproximada, nos enfrentamos al problema de cómo dar cabida al resto de factores explícitos y que, sin embargo afectan a la variable Y y que no hemos tenido en cuenta con el resto de variables. También tenemos que determinar cuál es la forma funcional que relaciona cada una de estas k variables con Y.

La relación entre las k variables X y la variable Y se establece a través de la **función de esperanza condicionada** (FEC). La función de esperanza condicionada es ideal para resumir la relación entre la Y todas las k variables. Son varios los motivos por los que la relación queda bien resumida.

$$\mathrm{E}\left(Y\middle|X_1,X_2,...X_k\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

A pesar de que en general $\mathrm{E}\left(Y\middle|X_1,X_2,...X_k\right)$ es no lineal, vamos a considerar que se puede aproximar mediante una función lineal.

El modelo de regresión lineal tiene dos partes diferenciadas. La primera parte de la ecuación es la **función de regresión poblacional (FRP)** y define la relación entre Y y X que se cumple en promedio para la población.

$$\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

Por tanto, si conociéramos los valores de las k variables X, podríamos predecir utilizando la recta poblacional el valor esperado de Y. Esta ecuación define la relación promedio entre las variables a la derecha del signo igual (las X_j) y la variable Y. Esta relación promedio, que se expresa mediante la esperanza condicionada, es la función de regresión poblacional (FRP), que indica el valor esperado de la variable Y condicionado a los valores que toman las variables explicativas X_j.

$$\mathrm{E}\left(Y\middle|X_1,X_2,...X_k\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

Esta expresión nos indica que estamos considerando que la FEC es lineal y que

$$\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

es la mejor aproximación que podemos hacer de la FEC, es decir, de $\mathrm{E}\left(Y\middle|X_1,X_2,...X_k\right)$ pese a que esta no sea lineal.

Como resumen diremos que interpretamos el modelo de regresión lineal como una aproximación lineal a la FEC y esta caracterización lineal que se usa permite capturar el *efecto parcial* (efecto *ceteris paribus*) de cada una de las variables k sobre la Y.

Efecto parcial de X_j sobre Y: el coeficiente de la pendiente de X_j, o parámetro β_j, captura el efecto que X_j tiene sobre Y teniendo en cuenta (controlando) los otros factores explicitados en la relación. Esta interpretación se obtiene de la FRP, para una variación de X_j (ΔX_j), mientras el resto de variables se mantienen constantes, este cambio de X_j hará que cambie la Y en una cierta cantidad ΔY.

- Por ejemplo un cambio en X_1, hará que cambie Y en una cantidad ΔY, operando

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 + \ldots + \beta_k X_k \quad \Rightarrow \quad \Delta Y = \beta_1 (\Delta X_1)$$

Obtenemos: $$\beta_1 = \frac{\Delta Y}{\Delta X_1}$$

que indica que el coeficiente poblacional β_1 es el efecto (cambio esperado) sobre Y ante un cambio en X_1 manteniendo fijas X_j, con j = 1, 2, 3, … , k.

Entonces la siguiente expresión indica que cada coeficiente poblacional β_j es el efecto sobre Y ante un cambio en cada variable X_j , manteniendo fijas el resto de variables X.

Efecto parcial de X_j sobre Y: $$\beta_j = \frac{\Delta Y}{\Delta X_j}$$

El término β_0 no está multiplicado por ninguna variable X, entonces su interpretación es más sencilla: es el valor esperado de Y, cuando $X_1 = X_2 = \ldots = X_k = 0$.

Así en la expresión de un modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Intervienen distintos tipos de variables y tiene diferentes funciones en el modelo.

- La variable objeto de estudio es Y y se la suele denominar **variable dependiente**.
- Las variables X_j son las **variables explicativas** de Y.

En la siguiente tabla vemos distintos nombres con los que se puede hacer referencia a estas variables:

Y	X
Variable explicada	Variable explicativa
Variable dependiente	Variable independiente
Regresada	Regresora
Endógena	Exógena
Variable respuesta	Variable de control
Predicha	Predictora

- La variable ε se denomina **término de error** y representa todos los otros factores que además de X_1, X_2,..., X_k determinan el valor de la variable dependiente Y para una observación concreta que llamamos observación *i*, de manera que para cada observación *i* habrá un error ε_i. Este término de error es una forma de incluir el resto de factores no incluidos expresamente y que afectan a la variable regresada; por tanto tiene un papel crucial en el modelo de regresión y se tendrá que analizar su comportamiento para evaluar el modelo en todo su conjunto.

Tanto el modelo de regresión múltiple como el de regresión simple contemplan relaciones lineales.

El caso de regresión lineal simple:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Esta ecuación es una ecuación en la que el término lineal, geométricamente hablando, se refiere a que la relación entre las variable explicativa, la variable dependiente y los parámetros es una recta.

Los modelos de regresión pueden ser lineales en las variables o lineales en los parámetros:

- **Lineales en las variables**: Para ser lineal en las variables, las variables X_j no puede estar elevadas a una potencia diferente de la unidad; tampoco puede estar ni multiplicado ni divididas por otra variable.
    - Ejemplo modelos **lineales en las variables**:

      $$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

      $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

    - Ejemplo modelos **no lineales en las variables**:

      $$Y = \beta_0 + \beta_1 X_1^2 + \varepsilon$$

      $$Y = \beta_0 + \beta_1 \left(1 / X_1\right) + \varepsilon$$

      Los modelos **no lineales en las variables se pueden linealizar** aplicando el cambio de variable apropiado ($Z_1 = X_1^2$ ó $Z_1 = 1/X_1$) en los respectivos ejemplos.

- **Lineales en los parámetros:** cuando los coeficientes β_j están multiplicados por las variables o por alguna transformación de estas, pero sin estar multiplicándose o dividiendo entre ellos, es decir sin que exista ninguna interacción entre los diferentes parámetros del modelo.

  Un modelo es no lineal en los parámetros cuando algún β_j aparece elevado a cualquier potencia distinta de la unidad o multiplicado o dividido por otro parámetro.

Entonces la no linealidad puede aparecer por dos motivos:

- No linealidad en las variables
- No linealidad en los parámetros.

La no linealidad en las variables se puede resolver mediante un cambio de variable y transformar el modelo en lineal. Sin embargo, la no linealidad en los parámetros no es posible de resolver. Por este motivo, el término lineal se utiliza solo para hablar de la linealidad en los parámetros.

2.2 MÍNIMOS CUADRADOS ORDINARIOS (MCO)

Una vez analizada la función de regresión poblacional (FRP), ahora el objetivo es encontrar la **función de regresión muestral** (FRM). Para ello necesitamos estimar los parámetros que aparecen en la relación poblacional. Se desea estimar los coeficientes β_j, para poder interpretar los efectos parciales de cada variable explicativa en el comportamiento esperado de la variable dependiente.

Para poder estimar los parámetros poblacionales necesitaríamos tener toda la información poblacional, pero solo tenemos los datos obtenidos de la muestra y con ellos solo podemos estimar la función de regresión muestral (FRM). En la medida en que la muestra sea más representativa de la población a estudio, los valores estimados serán más cercanos a los verdaderos valores poblacionales.

2.2.1 Regresión Simple

Supongamos que queremos estimar la función de regresión poblacional (FRP) del modelo de regresión lineal poblacional de:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

donde i recorre las n observaciones disponibles. Para cada observación de tenemos un valor observado Y_i y X_i (no importa si los datos proceden de una sección cruzada (transversal), subíndice i, o de una serie temporal, subíndice t). Queremos estimar los parámetros de la FRP, desconocidos y para ello utilizaremos los datos disponibles de ambas variables (X , Y).

Los coeficientes se estiman por alguna técnica estadística y serán los homólogos muestrales de los coeficientes poblacionales,

con ellos explicaremos la función de regresión muestral (FRM):

$$\hat{\beta}_0 + \hat{\beta}_1 X_{1i}$$

que es la homóloga de la función de regresión poblacional (FRP):

$$\beta_0 + \beta_1 X_1$$

A partir de la FRM puedo obtener, el valor estimado de Yi, valor de predicción a partir de la recta de regresión estimada, según el valor que toma X_{1i}

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$$

La diferencia entre el valor observado y el valor estimado o valor previsto es el **residuo de la regresión**, homólogo muestral del término poblacional error ε_i .

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}$$

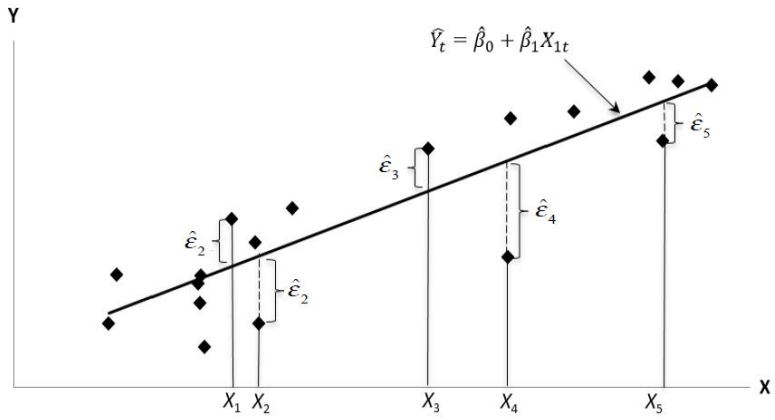
$$Y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_{1i}}_{\hat{Y}_i(\text{valor estimado})} + \underbrace{\hat{\varepsilon}_i}_{\text{residuos estimados}} = \hat{Y}_i + \hat{\varepsilon}_i$$

La técnica de los mínimos cuadrados ordinarios, MCO, permite estimar los parámetros o coeficientes que minimizan el cuadrado de la suma de las discrepancias que se producen entre los valores observados y los valores estimados (valores de predicción), es decir encuentra, para nuestra muestra, aquellos valores de los coeficientes que minimizan la expresión:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

$$\min \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} \right)^2$$

A continuación se muestra la recta de regresión que minimiza la suma de los cuadrados de los residuos y pueden observarse los errores ε_i en cada punto.



Si seleccionamos los valores paramétricos que minimizan la suma de los cuadrados de los residuos, se impide que los valores positivos (por encima de la recta) se compensen con los negativos (por debajo de la recta). Matemáticamente este problema se resuelve derivando e igualando a cero la expresión y eso nos lleva a las denominadas **ecuaciones normales**:

$$\frac{\partial\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)}{\partial\hat{\beta}_0}=-2\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}\right)=0;\quad \frac{\partial\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)}{\partial\hat{\beta}_1}=-2\sum_{i=1}^{n}X_{1i}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}\right)=0$$

Ecuaciones normales:

$$\sum_{i=1}^{n}\hat{\varepsilon}_i=\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}\right)=0$$

$$\sum_{i=1}^{n}X_{1i}\hat{\varepsilon}_i=\sum_{i=1}^{n}X_{1i}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1i}\right)=0$$

Las ecuaciones normales permiten calcular (dos ecuaciones con dos incógnitas) los parámetros de la regresión.

- La pendiente de la regresión:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_{1i}y_i}{\sum_{i=1}^{n}x_{1i}^2}=\frac{\sum_{i=1}^{n}\left(X_{1i}-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_{1i}-\bar{X}\right)^2}=\frac{\widehat{\mathrm{cov}}\left(X_1,Y\right)}{\widehat{\mathrm{var}}(X_1)}$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}X_{1i}Y_i}{\sum_{i=1}^{n}X_{1i}^2}=\frac{S_{X_1,Y}}{S_{X_1}^2}$$

Sumando ambas partes de la ecuación desde *i* = 1 hasta *n* y dividiendo por *n*, y teniendo en cuenta que el sumatorio de los errores es nulo (primera ecuación normal) obtenemos:

$$\bar{Y}=\hat{\beta}_0+\hat{\beta}_1\bar{X}_1$$

Esta ecuación indica que uno de los puntos por donde pasa la recta estimada coincide con las medias muestrales de las variables

- La constante u ordenada de la regresión:

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}_1$$

Resultados algebraicos de la regresión en el modelo de regresión simple

- Por la primera ecuación normal, deducimos que la **media de los residuos** estimados es **nula**.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

- A partir de: $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$

  deducimos que: $\overline{Y} = \overline{\hat{Y}} + \overline{\hat{\varepsilon}} = \overline{\hat{Y}}$

- A partir de: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\varepsilon}_i$ y $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1$

  Restando: $y_i = \hat{\beta}_1 x_{1i} + \hat{\varepsilon}_i$

  donde: $y_i = Y_i - \overline{Y}$ y $x_{1i} = X_{1i} - \overline{X}_1$

  De manera que, en desviaciones a las medias, la variable estimada por la regresión es $\hat{y}_i = \hat{\beta}_1 x_{1i}$

  Multiplicando ambas partes por los errores estimados y sumando desde i=1 hasta n y con la segunda ecuación normal se obtiene:

$$\sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i = \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} \hat{\varepsilon}_i = 0$$

- De manera que la covarianza $\text{cov}(\hat{Y}_i, \hat{\varepsilon}_i) = 0$. Y a partir de la segunda ecuación normal tenemos que:

$$\text{cov}(X_{1i}, \hat{\varepsilon}_i) = 0$$

  Entonces la variable independiente X_{1i} y los residuos $\hat{\varepsilon}_i$ están incorrelados.

- A partir de: $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$. La varianza es:

$$\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(\hat{\varepsilon}_i)$$

Coeficiente de Determinación R^2

El Coeficiente de Determinación es una medida estadística de bondad de ajuste o fiabilidad del modelo estimado a los datos. Este coeficiente indica cuál es la proporción de la variación total en la variable dependiente Y, que es explicada por el modelo de regresión estimada, es decir, mide la capacidad explicativa del modelo estimado.

Sabemos: $\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(\hat{\varepsilon}_i)$

Entonces el R^2 se define como la proporción de la varianza explicada por la regresión respecto de la varianza que queremos explicar:

$$R^2 = \frac{\text{var}(\hat{Y})}{\text{var}(Y)} = \frac{\text{var}(Y) - \text{var}(\hat{\varepsilon})}{\text{var}(Y)} = 1 - \frac{\text{var}(\hat{\varepsilon})}{\text{var}(Y)} \qquad (0 \le R^2 \le 1)$$

Este coeficiente siempre es positivo y menor o igual que 1.

- Si R^2 = 1, la regresión explica completamente la variación de la variable dependiente, es decir, todas las observaciones estarían sobre la recta de regresión.
- Si R^2 = 0, la regresión no explicaría nada sobre el comportamiento de la variable dependiente.

El R^2 multiplicado por 100 se interpreta como el porcentaje de la variable dependiente explicado por la regresión. En ningún caso un alto coeficiente de determinación garantiza que el modelo de regresión tenga necesariamente buenas características.

El coeficiente de determinación es igual al coeficiente de correlación lineal al cuadrado.

$$R^2 = (r_{XY})^2$$

$$r_{XY} = \frac{Cov(X,Y)}{S_X S_Y} = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2][n\sum Y_i^2 - (\sum Y_i)^2]}}$$

- Si r_{XY} = 1 tendremos una relación perfecta directa entre X e Y.
- Si r_{XY} = 0 no existirá relación lineal entre X e Y pero podrá existir otro tipo de relación no lineal entre ellas.
- Si r_{XY} = -1 tendremos una relación lineal inversa perfecta entre X e Y.
- Cuanto más cerca este el valor del coeficiente de los valores 1 ó -1 más intensa será la relación lineal que existe entre X e Y.

De la definición de varianza se deduce:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\frac{1}{n}SCT=\frac{1}{n}SCE+\frac{1}{n}SCR,$$

Donde:

- SCT es la suma cuadrática de la variable dependiente en desviaciones a las medias.
- SCE es la suma cuadrática de la variable estimada en desviaciones a las medias.
- SCR es la suma cuadrática de los residuos estimados.

Multiplicando por n a ambos lados tenemos que:

$$SCT=SCE+SCR$$

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Y el R^2, en términos de sumas cuadráticas, es:

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

Ejemplo 1: Demanda de tabaco

La ley de demanda nos dice que la cantidad demandada depende inversamente del precio del bien. El modelo de regresión simple se planeta de la siguiente forma:

$$cantidad = \beta_0 + \beta_1 precio + \varepsilon$$

disponemos de observaciones de precios medios de la cajetilla de 20 cigarrillos en euros y también del número de cajetillas consumidas por trimestre de la población española. La variable de cantidad se ha dividido por la población de cada año, por tanto estamos hablando de cajetillas de tabaco consumidas per cápita. La variable independiente está deflactada por el índice de precios al consumo (2005=1), de manera que el precio está en euros constantes de 2005. La regresión estimada es:

$$\widehat{(tabaco_t/pob_t)} = 33,48 - 8,98 \cdot (precio_t/ipc_t),$$
$$n = 32, R^2 = 0,906.$$

La interpretación de la regresión es clara, por ejemplo, a un precio de 2 euros el modelo predice un consumo medio de 16 cajetillas (33,48-8,98x2 =15,52) por persona y trimestre. Si el precio se incrementa en un euro, ceteris paribus, el consumo se reduce en 9 (8,98) unidades. El R^2 = 0, 906 indica que la regresión explica el 90 % de la varianza del consumo per cápita de tabaco.

2.2.2 Interpretación de los coeficientes del modelo: cambios de escala y relaciones no lineales

2.2.2.1 Cambios de escala

Son cambios en las unidades de medida. Se pueden representar como:

$$w_1 Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 w_2 X_{1i} + \tilde{\varepsilon}_i,$$

Donde w_1 es el cambio de escala de la variable dependiente y w_2 el cambio de la variable independiente.

Se puede comprobar que la pendiente se ve afectada por los cambios de escala de ambas variables. El término constante y los errores estimados solo quedan afectados por el cambio de escala de la variable explicada.

$$\tilde{\beta}_1 = \frac{w_1}{w_2}\hat{\beta}_1$$
$$\tilde{\beta}_0 = w_1\hat{\beta}_0,$$
$$\tilde{\varepsilon}_i = w_1\hat{\varepsilon}_i.$$

Ejemplo 2: Los salarios de alta dirección de las empresas españolas

A partir de la media de los salarios anuales (en miles de euros) de alta dirección de las empresas que cotizaban en el IBEX en 2010, y de los beneficios de las empresas (en millones de euros) de ese mismo año, nos planteamos si los salarios de alta dirección están relacionados con los beneficios de la empresa. El modelo poblacional es:

$$salario = \beta_0 + \beta_1 beneficios + \varepsilon,$$

La estimación de la FRM: $\widehat{salario}_i = 296,362 + 0,267993 \cdot beneficios_i,$

$$n = 31, R^2 = 0,786,$$

- El modelo explica el 78,6 % de los salarios. Un incremento de un millón de euros en los beneficios provoca un incremento de 0,268 miles de euros en los salarios.

➢ Si estimamos los beneficios en miles de millones:

$$\widehat{salario}_i = 296,362 + 267,99 \cdot (beneficios_i/1000),$$
$$n = 31, R^2 = 0,786.$$

- Se ha hecho un cambio de escala (dividir por 1000) en la variable independiente (beneficios), esto provoca una modificación de la pendiente de la variable beneficios, ésta queda multiplicada por mil, el incremento de mil millones de euros de beneficios produce un

incremento salarial de 267,99 miles de euros. El término constante y el coeficiente R^2 comprobamos que no varía.

- Si expresamos la variable dependiente en euros y los beneficios en millones entonces:

$$\widehat{(salario_i \cdot 1000)} = 296362 + 267,99 \cdot (beneficios_i),$$

El término constante se ha multiplicado por mil (296,362x1000 = 296.362), la pendiente se ha multiplicado por mil respecto de la expresiones anteriores. El modelo prevé un sueldo medio de alta dirección de 296.362 euros, y además nos indica que en promedio el sueldo aumenta 267,99 euros por cada millón de euros de beneficios.

2.2.2.2 Forma funcional

Modelo de regresión es flexible y contempla relaciones no lineales. Los modelos de regresión no lineales en las variables los podemos linealizar mediante cambios de variable y el muy habitual realizar transformaciones de las variables en econometría.

Las transformaciones más comunes son:

- Modelos logarítmicos o de elasticidad constante (log-log)
- Modelos semilogarítmicos:
  - Logarítmicos lineales (log-nivel)
  - Lineales logarítmicos (nivel-log)
- Modelos recíprocos

Existen modelos que no son lineales pero que si son intrínsecamente lineales, esto significa, que mediante una transformación matemática se podrían convertir en lineales. En Econometría se utiliza frecuentemente la transformación logarítmica para conseguir modelos del tipo **log-log** o de **elasticidad constante**, cuya ventaja es que los coeficientes de las variables se pueden interpretar directamente como las elasticidades.

Por ejemplo, cuando la relación entre las variables es exponencial del tipo:

$$Y = \beta_0 X^{\beta_1} e^{\varepsilon}$$

si tomamos logaritmos y operamos, entonces se puede expresar como:

$$\ln Y = \ln \beta_0 + \beta_1 \ln X + \varepsilon = \alpha_0 + \beta_1 \ln X + \varepsilon$$

puesto que ln β_0 es una constante podemos hacer el cambio (ln β_0= α_0) y así tenemos un modelo lineal a partir de un modelo que inicialmente no lo era y ya podríamos aplicar MCO.

- Análisis del modelo logarítmico (log-log)

En los modelos log-log todas las variables están transformadas logarítmicamente,

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

si ahora consideramos la variación en lnY ante un cambio en la variable en lnX, diferenciando tenemos:

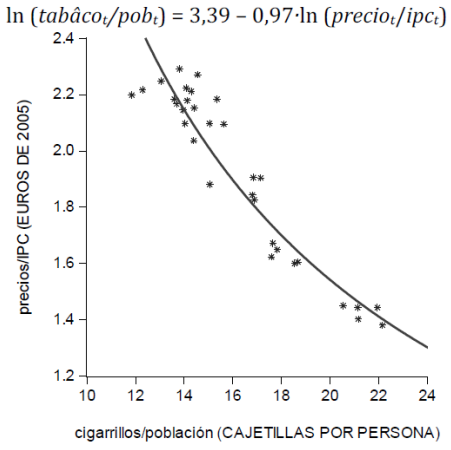
$$\frac{dY}{Y} = \beta_1 \frac{dX}{X} \qquad \text{y operando:} \qquad \beta_1 = \frac{\frac{dY}{Y}}{\frac{dX}{X}} \approx \frac{\frac{\Delta Y}{Y}}{\frac{\Delta X}{X}}$$

de manera que cuando las variables están en logaritmos la pendiente es directamente la elasticidad media:

$$\Delta Y\% = \beta \Delta X\%$$

En este modelo una variación de un 1% en la variable explicativa está asociada con una variación en la variable dependiente de un $\beta\%$

Ejemplo 3: Demanda de tabaco



MODELO: $\ln (tabâco_t/pob_t) = 3,39 - 0,97 \cdot \ln (precio_t/ipc_t)$

donde $\hat{\beta}_1$ = - 0,97

INTERPRETACIÓN: Si el precio crece un 1% la cantidad consumida disminuye un 0,97%

- Análisis del modelo logarítmico lineal (log-lin o log-nivel)

Cuando la variable endógena Y está en logaritmos y la variable explicativa X está en niveles. Su forma general es:

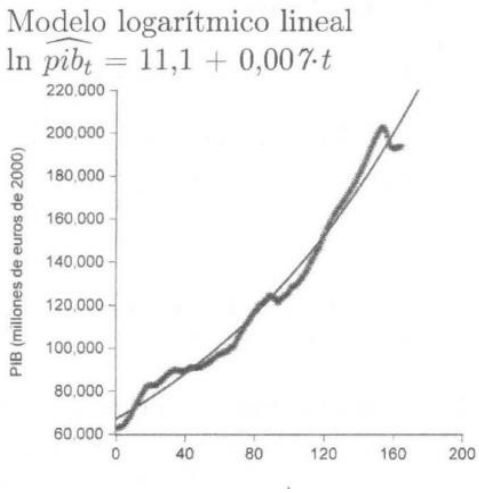
$$\ln Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

donde la pendiente β_1 multiplicada por 100 es aproximadamente la tasa porcentual de cambio de la variable dependiente (semielasticidad):

$$\Delta Y\% = 100 \beta \Delta X$$

Si X cambia una unidad (cambio unitario), este cambio está asociado a un cambio de100x $\beta\%$

Ejemplo 4: Crecimiento de la economía española



MODELO: $\widehat{\ln pib_t} = 11,11444 + 0,006833 \cdot t$

INTERPRETACIÓN: la tasa de variación en estos modelos es la pendiente multiplicada por 100, en este caso la tasa de variación es 0,6833% (100x0,006833)

- Análisis del modelo lineal logarítmico (lin-log o nivel-log)

Cuando la variable dependiente Y está en niveles y la independiente X en logaritmos. Su forma general es:

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

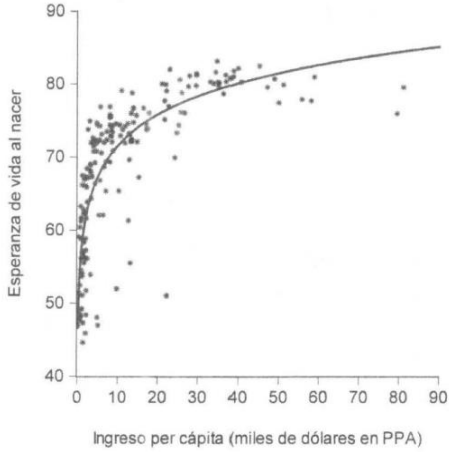
donde la pendiente β_1 dividida por 100 es la tasa porcentual de cambio de la variable explicada:

$$\Delta Y = (\beta/100)\Delta X\%$$

Si cambia X en 1%, es decir, si ΔX/X=0,01, entonces dicho cambio tiene asociado en este modelo una variación en Y de 0,01x β_1

Ejemplo 5: Esperanza de vida e ingresos

Ajuste lineal logarítmico: $\widehat{esperanza}_i = 57{,}27 + 6{,}197 \cdot (\ln\ ingresos_i)$



MODELO: $\widehat{esperanza}_i = 57,27 + 6,197 \cdot (\ln ingresos_i)$

INTERPRETACIÓN: un incremento de un 1% en los ingresos per cápita (PPA) propicia un incremento de 0,06197 (6,197/100) años en la esperanza de vida.

- Análisis del modelo recíproco

Se conoce como modelo recíproco aquel en el que la variable independiente aparece en su forma inversa:

$$Y = \beta_0 + \beta_1 \left(1 / X_1\right) + \varepsilon$$

A medida que aumenta X, la variable independiente disminuye 1/X, y en el límite se va acercando a cero, momento en el que la variable explicada Y se hace igual al término constante. Este tipo de modelos tiene sentido cuando la variable dependiente tiene límite asintótico β_0

Ejemplo 6: Mortalidad infantil y años de estudio

Datos de mortalidad infantil por cada cien mil habitantes y años de estudios en promedio de 185 países.

$$\widehat{mortalidad} = -1,56 + 292,78\ (1/estudios)$$

INTERPRETACIÓN: a medida que aumentan los años de estudios disminuye la tasa de mortalidad infantil, si los años de estudio son igual a uno, entonces el modelo predice una tasa de mortalidad por cien mil de 291,22 (292,78-1,56).

La forma funcional en los modelos de regresión simple es fácil de determinar, el problema radica cuando se introducen más variables (regresión lineal múltiple), ya que para determinar la forma funcional de estos modelos más complejos es útil basarse en la teoría económica o calcular las tasas de cambio y las elasticidades para ver cuál sería la forma funcional más adecuada.

FORMAS FUNCIONALES HABITUALES

Formas funcionales usuales

Modelo	*Variable dependiente*	*Variable independiente*	*Interpretación del cambio*	*Elasticidad*
Nivel-nivel	Y	X	$\Delta Y = \beta \Delta X$	$\beta(X/Y)$
Nivel-log	Y	$\ln X$	$\Delta Y = (\beta/100)\Delta X\%$	$\beta(1/Y)$
Log-nivel	$\ln Y$	X	$\Delta Y\% = 100\beta \Delta X$	βX
Log-log	$\ln Y$	$\ln X$	$\Delta Y\% = \beta \Delta X\%$	β

2.3 REGRESIÓN MÚLTIPLE

2.1.1 Estimación por MCO de la función de regresión

El modelo de regresión simple es un caso particular del modelo de regresión lineal múltiple que facilita la comprensión del método mínimo cuadrático pero tiene la limitación de admitir solo una variable independiente. En el modelo de regresión múltiple se extiende el modelo de regresión simple para incluir variables adicionales como regresores.

El modelo de regresión lineal normal clásico (MRLM), que se va a estudiar, considera que la relación entre la variable dependiente (Y) y las independientes (X_1 ,X_2, ... , X_k).

El modelo de regresión múltiple poblacional se puede formular a partir de la siguiente expresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

Se puede observar que ente modelo:

- El número de regresores X (variables explicativas) es igual a k
- El número coeficientes β es igual a k+1

- La variable a explicar (Y) depende, y por tanto varía, en función del valor que tomen varias variables (X_1 ,X_2, ... , X_k).

- Los coeficientes en términos relativos capturan el efecto parcial, esto es, el efecto esperado sobre *Y* ante un cambio en una de las variables explicativas, cuando el valor de las otras variables explicativas toma un valor determinado y por tanto a esto efecto fijo.

- La variable ε se denomina **término de error** y representa todos los otros factores que además de X_1, X_2,..., X_k determinan el valor de la variable dependiente Y para una observación concreta que llamamos observación i, de manera que para cada observación i habrá un error ε_i.

El modelo no es observable directamente puesto que solo tenemos acceso a una muestra y no a la población. Siempre podemos definir el modelo estimable a partir de

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

y despejando:

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i$$

Como en el caso de regresión lineal simple, se trata de localizar los parámetros que permiten minimizar la suma de los cuadrados de los residuos.

Sabiendo que:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$$

Y calculando los errores a partir de:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

Obtenemos que la suma de los cuadrados de los residuos en regresión lineal múltiple es:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right)^2$$

Y el mínimo de esta función:

$$\min_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right)^2$$

se consigue derivando respecto a cada parámetro e igualando a cero y operando se llega a las k+1 **ecuaciones normales**:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right) = \sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

$$\sum_{i=1}^{n} X_{ji} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right) = \sum_{i=1}^{n} X_{ij} \hat{\varepsilon}_i = 0$$

Con estas ecuaciones se pueden calcular los coeficientes o parámetros de la regresión y deducir algunas ecuaciones útiles como se hizo en regresión lineal simple:

- $\overline{\hat{\varepsilon}} = 0$
- $\overline{Y} = \overline{\hat{Y}}$
- $\text{cov}(\hat{Y}_i, \hat{\varepsilon}_i) = 0$
- $\text{cov}(X_j, \hat{\varepsilon}_i) = 0$

Dividiendo la primera ecuación normal por n en ambas partes y operando llegamos a esta expresión:

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1 + \hat{\beta}_2 \overline{X}_2 + \ldots + \hat{\beta}_k \overline{X}_k$$

cuando la regresión pasa por las medias de las variables independientes, los errores se anulan (la relación es exacta en las medias).
El mismo modelo centrado en sus medias:

$$(Y_i - \overline{Y}) = \hat{\beta}_1 (X_{1i} - \overline{X}_1) + \hat{\beta}_2 (X_{2i} - \overline{X}_2) + \cdots + \hat{\beta}_k (X_{ki} - \overline{X}_k) + \hat{\varepsilon}_i.$$

Y si seguimos operando:

$$y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + \hat{\varepsilon}_i.$$

A partir de las ecuaciones normales se despejan los estimadores de los coeficientes y tenemos:

$$\begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \ldots & \sum X_{1i}X_{ki} \\ \sum X_{2i} & \sum X_{2i}X_{1i} & \sum X_{2i}^2 & \ldots & \sum X_{2i}X_{ki} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sum X_{ki} & \sum X_{ki}X_{1i} & \sum X_{ki}X_{2i} & \ldots & \sum X_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \ldots & X_{kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}$$

$$(\boldsymbol{X'X}) \qquad \hat{\boldsymbol{\beta}} = \qquad \boldsymbol{X'} \qquad \boldsymbol{y}$$

$$\hat{\boldsymbol{\beta}}_{\mathbf{MCO}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$\hat{\boldsymbol{\beta}}_{\mathbf{MCO}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y} = \begin{pmatrix} n & \sum_{i=1}^{n} X_{1i} & \cdots & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}^2 & \cdots & \sum_{i=1}^{n} X_{1i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \cdots & \sum_{i=1}^{n} X_{ki}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i X_{1i} \\ \vdots \\ \sum_{i=1}^{n} Y_i X_{ki} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

Donde $\widehat{\boldsymbol{\beta}}_{\mathbf{MCO}}$ es el vector columna de los parámetros estimados. Para la estimación de cada parámetro se tiene en cuenta no solo la relación entre variable regresora y regresada, como ocurría en el análisis de regresión simple, sino que en la regresión múltiple se tienen en cuenta la relación entre todas las regresaras para el cálculo de cada parámetro. De manera que la introducción de una nueva variable explicativa hará que el resto de parámetros se modifiquen salvo que la nueva variable no esté correlacionada con el resto, en cuyo caso los parámetros son nulos; o bien que el parámetro de la nueva variable introducida sea nulo.

Todos estos cálculos realizan con la forma matricial de modelo de regresión múltiple que desarrollará en el apéndice.

La matriz X'X tiene dimensiones (k+1)x(k+1)

$$\mathbf{X'X} = \begin{pmatrix} n & \sum_{i=1}^{n} X_{1i} & \cdots & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}^2 & \cdots & \sum_{i=1}^{n} X_{1i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \cdots & \sum_{i=1}^{n} X_{ki}^2 \end{pmatrix}$$

- Donde n es el tamaño muestral.
- Las expresiones: $\sum_{i=1}^{n} X_{1i}^2 \quad \cdots \quad \sum_{i=1}^{n} X_{ki}^2$ de la diagonal principal son las sumas de los cuadrados de las variable explicativas.
- Las expresiones: $\sum_{i=1}^{n} X_{1i}X_{ki} \quad \ldots \quad \sum_{i=1}^{n} x_{ki}x_{1i}$ son las sumas de los productos cruzados de las variables explicativas.

En general las características que vimos en el análisis de regresión lineal simple se pueden extender al múltiple:

- $\sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i = 0$ La estimación de la variable regresada y los residuos no están correlados.

Lo que implica que la covarianza es nula $\text{cov}(\hat{Y}_i, \hat{\varepsilon}_i) = 0$

- La variables independientes X y los residuos también están incorrelacionados.

$\text{cov}(X_{1i}, \hat{\varepsilon}_i) = 0$

Con los supuestos del modelo clásico de regresión lineal estos estimadores de MCO son lineales e insesgados, y dentro de éstos son los de mínima varianza (MELI).

Recordando la expresión: $\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(\hat{\varepsilon}_i)$

Calculamos el coeficiente de determinación R^2 , que se interpreta igual que en el modelo de regresión lineal simple pero además incluiremos la formulación matricial.

$$R^2 = \frac{var(\hat{Y})}{var(Y)} = \frac{var(Y) - var(\hat{\varepsilon})}{var(Y)} = \frac{SCT - SCR}{SCT} = \frac{\mathbf{y'y} - (\mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{X'y})}{\mathbf{y'y}} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X'y}}{\mathbf{y'y}}.$$

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{Y}^2}{\boldsymbol{y}'\boldsymbol{y} - n\bar{Y}^2}$$

Existe una relación entre el R^2 y la varianza de los estimadores $\hat{\beta}_j$:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1 - R_j^2}\right) = \frac{\sigma^2}{SCT_j(1 - R_j^2)}$$

Donde R_j^2 es el coeficiente de determinación de la regresión de X_j sobre las (k - 1) variables independientes restantes y SCT_j es la suma cuadrática total de la variable independiente j.

El coeficiente de determinación indica que proporción de variabilidad total queda explicada por la regresión. Si el modelo tiene término independiente, entonces R^2 toma valores entre 0 y 1. Este coeficiente permite, además, seleccionar entre modelos clásicos que tengan el mismo número de regresores, ya que la capacidad explicativa de un modelo es mayor cuanto más elevado sea el valor que tome este coeficiente.

Por otra parte el valor coeficiente de determinación crece con el número de regresores del modelo. Por ello, si los modelos que se comparan tienen distinto número de regresores, no puede establecerse comparación entre sus R^2. En este caso debe emplearse el coeficiente de determinación corregido $\overline{R}^2$, que depura el incremento que experimenta el coeficiente de determinación cuando el número de regresores es mayor.

Por ello, si los modelos que se comparan tienen distinto número de regresores, no puede establecerse comparación entre sus R^2.

En práctica, el uso de R^2 presenta algunas limitaciones a la hora de comparar varios modelos desde la perspectiva de bondad del ajuste. En efecto, cuantas más variables explicativas incorporamos al modelo, mayor será el coeficiente de determinación, pues la SCE disminuye conforme aumenta el número de variables explicativas.

Como el coeficiente de determinación crece con el número de regresores del modelo, necesitamos una medida de bondad de ajuste que tenga en cuenta el ajuste en función del número de variables.

Por tanto, cuando queremos llevar a cabo un análisis comparativo entre varios modelos restringidos, utilizamos R^2 corregido.

2.1.2 Coeficiente de determinación R^2 corregido

Una característica del modelo de regresión múltiple es que a medida que aumentamos el número de regresores X_j el coeficiente de determinación R^2 necesariamente aumenta salvo que el coeficiente estimado sea *exactamente* nulo. Entonces el R^2 nunca disminuye al incorporar nuevos regresores. Así que un incremento del R^2 no significa necesariamente que añadir una nueva variable realmente haya mejorado la calidad del ajuste de nuestro modelo. En realidad incluso si la nueva variable incluida en el modelo mejora nuestro ajuste, sabemos que necesariamente el R^2 de la nueva regresión estará artificialmente «inflado» por el mero hecho de incorporar un nuevo regresor. Por este motivo se utiliza el **R^2 corregido,** que ajusta por el número de coeficientes estimados y cuya definición es:

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{n-k-1}}{\frac{SCT}{n-1}} = 1 - \frac{\hat{\sigma}^2}{S_Y^2}$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Donde:

- $\hat{\sigma}^2$ es el estimador ingresado de la verdadera varianza de los residuos.
- S_Y^2 es la varianza muestral de Y.

A tener en cuenta:

- $\bar{R}^2 < R^2$ El coeficiente de determinación corregido siempre será menor que el coeficiente de determinación.
- Añadir un nuevo regresor tiene dos efectos opuestos; el resultado final sobre $\bar{R}^2$, el resultado final dependerá de cuál de los dos efectos es mayor:
  - Al disminuir la SCR → se incrementa la $\bar{R}^2$
  - El cociente (n-1)/(n-k-1) aumenta
- $\bar{R}^2$ **puede ser negativo** si los regresores en conjunto reducen la SCR en una cantidad tan pequeña que dicha reducción no logre superar el efecto del factor (n-1)/(n-k-1).

Ejemplo 7: Consumo de las familias catalanas dedicadas a la hostelería y el turismo

Con datos de Cataluña y del sector de la hostelería , nos proponemos analizar el consumo de las familias catalanas.

Modelo poblacional: $\ln consumo = \beta_0 + \beta_1(\ln ingresos) + \varepsilon$

Y su estimación (FRM): $\widehat{\ln consumo} = 3,89 + 0,615(\ln ingresos)$

$$n = 95, R^2 = 0,3292.$$

Este modelo es una regresión lineal simple. Es un modelo log-log, por tanto los coeficiente se interpretan directamente como las elasticidades. Un incremento del 1 % en los ingresos provoca que el consumo se incremente un 0,615 %.

Ampliando el modelo, ya que es muy probable que el consumo también tenga relación con el número de miembros en la familia (tamaño), ya que se espera que a medida que el tamaño aumenta, el consumo también aumenta, la nueva estimación:

$$\widehat{\ln consumo}_i = 5,15 + 0,443 \cdot (\ln ingreso_i) + 0,1420 \cdot tamaño_i,$$

$$n = 95, R^2 = 0,4149.$$

Al introducir la nueva variable, observamos que los coeficientes han cambiado, como hemos explicado anteriormente. Así que el término independiente y la elasticidad la variable ingresos cambia.

La nueva estimación nos aporta información sobre cómo influye el incremento, o decremento, del número de miembros en el consumo de las familias, dado un nivel determinado de ingresos.

Respecto a la variable tamaño el modelo es un modelo log-lineal así que la interpretación de los coeficientes estimados nos indica que, manteniendo constante el nivel de ingresos, es decir controlando el efecto del ingreso en el consumo, entonces el incremento de un miembro en la familia se prevé un incremento medio del 14,20 % del consumo familiar (100x0,1420 = 14,20).

Por otro lado, el incremento de los ingresos en un 1 %, dado un tamaño familiar determinado, solo produce un incremento del 0,443 % del consumo, que contrasta con el 0,615 % de la expresión del modelo anterior de regresión lineal simple. Por tanto la introducción de nuevas variables (*tamaño*) afecta al resto de coeficientes de las variables del modelo (*Ingreso).*

Comparando los dos modelos podemos ver que el coeficiente de determinación del segundo modelo R^2 = 0,4149 es mayor que el del primero R^2 = 0,3239. Esto no podía ser de otro modo ya que como hemos comentado anteriormente el coeficiente de determinación aumenta al añadir regresores al modelo. Así que en este caso, no tiene sentido comparar estos coeficientes para decidir si la nueva variable añadida (tamaño) ha mejorado el ajuste del modelo, para poder determinar si el nuevo ajuste es mejor, necesitaríamos conocer el coeficiente de determinación corregido.

2.1.3 Formas funcionales cuadráticas

La regresión múltiple permite establecer relaciones funcionales de una variable que no se pueden tratar o modelizar mediante la regresión simple.

Supongamos una relación cuadrática del siguiente tipo:

$$Y = \beta_0 - \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

Para estudiar la variación esperada en Y tras un cambio unitario en X_1 derivamos:

$$dY = -\beta_1 dX_1 + 2\beta_2 X_1 dX_1 = dX_1\left(-\beta_1 + 2\beta_2 X_1\right)$$

y sustituyendo los diferenciales por incrementos

$$\Delta Y = \left(-\beta_1 + 2\beta_2 X_1\right)\Delta X_1$$

de manera que el incremento de la variables explicada depende del incremento de la variable independiente pero también de su nivel.

Igualando a cero obtenemos su máximo o mínimo:

$$-\beta_1 + 2\beta_2 X_1 = 0;.$$

Para ese nivel inicial de X_1 existe:

- Un Mínimo: si $\beta_2 > 0$
- Un Máximo: si $\beta_2 < 0$

EJERCICIO 1: Salarios en el sector turístico español

Estimamos el modelo en que el salario hora en el sector turístico español depende, con una relación del nivel de estudios acabados, y también de la antigüedad en la empresa.

Modelo poblacional:

$$salario = \beta_0 + \beta_1 estudios + \beta_2 estudios^2 + \beta_3 antigüedad + \beta_4 antigüedad^2 + \varepsilon,$$

Y su estimación (FRM):

$$\widehat{salario}_i = 8,04 - 0,385 \cdot estudios_i + 0,189 \cdot estudios_i^2 + 0,299 \cdot antigüedad_i - 0,0017 \cdot antigüedad_i^2,$$

$$n = 5286, R^2 = 0,2165, \bar{R}^2 = 0,2159.$$

a) ¿En qué nivel de estudios se alcanza el salario mínimo fijada la antigüedad?

b) ¿Para cuántos años de antigüedad alcanza el salario su máximo fijados los estudios?

c) ¿En cuánto se incrementa el salario por cada nivel de estudios superado, ceteris paribus la antigüedad?

d) ¿En cuánto se incrementa el salario medio cuando se pasa de uno a dos años de antigüedad? ¿y cuando se pasa de 29 a 30 años? ¿Por qué se produce distinto incremento?

2.1.4 Términos de interacción

En ocasiones es adecuado para dotar de mayor realismo o afinación al modelo previsto hacer que una variable explicativa dependa de la magnitud o nivel que alcanza otra variable independiente. Es como si ambas variables explicativas tuvieran un efecto parcial no solo aisladamente, sino también conjuntamente. Este tipo de interacción se puede considerar introduciendo en el modelo un término nuevo que actúe como término de interacción:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Diferenciando en en ambas partes respecto a Y y X_1 tenemos:

$$dY = \beta_1 dX_1 + \beta_3 X_2 dX_1 = dX_1 \left(\beta_1 + \beta_3 X_2\right)$$

y sustituyendo los diferenciales por incrementos:

$$\Delta Y \doteq \left(\beta_1 + \beta_3 X_2\right) \Delta X_1$$

de manera que el incremento de la variables explicada depende del incremento de la variable independiente pero también de su nivel de la variable con la que interacciona.

EJERCICIO 2: Usuarios de internet

Nos plantemos si los ingresos per cápita y los años de estudio influyen en la proporción de la población usuaria de internet. Consideramos además que el efecto sobre los usuarios de internet de una variación porcentual en los ingresos depende de los años de educación.

Su estimación (FRM):

$$\begin{aligned} \widehat{internet} &= 52{,}608 - 6{,}26 \ln(ingresos) - 19{,}08 \cdot estudios \\ &\quad + 2{,}511 \left[\ln(ingresos) \cdot estudios\right] \\ n &= 169,\ R^2 = 0{,}8024,\ \bar{R}^2 = 0{,}7988. \end{aligned}$$

a) ¿Cuál es el efecto parcial sobre internet de los ingresos, si consideramos que el valor del nivel de estudios es 7,59?

b) ¿Cuál será el efecto parcial sobre internet de los estudios, fijando el valor de los ingresos en el ingreso medio per cápita en términos de PPA en logaritmos de la muestra que es 8,8.

2.1.5 Regresiones con variables estandarizadas

Cuando alguna de las variables tiene una escala de valores de difícil interpretación puede ser interesante medirla en términos tipificados o estandarizados. Tipificar no es más que restar la media a todos los valores de la variable y dividirla por su desviación típica o error estándar.

$$Z = \left(X - \overline{X}\right) / S_X$$

Puede resultar adecuado expresar todo el modelo estandarizado, en este caso se llama modelo de coeficientes beta.

Tipificando las variables del modelo tenemos que la estimación MCO es:

$$\frac{y_i}{S_Y} = \left(\frac{S_{X_1}}{S_Y}\right)\hat{\beta}_1\left(\frac{x_{1i}}{S_{X_1}}\right) + \left(\frac{S_{X_2}}{S_Y}\right)\hat{\beta}_2\left(\frac{x_{2i}}{S_{X_2}}\right) + \ldots + \left(\frac{S_{X_k}}{S_Y}\right)\hat{\beta}_k\left(\frac{x_{ki}}{S_{X_k}}\right) + \frac{\hat{\varepsilon}_i}{S_Y}$$

Y podemos expresar el modelo en función de la variable tipificada z:

$$Z_y = \tilde{\beta}_1 Z_1 + \tilde{\beta}_2 Z_2 + \ldots \tilde{\beta}_k Z_k + \tilde{\varepsilon}$$

Donde:

$$\tilde{\beta}_2 = \left(S_{x_2} / S_Y\right)\hat{\beta}_2, \ldots, \tilde{\beta}_k = \left(S_{x_k} / S_Y\right)\hat{\beta}_1$$

Una de las ventajas de los coeficientes beta es que no dependen de las unidades de medida utilizadas y permiten determinar la influencia de las variables explicativas sobre la explicada a partir de la magnitud del coeficiente, lo que normalmente no ocurre en los otros casos en que los coeficientes pueden modificarse cambiando las unidades de medida de las variables.

2.4 MODELIZACIÓN

Nunca sabremos la forma funcional del verdadero modelo, es decir, la verdadera relación funcional entre las variables socio-económicas. Nuestro modelo seleccionado, tras haber realizado suficientes pruebas y comprobaciones es una aproximación.

En términos generales, la guía más natural para elegir la forma funcional , si bien no es la única y podría matizarse en función del problema a tratar, consistiría en:

- optar por una forma que sea consistente con lo que indica la teoría económica sobrea relación.
- elegir una forma que sea suficientemente flexible para ajustar los datos.
- elegir una forma funcional que (mejor) asegure que los supuestos que veremos en temas posteriores son satisfechos, de modo que los estimadores -en este caso MCO- tengan igualmente las propiedades deseadas para un estimador.

Estas propiedades también se verán en los siguientes temas. Comprobaremos entonces que el análisis de los residuos del modelo estudiado será determinante para determinar la calidad del modelo, de las propiedades que cumplan los residuos dependen las características de los estimadores.