El modelo Lineal General

Román Salmerón Gómez

Universidad de Granada

Especificación del modelo

El modelo lineal uniecuacional múltiple analiza la relación lineal entre una variable dependiente, Y, y más de una variable independiente, X_i, $i = 1, \ldots, k$, $k > 1$, más un término aleatorio, u.

Así, a partir de n observaciones para cada variable, el modelo puede ser expresado como:

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \cdots + \beta_k X_{tk} + u_t, \quad t = 1, \ldots, n, \tag{1}$$

donde se ha considerado que hay término constante, es decir, $X_{1t} = 1, \forall t$.

El objetivo será estimar (es decir, obtener una aproximación numérica) aquellas cantidades constantes presentes en el modelo (1), así como la bondad de la estimación realizada. En primer lugar, se escribe dicho modelo para todas y cada una de las observaciones:

$$\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{12} + \beta_3 X_{13} + \cdots + \beta_k X_{1k} + u_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{23} + \cdots + \beta_k X_{2k} + u_2 \\
&\vdots \\
Y_n &= \beta_1 + \beta_2 X_{n2} + \beta_3 X_{n3} + \cdots + \beta_k X_{nk} + u_n
\end{aligned}$$

Que nos conduce a la siguiente forma matricial:

$$y_{n\times 1} = X_{n\times k} \cdot \beta_{k\times 1} + u_{n\times 1}, \tag{2}$$

donde:

$$y_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \beta_{k\times 1} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad u_{n\times 1} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix},$$

$$X_{n\times k} = \begin{pmatrix} 1 & X_{12} & \dots & X_{1k} \\ 1 & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \dots & X_{nk} \end{pmatrix}.$$

Consideraremos las siguientes hipótesis básicas en el modelo lineal uniecuacional múltiple:

- El vector y se puede expresar como combinación lineal de las variables explicativas más un vector de perturbación.
- La perturbación aleatoria está centrada $(E[u_t] = 0,\ t = 1, \ldots, n)$, es homocedástica $\left(Var(u_t) = E[u_t^2] = \sigma^2,\ t = 1, \ldots, n\right)$ e incorrelada $(Cov(u_t, u_s) = E[u_t \cdot u_s] = 0,\ \forall t \neq s,\ t, s = 1, \ldots, n)$. En tal caso se dice que las perturbaciones son esféricas y se verifica que $E[u] = 0_{n\times 1}$ y $Var(u) = E[u \cdot u^t] = \sigma^2 \cdot I_{n\times n}$.
- La matriz X es no estocástica y de rango completo por columnas, es decir, $rg(X) = k$ (como consecuencia $n > k$ y las columnas de X, es decir, $X_i, i = 1, \ldots, n$, son linealmente independientes).
- No hay relación entre variables independientes y la perturbación aleatoria:

$$
\begin{aligned}
Cov(u_{n\times 1}, X_i) &= E\left[(u - E[u]) \cdot (X_i - E[X_i])^t\right] \\
&= E\left[u \cdot (X_i - X_i)^t\right] = E[u_{n\times 1} \cdot 0_{1\times n}] = 0_{n\times n}.
\end{aligned}
$$

Estimación del modelo

Estimación mínimo cuadrática de los coeficientes del modelo

Definiendo los errores o residuos, e, del modelo lineal uniecuacional múltiple como la diferencia entre los verdaderos valores de la variable dependiente y su estimación, esto es

$$e = y - \widehat{y},$$

donde $\widehat{y} = X\widehat{\beta}$, y siguiendo la premisa de minimizar la suma de los cuadrados de los residuos

$$e^t e = (y - X\widehat{\beta})^t \cdot (y - X\widehat{\beta}) = y^t y - 2\widehat{\beta}^t X^t y + \widehat{\beta}^t X^t X \widehat{\beta},$$

se obtiene la estimación del parámetro β como

$$\widehat{\beta} = \left(X^t X\right)^{-1} \cdot X^t y.$$

Dicho método recibe el nombre de mínimos cuadrados ordinarios, MCO, por lo que los estimadores obtenidos a partir de dicho método reciben el nombre de estimadores de mínimos cuadrados ordinarios, EMCO.

Como consecuencias de dicha estimación se verifica que $X^t \cdot e = 0_{k\times 1}$, $i^t \cdot e = 0_{1\times 1}$, $i^t \cdot \widehat{y} = i^t \cdot y$ y $\widehat{y}^t \cdot e = 0_{1\times 1}$ donde $i^t = (1\ 1\ \ldots\ 1)_{1\times n}$.

Estimación mínimo cuadrática de los coeficientes del modelo

Adviértase que:

$$X^t X = \begin{pmatrix} n & \sum_{t=1}^{n} X_{t2} & \cdots & \sum_{t=1}^{n} X_{tk} \\ \sum_{t=1}^{n} X_{t2} & \sum_{t=1}^{n} X_{t2}^2 & \cdots & \sum_{t=1}^{n} X_{t2} X_{tk} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=1}^{n} X_{tk} & \sum_{t=1}^{n} X_{tk} X_{t2} & \cdots & \sum_{t=1}^{n} X_{tk}^2 \end{pmatrix},$$

y

$$X^t y = \begin{pmatrix} \sum_{t=1}^{n} Y_t \\ \sum_{t=1}^{n} X_{t2} Y_t \\ \vdots \\ \sum_{t=1}^{n} X_{tk} Y_t \end{pmatrix}.$$

Teorema 1 (Teorema de Gauss-Markov) *Los estimadores de mínimos cuadrados ordinarios son lineales, insesgados y óptimos (ELIO), es decir, tienen varianza mínima entre la clase de los estimadores lineales e insesgados.*

En efecto, por la forma de escribirse el estimador es evidente que es lineal. Así, llamando:

$$C_{k\times n} = \left(X^t X\right)^{-1}_{k\times k} \cdot X^t_{k\times n} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{kn} \end{pmatrix},$$

se tiene que $\widehat{\beta}$ se expresa como combinación lineal del vector y:

$$\widehat{\beta}_{k\times 1} = C_{k\times n} \cdot y_{n\times 1} = \begin{pmatrix} c_{11}Y_1 + c_{12}Y_2 + \ldots + c_{1n}Y_n \\ c_{21}Y_1 + c_{22}Y_2 + \ldots + c_{2n}Y_n \\ \vdots \\ c_{k1}Y_1 + c_{k2}Y_2 + \ldots + c_{kn}Y_n \end{pmatrix}.$$

Para que el estimador $\widehat{\beta}$ de β sea insesgado se ha de cumplir que $E[\widehat{\beta}] = \beta$. En efecto, sustituyendo $y = X\beta + u$ en $\widehat{\beta}$:

$$\begin{aligned} \widehat{\beta} &= \left(X^t X\right)^{-1} \cdot X^t y = \left(X^t X\right)^{-1} \cdot X^t (X\beta + u) \\ &= \beta + \left(X^t X\right)^{-1} \cdot X^t u \longrightarrow \widehat{\beta} = \beta + \left(X^t X\right)^{-1} \cdot X^t u. \end{aligned}$$

Entonces, teniendo en cuenta que $E[u] = 0$:

$$E[\widehat{\beta}] = E\left[\beta + \left(X^t X\right)^{-1} \cdot X^t u\right] = \beta + \left(X^t X\right)^{-1} \cdot X^t \cdot E[u] = \beta.$$

Por otro lado, la matriz de varianzas-covarianzas de $\widehat{\beta}$:

$$\begin{aligned} Var\left(\widehat{\beta}\right) &= E\left[\left(\widehat{\beta} - E[\widehat{\beta}]\right) \cdot \left(\widehat{\beta} - E[\widehat{\beta}]\right)^t\right] = E\left[\left(\widehat{\beta} - \beta\right) \cdot \left(\widehat{\beta} - \beta\right)^t\right] \\ &= E\left[\left(X^t X\right)^{-1} X^t u \cdot u^t X \left(X^t X\right)^{-1}\right] \\ &= \left(X^t X\right)^{-1} X^t \cdot E[u \cdot u^t] \cdot X \left(X^t X\right)^{-1} \\ &= \sigma^2 \cdot \left(X^t X\right)^{-1} X^t X \left(X^t X\right)^{-1} = \sigma^2 \cdot \left(X^t X\right)^{-1}, \end{aligned}$$

donde se ha tenido en cuenta que $\widehat{\beta}$ es insesgado, $\widehat{\beta} - \beta = \left(X^t X\right)^{-1} X^t u$ y $Var(u) = E[u \cdot u^t] = \sigma^2 \cdot I_{n \times n}$.

Para demostrar que $\widehat{\beta}$ es de mínima varianza consideraremos otro estimador, β^*, de β lineal e insesgado de forma que $Var\left(\widehat{\beta}\right) < Var\left(\beta^*\right)$.

En efecto, $\beta^* = D_{k \times n} \cdot y_{n \times 1}$ tal que $D \cdot X = I_{k \times k}$ es lineal e insesgado. Además, $Var\left(\beta^*\right) = \sigma^2 \cdot DD^t$.

En tal caso, puesto que podemos escribir $D = \left(X^t X\right)^{-1} X^t + W$ con $W \neq 0_{k \times n}$, se tiene que $DD^t = \left(X^t X\right)^{-1} + WW^t$, y en tal caso:

$$Var\left(\beta^*\right) = \sigma^2 \cdot DD^t = \sigma^2 \cdot \left(X^t X\right)^{-1} + \sigma^2 \cdot WW^t = Var\left(\widehat{\beta}\right) + \sigma^2 \cdot WW^t,$$

esto es, $Var\left(\beta^*\right) - Var\left(\widehat{\beta}\right) = \sigma^2 \cdot WW^t$.

Y como WW^t es definida positiva: $Var\left(\beta^*\right) - Var\left(\widehat{\beta}\right) > 0$, y en tal caso:

$$Var\left(\beta^*\right) > Var\left(\widehat{\beta}\right).$$

Estimación de la varianza de la perturbación aleatoria

Además de los coeficientes de las variables independientes, hay en el modelo otra cantidad constante que habrá que estimar: la varianza de la perturbación aleatoria, σ^2.

Un estimador insesgado de σ^2 es:

$$\widehat{\sigma}^2 = \frac{e^t e}{n-k},$$

ya que $E[e^t e] = (n-k) \cdot \sigma^2$.

Para calcular dicho estimador se dispone de la expresión:

$$\widehat{\sigma}^2 = \frac{y^t y - \widehat{\beta}^t X^t y}{n-k}.$$

En consecuencia, la estimación de la matriz de varianzas-covarianzas de $\widehat{\beta}$ es:

$$\widehat{Var\left(\widehat{\beta}\right)} = \widehat{\sigma}^2 \cdot \left(X^t X\right)^{-1}.$$

Validación del modelo

Bondad de ajuste: Coeficiente de determinación

Una vez estimado el modelo lineal uniecuacional multiple, es decir, una vez obtenidas las estimaciones de β y σ^2, el siguiente paso será estudiar la calidad de dichas estimaciones.

Así, a continuación, obtendremos el coeficiente de determinación, que no es más que una medida para estudiar la bondad del ajuste lineal determinado por los estimadores por mínimos cuadrados ordinarios.

Dicho coeficiente de determinación, que se denota por R^2, se define como el porcentaje de variabilidad explicada por el modelo. Por tanto, éste se obtendrá como el cociente entre la varianza explicada por la estimación y la total:

$$R^2 = \frac{\frac{1}{T} \cdot \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{\frac{1}{T} \cdot \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}.$$

Como se observa, el coeficiente de determinación queda expresado en función de la suma de cuadrados explicados (SCE) y los totales (SCT).

Luego, teniendo en cuenta la descomposición

$$SCT = SCE + SCR,$$

se tiene que

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

Entonces, para calcular dicho coeficiente se dispone de la expresión:

$$R^2 = \frac{\widehat{\beta}^t X^t y - n \cdot \overline{Y}^2}{y^t y - n \cdot \overline{Y}^2} = 1 - \frac{y^t y - \widehat{\beta}^t X^t y}{y^t y - n \cdot \overline{Y}^2}.$$

Adviértase que, siempre que el modelo lineal tenga término independiente, el coeficiente de determinación varía entre 0 y 1. El valor 0 lo toma cuando la SCE es nula y, por tanto, el modelo no es adecuado; mientras que toma el valor 1 cuando la SCR es nula y, por tanto, el modelo es adecuado.

Coeficiente de determinación corregido

Puesto que a medida que vamos incluyendo variables en el modelo el coeficiente de determinación aumenta aunque las variables que incluyamos no sean significativas, esto supone un problema.

El coeficiente de determinación corregido, $\overline{R}^2$, viene a resolver este problema del coeficiente de determinación. Dicho coeficiente mide el porcentaje de variación de la variable dependiente (al igual que el coeficiente de determinación) pero teniendo en cuenta el número de variables incluidas en el modelo. Se define como:

$$\overline{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}.$$

En cualquier caso, estas medidas de bondad del ajuste no deben de ser sobrevaloradas. Obtener un R^2 o $\overline{R}^2$ cercano a 1 no indica que los resultados sean fiables, ya que, por ejemplo, puede ser que no se cumpla alguna de las hipótesis básicas y los resultados no ser válidos. Por tanto, estos indicadores han de ser considerados como una herramienta más a tener en cuenta dentro del análisis.

Por otro lado, se podría pensar en usar el coeficiente de determinación para comparar distintos modelos. En tal caso, estos deben de tener la misma variable dependiente ya que así tendrán la misma suma de cuadrados totales. Y aún así, habría que tener cuidado con el problema ya comentado: aumenta su valor al añadir una nueva variable explicativa, sea cual sea su aportación al modelo.

Para evitar tales problemas, a la hora de comparar modelos para elegir uno de ellos se usan los criterios de selección de modelos. Más concretamente, estudiaremos los criterios de información de Akaike (AIC), el bayesiano de Schwarz (BIC) y el de Hannan-Quinn (HQC).

Estos criterios se obtienen a partir de la suma de cuadrados de los residuos y de un factor que penaliza la inclusión de parámetros. Así, un modelo más complejo (con más variables explicativas) reducirá la suma de cuadrados de los residuos pero aumentará el factor de penalización.

Utilizando estos criterios se escogería aquel modelo con un menor valor de AIC, BIC o HQC.

Teniendo en cuenta que:

$$\mathfrak{L} = -\frac{n}{2} \cdot (1 + \ln(2 \cdot \pi) - ln(n)) - \frac{n}{2} \cdot \ln(SCR),$$

el criterio de información de Akaike responde a la expresión:

$$AIC = -2 \cdot \mathfrak{L} + 2 \cdot k,$$

el de Schwarz a:

$$BIC = -2 \cdot \mathfrak{L} + k \cdot \ln(n),$$

y el de Hannan-Qinn:

$$HQC = -2 \cdot \mathfrak{L} + 2 \cdot k \cdot \ln\left(\ln(n)\right).$$

Distribución en el muestreo de los estimadores MCO

Introduciendo la hipótesis de que la perturbación aleatoria sigue una distribución normal, esto es:

$$u_{n\times 1} \sim N(0_{n\times 1}, \sigma^2 \cdot I_{n\times n}).$$

En consecuencia, $\widehat{\beta}_{k\times 1} \sim N(\beta, \sigma^2 \cdot (X^tX)^{-1})$, ya que:

- $\widehat{\beta}$ sigue una distribución normal ya que se puede expresar en función de una normal: $\widehat{\beta} = \beta + (X^tX)^{-1} \cdot X^t u$.
- se tienen calculados el vector de medias, $E\left[\widehat{\beta}\right] = \beta$, y matriz de varianzas-covarianzas, $Var\left(\widehat{\beta}\right) = \sigma^2 \cdot (X^tX)^{-1}$.

Por otro lado, ya que $e^te = u^tMu$ siendo $M_{n\times n} = I - X\left(X^tX\right)^{-1}X^t$ simétrica, idempotente y con $rg(M) = n-k < k$ se tiene que $\frac{u^tMu}{\sigma^2} \sim \chi^2_{n-k}$, lo que se traduce en que

$$\frac{(n-k)\cdot\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}.$$

A continuación abordaremos la especificación de contrastes sobre un conjunto de hipótesis lineales sobre los coeficientes del modelo. Concretamente, suponiendo q restricciones lineales independientes entre sí:

$$
\begin{array}{rcl}
a_{11}\beta_1 + a_{12}\beta_2 + \cdots + a_{1k}\beta_k & = & b_1 \\
a_{21}\beta_1 + a_{22}\beta_2 + \cdots + a_{2k}\beta_k & = & b_2 \\
\vdots \qquad\qquad\qquad \vdots & = & \vdots \\
a_{q1}\beta_1 + a_{q2}\beta_2 + \cdots + a_{qk}\beta_k & = & b_q
\end{array}
$$

Plantearemos contrastar la hipótesis nula $H_0 : R\beta = r$ donde

$$
R_{q\times k} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \ldots & a_{qk} \end{pmatrix}, \qquad r_{q\times 1} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{pmatrix}.
$$

Contraste de un conjunto de hipótesis lineales

Usando la distribución

$$\left(R\widehat{\beta} - R\beta\right)^t \cdot \frac{\left[R\left(X^tX\right)^{-1}R^t\right]^{-1}}{q \cdot \widehat{\sigma}^2} \cdot \left(R\widehat{\beta} - R\beta\right) \sim F_{q,n-k},$$

rechazaremos la hipótesis nula al nivel de significación α si

$$\left(R\widehat{\beta} - r\right)^t \cdot \frac{\left[R\left(X^tX\right)^{-1}R^t\right]^{-1}}{q \cdot \widehat{\sigma}^2} \cdot \left(R\widehat{\beta} - r\right) > F_{q,n-k}(1-\alpha),$$

donde $F_{q,n-k}(1-\alpha)$ es el punto de una F de Senedecor de q y $n-k$ grados de libertad que deja por debajo suyo una probabilidad $1-\alpha$.

Un caso particular de suma importancia será aquel en el que se desee contrastar la hipótesis nula $H_0 : \beta_i = b_i, i = 1, \ldots, k$.

En tal caso, $q = 1$, $R = (0\ 0\ \ldots\ 1^{i)}\ \ldots 0)$ y $r = b_i$, por lo que la distribución anterior queda simplificada como

$$\frac{\left(\widehat{\beta}_i - b_i\right)^2}{\widehat{\sigma}^2 \cdot w_i} \sim F_{1,n-k},$$

donde w_i es el elemento (i,i) de la matriz $\left(X^t X\right)^{-1}$, o lo que es lo mismo, $\widehat{\sigma}^2 \cdot w_i$ es el elemento (i,i) de $\widehat{\sigma}^2 \cdot \left(X^t X\right)^{-1} = \widehat{Var\left(\widehat{\beta}\right)}$, esto es, la varianza estimada de $\widehat{\beta}_i$.

Teniendo en cuenta que la raíz cuadrada de una F-Snedecor con 1 y n grados de libertad es una t-Student con n grados de libertad se tiene que

$$\frac{\widehat{\beta}_i - b_i}{\widehat{\sigma} \cdot \sqrt{w_i}} \sim t_{n-k},$$

y en tal caso rechazaremos $H_0 : \beta_i = b_i$ al nivel de significación α si

$$\left| \frac{\widehat{\beta}_i - b_i}{\widehat{\sigma} \cdot \sqrt{w_i}} \right| > t_{n-k}\left(1 - \frac{\alpha}{2}\right),$$

donde $t_{n-k}\left(1 - \frac{\alpha}{2}\right)$ es el punto de una distribución t de student con $n-k$ grados de libertad que deja por debajo suya una probabilidad $1 - \frac{\alpha}{2}$.

Este caso particular es de vital importancia cuando $b_i = 0$, ya que entonces estaremos contrastando si el coeficiente de la variable independiente X_i es o no nulo. De forma que al rechazar dicha hipótesis tenemos garantizado que la variable X_i ha de estar en el modelo, por lo que sus variaciones influyen en la variable dependiente. En tal caso se dice que dicha variable es significativa y que el contraste es un contraste de significación individual.

En el caso en el que no se rechace la hipótesis nula $H_0 : R\beta = r$, sería deseable incorporar dicha información al modelo. En tal caso, se obtiene un nuevo estimador:

$$\widehat{\beta}_R = \widehat{\beta} + \left(X^t X\right)^{-1} R^t \left[R \left(X^t X\right)^{-1} R^t\right]^{-1} \cdot \left(r - R\widehat{\beta}\right),$$

que recibe el nombre de mínimos cuadrados restringidos ya que se ha obtenido con la restricción de que ha de verificar que $R\widehat{\beta}_R = r$.

Dicho estimador es lineal, insesgado siempre que la hipótesis nula $H_0 : R\beta = r$ sea cierta y óptimo. Es decir, el estimador por mínimos cuadrados restringidos tiene menor varianza que el estimador mínimo cuadrático ordinario siempre y cuando la restricción (hipótesis nula) sea cierta.

Luego, cuando una restricción lineal sobre los coeficientes de las variables independientes es cierta, el estimador por mínimos cuadrados ordinarios deja de ser óptimo y habrá que usar el estimador por mínimos cuadrados restringidos.

Además se verifica que:

$$SCR_R \geq SCR, \quad R_R^2 \leq R^2.$$

El análisis de la varianza aborda el contraste que tiene por hipótesis nula que todos los coeficientes de las variables independientes son nulos simultáneamente, esto es, $H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$.

Salta a la vista que estamos ante un caso particular de un contraste sobre $k-1$ restricciones lineales de los coeficientes de las variables independientes. En este caso, rechazaremos la hipótesis nula al nivel de significación α si

$$F_{exp} = \frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}} > F_{k-1,n-k}(1-\alpha).$$

Para calcular dicho estadístico se suele resumir la información anterior en una tabla, conocida como tabla de análisis de la varianza (tabla ANOVA) ya que en ella se recogen las fuentes de variación de la varianza:

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Medias
Explicada	$SCE = \widehat{\beta}^t X^t y - n\overline{Y}^2$	$k-1$	$\frac{SCE}{k-1}$
Residuos	$SCR = y^t y - \widehat{\beta}^t X^t y$	$n-k$	$\frac{SCR}{n-k}$
Total	$SCT = y^t y - n\overline{Y}^2$	$n-1$	

Adviértase que rechazar H_0 implica que hay al menos un coeficiente no nulo, por lo que la relación existente entre las variables independientes y la dependiente no se debe al azar, lo cual valida el modelo en su conjunto.

Por otro lado, sin más que dividir la región de rechazo por SCT tanto en el numerador como en el denominador se obtiene la expresión equivalente:

$$\frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}} > F_{k-1,n-k}(1-\alpha).$$

La importancia de esta nueva expresión para la región de rechazo es que permite calcular una cota, sin más que despejar R^2, a partir de la cual el coeficiente de determinación es significativo. Esto es, el coefciente de determinación es significativo al nivel de significación α si

$$R^2 > \frac{\frac{k-1}{n-k}\cdot F_{k-1,n-k}(1-\alpha)}{1+\frac{k-1}{n-k}\cdot F_{k-1,n-k}(1-\alpha)}.$$

A partir de las distribuciones en el muestreo para los estimadores estudiados es inmediato obtener los siguientes intervalos de confianza al nivel $1-\alpha$:

Intervalo de confianza para β_i

$$\widehat{\beta}_i \pm t_{n-k}\left(1-\frac{\alpha}{2}\right)\cdot\widehat{\sigma}\cdot\sqrt{w_i},\quad i=1,\ldots,k.$$

Intervalo de confianza para σ^2

$$\left[\frac{(n-k)\cdot\widehat{\sigma}^2}{\chi^2_{n-k}\left(1-\frac{\alpha}{2}\right)},\ \frac{(n-k)\cdot\widehat{\sigma}^2}{\chi^2_{n-k}\left(\frac{\alpha}{2}\right)}\right],$$

donde $\chi^2_{n-k}\left(1-\frac{\alpha}{2}\right)$ y $\chi^2_{n-k}\left(\frac{\alpha}{2}\right)$ son los puntos de una distribución chi-cuadrado con $n-k$ grados de libertad que dejan a su izquierda, respectivamente, una probabilidad $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$.

Una forma alternativa de contrastar hipótesis es usando los intervalos de confianza. De manera que para contrastar $H_0 : R\beta = r$ se calculará la región de confianza para $R\beta$ y si r pertenece a dicha región, no se rechazará la hipótesis nula.

Explotación del modelo

Una vez validado el modelo, la siguiente fase de un modelo econométrico es la explotación, siendo entonces la predicción o la permanencia estructural algunos de sus objetivos.

La predicción se realiza desde dos puntos de vista: a) por un lado realizaremos una predicción puntual dando un único valor de predicción para un instante en concreto; b) por otra parte, puesto que Y es una variable aleatoria, podemos calcular su esperanza dado un valor en concreto de las variables independientes.

Siguiendo las directrices anteriores se llega a la misma expresión algebráica en ambos casos:

$$p_0 = x_0^t \cdot \widehat{\beta},$$

donde $x_0^t = (1 \ X_{02} \ X_{03} \ \ldots \ X_{0k})$ contiene los valores de las variables independientes para los que se quiere obtener la predicción.

Este predictor, p_0, mínimo cuadrático (ya que se obtiene a partir del estimador por mínimos cuadrados ordinarios de β) es lineal, insesgado y óptimo (en el sentido de mínima varianza).

En este apartado calcularemos el intervalo de confianza para el valor esperado de Y dado x_0, es decir, para $E[Y_0/x_0] = x_0^t \cdot \beta$.

Como $x_0^t \cdot \widehat{\beta}$ se distribuye según una normal (ya que está en función de $\widehat{\beta}$) y

- $E[x_0^t \cdot \widehat{\beta}] = x_0^t\beta$, ya que es insesgado.
- $Var\left(x_0^t \cdot \widehat{\beta}\right) = E\left[\left(x_0^t \cdot \widehat{\beta} - x_0^t \cdot \beta\right) \cdot \left(x_0^t \cdot \widehat{\beta} - x_0^t \cdot \beta\right)\right] = x_0^t \cdot E\left[\left(\widehat{\beta} - \beta\right) \cdot \left(\widehat{\beta} - \beta\right)^t\right] \cdot x_0 = x_0^t \cdot Var\left(\widehat{\beta}\right) \cdot x_0 = \sigma^2 \cdot x_0^t \left(X^t X\right)^{-1} x_0$.

se tiene que

$$x_0^t \cdot \widehat{\beta} \sim N\left(x_0^t \cdot \beta, \sigma^2 \cdot x_0^t \left(X^t X\right)^{-1} x_0\right).$$

Ahora bien, esta distribución no es apta para hacer inferencia puesto que depende de la cantidad desconocida σ^2. Para resolver este problema, tipificaremos la anterior distribución normal y la dividiremos entre la raíz cuadrada de la siguiente distribución chi-cuadrado

$$\frac{(n-k)\cdot\widehat{\sigma}^2}{\sigma^2}\sim\chi^2_{n-k},$$

dividida a su vez entre sus grados de libertad, obteniendo la siguiente distribución t-Student:

$$\frac{x_0^t\cdot\widehat{\beta}-x_0^t\cdot\beta}{\widehat{\sigma}\cdot\sqrt{x_0^t\left(X^tX\right)^{-1}x_0}}\sim t_{n-k}.$$

A partir de esta distribución, el intervalo de confianza al nivel $1-\alpha$ para $E[Y_0/x_0]=x_0^t\cdot\beta$ es:

$$x_0^t\cdot\widehat{\beta}\pm t_{n-k}\left(1-\frac{\alpha}{2}\right)\cdot\widehat{\sigma}\cdot\sqrt{x_0^t\left(X^tX\right)^{-1}x_0},$$

donde $t_{n-k}\left(1-\frac{\alpha}{2}\right)$ es el punto de una distribución t de Student con $n-k$ grados de libertad que deja a su izquierda una probabilidad $1-\frac{\alpha}{2}$.

Al explotar el modelo mediante la predicción se está presuponiendo que la relación estimada se mantiene para la información no presente en la muestra observada. Para confirmar este aspecto, calcularemos el intervalo de confianza para Y dado x_0, de forma que si la nueva información pertenece a dicho intervalo, la estructura del modelo estimado permanecerá.

Partiendo de que

$$Y_0 - \widehat{Y}_0 = u_0 - x_0^t\left(\widehat{\beta} - \beta\right) \sim N\left(0, \sigma^2 \cdot \left(1 + x_0^t\left(X^tX\right)^{-1} x_0\right)\right),$$

se llega de forma análoga a la anterior a la distribución

$$\frac{Y_0 - \widehat{Y}_0}{\widehat{\sigma} \cdot \sqrt{1 + x_0^t\left(X^tX\right)^{-1} x_0}} \sim t_{n-k},$$

donde $\widehat{Y}_0 = x_0^t \cdot \widehat{\beta}$. Por tanto, el intervalo de confianza al nivel $1-\alpha$ para Y_0 es:

$$x_0^t \cdot \widehat{\beta} \pm t_{n-k}\left(1 - \frac{\alpha}{2}\right) \cdot \widehat{\sigma} \cdot \sqrt{1 + x_0^t\left(X^tX\right)^{-1} x_0}.$$

Ejemplos

A continuación vamos a realizar un análisis exhaustivo del modelo

$$Y_t = \beta_1 + \beta_2 \cdot X_{t2} + \beta_3 \cdot X_{t3} + u_t,$$

a partir de las siguiente información muestral:

Observación	Y_t	X_{t2}	X_{t3}
1	16	1	1
2	26	3	2
3	30	5	-1
4	44	7	3
5	56	8	-2
6	64	10	0
7	68	10	1
8	72	12	4

En primer lugar calcularemos la estimación por mínimos cuadrados ordinarios de los coeficientes de las variables a partir de la expresión

$$\widehat{\beta} = \left(X^t X\right)^{-1} X^t y. \tag{3}$$

A partir de la información muestral anterior es claro que:

$$y = \begin{pmatrix} 16 \\ 26 \\ 30 \\ 44 \\ 56 \\ 64 \\ 68 \\ 72 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 5 & -1 \\ 1 & 7 & 3 \\ 1 & 8 & -2 \\ 1 & 10 & 0 \\ 1 & 10 & 1 \\ 1 & 12 & 4 \end{pmatrix},$$

de forma que:

$$X^t X = \begin{pmatrix} 8 & 56 & 8 \\ 56 & 492 & 65 \\ 8 & 65 & 36 \end{pmatrix}, \qquad X^t y = \begin{pmatrix} 376 \\ 3184 \\ 414 \end{pmatrix},$$

y entonces a partir de la fórmula (3):

$$\begin{aligned}
\widehat{\beta} &= \begin{pmatrix} 8 & 56 & 8 \\ 56 & 492 & 65 \\ 8 & 65 & 36 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 376 \\ 3184 \\ 414 \end{pmatrix} \\
&= \begin{pmatrix} 0'62 & -0'0688 & -0'0136 \\ -0'0688 & 0'0103 & -0'0033 \\ -0'0136 & -0'0033 & 0'0368 \end{pmatrix} \cdot \begin{pmatrix} 376 \\ 3184 \\ 414 \end{pmatrix} \\
&= \begin{pmatrix} 8'5189 \\ 5'5587 \\ -0'4296 \end{pmatrix}.
\end{aligned}$$

Es decir, $\widehat{\beta}_1 = 8'5189$, $\widehat{\beta}_2 = 5'5587$ y $\widehat{\beta}_3 = -0'4296$. Lo cual se traduce en la siguiente estimación del modelo considerado:

$$\widehat{Y}_t = 8'5189 + 5'5587 X_{t2} - 0'4296 X_{t3}.$$

A partir de estas estimaciones es sencillo obtener las estimaciones de Y:

$$\widehat{y} = X\widehat{\beta} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 5 & -1 \\ 1 & 7 & 3 \\ 1 & 8 & -2 \\ 1 & 10 & 0 \\ 1 & 10 & 1 \\ 1 & 12 & 4 \end{pmatrix} \cdot \begin{pmatrix} 8'5189 \\ 5'5587 \\ -0'4296 \end{pmatrix} = \begin{pmatrix} 13'6480 \\ 24'3358 \\ 36'7420 \\ 46'1410 \\ 53'8477 \\ 64'1059 \\ 63'6763 \\ 73'5049 \end{pmatrix},$$

y los residuos del modelo:

$$e = y - \widehat{y} = \begin{pmatrix} 16 \\ 26 \\ 30 \\ 44 \\ 56 \\ 64 \\ 68 \\ 72 \end{pmatrix} - \begin{pmatrix} 13'6480 \\ 24'3358 \\ 36'7420 \\ 46'1410 \\ 53'8477 \\ 64'1059 \\ 63'6763 \\ 73'5049 \end{pmatrix} = \begin{pmatrix} 2'3520 \\ 1'6642 \\ -6'7420 \\ -2'1410 \\ 2'1523 \\ -0'1059 \\ 4'3237 \\ -1'5049 \end{pmatrix}.$$

Desde un punto de vista teórico, dichos residuos han de sumar cero, si bien en este caso la suma del vector anterior es igual a $-0'0016$. De igual forma, a partir de dichos residuos se puede obtener fácilmente la estimación de la varianza de la perturbación aleatoria, ya que por definición:

$$\widehat{\sigma}^2 = \frac{e^t e}{n-k}, \tag{4}$$

donde $e^t e$ es la suma de los cuadrados de los residuos, n el número de observaciones del modelo y k el número de variables presentes en el mismo. En este caso:

$$\widehat{\sigma}^2 = \frac{83'8472}{8-3} = 16'76944.$$

Otra forma equivalente de obtener la estimación anterior es:

$$\widehat{\sigma}^2 = \frac{y^t y - \widehat{\beta}^t X^t y}{n-k}. \tag{5}$$

Puesto que

$$y^t y = 20808, \quad \widehat{\beta}^t X^t y = (8'5189 \; 5'5587 \; -0'4296) \begin{pmatrix} 376 \\ 3184 \\ 414 \end{pmatrix} = 20724'1528,$$

es claro que

$$\widehat{\sigma}^2 = \frac{20808 - 20724'1528}{8-3} = \frac{83'8472}{5} = 16'76944.$$

Y a partir de esta estimación se puede obtener la estimación de la matriz de varianzas-covarianzas de $\widehat{\beta}$ mediante:

$$\begin{aligned} \widehat{Var\left(\widehat{\beta}\right)} &= \widehat{\sigma}^2 \cdot \left(X^t X\right)^{-1} = 16'7694 \cdot \begin{pmatrix} 0'62 & -0'0688 & -0'0136 \\ -0'0688 & 0'0103 & -0'0033 \\ -0'0136 & -0'0033 & 0'0368 \end{pmatrix} \\ &= \begin{pmatrix} 10'3976 & -1'1533 & -0'2282 \\ -1'1533 & 0'1727 & -0'0555 \\ -0'2282 & -0'0555 & 0'6168 \end{pmatrix}, \end{aligned} \tag{6}$$

que será usada para calcular la región de rechazo de los contrastes de significación individual así como para los intervalos de confianza de cada coeficiente de la regresión.

Para medir la bondad del ajuste realizado mediante la estimación anterior calcularemos el coeficiente de determinación:

$$R^2 = \frac{\widehat{\beta}^t X^t y - n\overline{Y}}{y^t y - n\overline{Y}} = 1 - \frac{y^t y - \widehat{\beta}^t X^t y}{y^t y - n\overline{Y}}. \tag{7}$$

Para la primera expresión de (7), teniendo en cuenta que:

$$\widehat{\beta}^t X^t y - n\overline{Y} = 20724'1528 - 8 \cdot 47^2 = 20724'1528 - 17672 = 3052'1528,$$

$$y^t y - n\overline{Y} = 20808 - 17672 = 3136,$$

se tiene que

$$R^2 = \frac{3052'1528}{3136} = 0'97326301.$$

Además, en tal caso: $\overline{R}^2 = 1 - (1 - 0'97326301) \cdot \frac{7}{5} = 0'9625682.$

Mientras que para la segunda expresión:

$$R^2 = 1 - \frac{83'8472}{3136} = 1 - 0'02673699 = 0'97326301.$$

A partir de este coeficiente podemos afirmar que el ajuste realizado permite explicar un $97'326301\%$ de la variabilidad de la variable dependiente, que si bien se encuentra muy próximo al 100%, más adelante comprobaremos si es significativo y, por tanto, si es suficiente para validar el modelo.
Una vez estimadas las cantidades constantes del modelo, a continuación se estudiará la validez del mismo a partir de:

- contrastes de significación individual.
- contraste de significación conjunta.
- significación del coeficiente de determinación.

Para abordar los contrastes de significación individual tendremos en cuenta que se rechaza $H_0 : \beta_i = 0$ si

$$t_{exp} = \left| \frac{\widehat{\beta}_i}{\widehat{\sigma} \cdot \sqrt{w_i}} \right| > t_{n-k}\left(1 - \frac{\alpha}{2}\right), \quad \forall i,$$

donde w_i es el elemento $(i,\ i)$ de la matriz $(X^t X)^{-1}$ o, lo que es lo mismo, $\widehat{\sigma} \cdot \sqrt{w_i}$ es la raíz cuadrada del elemento $(i,\ i)$ de la matriz $\widehat{\sigma}^2 \cdot (X^t X)^{-1} = \widehat{Var}\left(\widehat{\beta}\right)$.

Observando (6) es claro que $\widehat{\sigma} \cdot \sqrt{w_2} = \sqrt{0'1727} = 0'4156$ y $\widehat{\sigma} \cdot \sqrt{w_3} = \sqrt{0'6168} = 0'7854$. Teniendo en cuenta que $t_{n-k}\left(1 - \frac{\alpha}{2}\right) = t_5(0'975) = 2'57$, se obtiene que:

- rechazo $H_0 : \beta_2 = 0$ si $t_{exp} = \frac{5'5587}{0'4156} = 13'376 > 2'57$.
- rechazo $H_0 : \beta_3 = 0$ si $t_{exp} = \left|\frac{-0'4296}{0'7854}\right| = 0'547 > 2'57$.

Como es evidente, rechazamos $H_0 : \beta_2 = 0$ y no rechazamos $H_0 : \beta_3 = 0$, es decir, la variable X_{t2} influye en Y_t, mientras que la X_{t3} no lo hace. En tal situación se dice que la segunda variable es significativa y que la tercera no es significativa.

Ejemplo 1

Para el contraste de significación conjunta, $H_0 : \beta_2 = \beta_3 = 0$, se rechaza la hipótesis nula si

$$F_{exp} = \frac{SCE/k-1}{SCR/n-k} > F_{k-1,n-k}(1-\alpha),$$

donde $F_{k-1,n-k}(1-\alpha)$ es el punto de una F de Snedecor con $k-1$ y $n-k$ grados de libertad que deja a su izquierda una probabilidad $1-\alpha$, SCE denota a la suma de cuadrados explicada y SCR a la suma de los cuadrados de los residuos (cantidades que ya han sido calculadas con anterioridad al obtener el coeficiente de determinación).

En este caso, para calcular la región de rechazo recurriremos a la tabla ANOVA:

Fuentes de variación	Sumas de cuadrados	Grados de libertad	Medias
Explicada	$SCE = 3052'1528$	$k-1=2$	$1526'0764$
Residual	$SCR = 83'8472$	$n-k=8-3=5$	$16'76944$
Total	$SCT = 3136$		

Luego $F_{exp} = \frac{1526'0764}{16'76944} = 91'00342$.

Y como $F_{k-1,n-k}(1-\alpha) = F_{2,5}(0'95) = 5'78$, es evidente que se rechaza la hipótesis nula. Esto es, existe al menos un coeficiente que es no nulo de manera que entonces se puede afirmar que hay algún tipo de asociación (que no se debe al azar) entre las variables independientes y la dependiente.

Para terminar con la validación del modelo, estuadiaremos si el coeficiente de determinación obtenido con anterioridad es significativo o no. Teniendo en cuenta que:

$$\frac{SCE/k-1}{SCR/n-k} = \frac{R^2/k-1}{(1-R^2)/n-k},$$

la región de rechazo anterior se puede expresar como:

$$\frac{R^2/k-1}{(1-R^2)/n-k} > F_{k-1,n-k}(1-\alpha),$$

y sin más que despejar el coeficiente de determinación, se obtiene que el modelo es significativo si

$$R^2 > \frac{\frac{k-1}{n-k} \cdot F_{k-1,n-k}(1-\alpha)}{1 + \frac{k-1}{n-k} \cdot F_{k-1,n-k}(1-\alpha)} = R^2_{sig}.$$

Esto es, se tiene una cota, R^2_{sig}, a partir de la cual el coeficiente de determinación es significativo.
Puesto que en este caso:

$$\left.\begin{array}{r} \frac{k-1}{n-k} = \frac{2}{5} = 0'4 \\ F_{k-1,n-k}(1-\alpha) = F_{2,5}(0'95) = 5'78 \end{array}\right\} \rightarrow \frac{k-1}{n-k} \cdot F_{k-1,n-k}(1-\alpha) = 0'4 \cdot 5'78 = 2'312$$

$$\rightarrow R^2_{sig} = \frac{2'312}{3'312} = 0'6981.$$

Recordemos que $R^2 = 0'97326301$, que claramente es significativo al ser superior a la cota inferior de significación $R^2_{sig} = 0'6981$. Esto es, el coeficiente de determinación obtenido implica que el modelo es explicativo.
Por todo lo anterior, parece claro que el modelo es válido y, por tanto, apto para la predicción.
Supongamos ahora que se tiene nueva información para las variables independientes ($X_{02} = 2$ y $X_{03} = 3$) y que se desea obtener una predicción puntual y por intervalo a partir de ella para la variable dependiente.

A partir de dicha información, la predicción puntual óptima será

$$x_0^t \widehat{\beta} = (1\ 2\ 3) \cdot \begin{pmatrix} 8'5189 \\ 5'5587 \\ -0'4296 \end{pmatrix} = 18'3475.$$

Mientras que para la predicción por intervalo será necesario calcular:

$$x_0^t \left(X^t X\right)^{-1} x_0 = (1\,2\,3) \cdot \begin{pmatrix} 0'62 & -0'0688 & -0'0136 \\ -0'0688 & 0'0103 & -0'0033 \\ -0'0136 & -0'0033 & 0'0368 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 0'596,$$

de forma que el intervalo de confianza para el valor esperado de Y será:

$$x_0^t \widehat{\beta} \pm t_{n-k} \left(1 - \frac{\alpha}{2}\right) \cdot \widehat{\sigma} \cdot \sqrt{x_0^t \left(X^t X\right)^{-1} x_0}$$

$$= 18'3475 \pm 2'57 \cdot 4'095051 \cdot \sqrt{0'596} = (10'221,\ 26'4742).$$

y el intervalo de confianza para Y será:

$$x_0^t\widehat{\beta} \pm t_{n-k}\left(1-\frac{\alpha}{2}\right)\cdot\widehat{\sigma}\cdot\sqrt{1+x_0^t\left(X^tX\right)^{-1}x_0}$$

$$= 18'3475 \pm 2'57\cdot 4'095051\cdot\sqrt{1'596} = (5'04887,\ 31'64613).$$

Además, a partir de este último intervalo (conocido como permanencia estructural), si se sabe que acompañando a x_0 se tiene $Y_0 = 6$, puesto que este valor pertenece al intervalo calculado, se puede afirmar (al nivel de confianza considerado) que la relación estimada para las variables se sigue verificando (permanece la estructura) para la nueva información.

Por último, con el objetivo de aplicar la estimación con información a priori al modelo considerado vamos contrastar la hipótesis nula $H_0 : \beta_2 + \beta_3 = 5$. Así, en el caso de no rechazarla obtendremos el estimador por mínimos cuadrados restringidos.

Como es sabido, se rechazará la hipótesis nula si

$$F_{exp} = \left(R\widehat{\beta}-r\right)^t\cdot\frac{\left[R\left(X^tX\right)^{-1}R^t\right]^{-1}}{q\cdot\widehat{\sigma}^2}\cdot\left(R\widehat{\beta}-r\right) > F_{q,n-k}(1-\alpha),$$

donde $F_{q,n-k}(1-\alpha)$ es el punto de una F de Snedecor con q y $n-k$ grados de libertad que deja a su izquierda una probabilidad $1-\alpha$.
A partir de $\beta_2 + \beta_3 = 5$ se obtiene que $q = 1$, $r = 5$ y $R = (0\ 1\ 1)$, de forma que

$$R\widehat{\beta} - r = (0\ 1\ 1) \cdot \begin{pmatrix} 8'5189 \\ 5'5587 \\ -0'4296 \end{pmatrix} - 5 = 5'5587 - 0'4296 - 5 = 0'1291,$$

$$R\left(X^tX\right)^{-1}R^t = (0\ 1\ 1) \cdot \begin{pmatrix} 0'62 & -0'0688 & -0'0136 \\ -0'0688 & 0'0103 & -0'0033 \\ -0'0136 & -0'0033 & 0'0368 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = 0'0405.$$

Y en tal caso:

$$F_{exp} = \frac{0'1291^2}{0'0405 \cdot 16'76944} = 0'02454025,$$

donde recordemos que $\widehat{\sigma}^2 = 16'76944$.

Por otro lado, puesto que $F_{q,n-k}(1-\alpha) = F_{1,5}(0'95) = 6'61$, es evidente que no se rechaza la hipótesis nula, es decir, no rechazo que los coeficientes de las variables verifiquen la relación $\beta_2 + \beta_3 = 5$.
En tal caso, habrá que incorporar dicha información al modelo con el fin de obtener un mejor estimador (cuando se dispone de información a priori el estimador por mínimos cuadrados ordinarios ya no es óptimo). En esta situación el estimador insesgado con mínima varianza es el de mínimos cuadrados restringidos, el cual responde a la siguiente expresión:

$$\widehat{\beta}_R = \widehat{\beta} + \left(X^t X\right)^{-1} R^t \left[R\left(X^t X\right)^{-1} R^t\right]^{-1} \left(r - R\widehat{\beta}\right). \tag{8}$$

De la expresión anterior se conoce:

$$\widehat{\beta} = \begin{pmatrix} 8'5189 \\ 5'5587 \\ -0'4296 \end{pmatrix}, \quad \left[R\left(X^t X\right)^{-1} R^t\right]^{-1} = \frac{1}{0'0405}, \quad r - R\widehat{\beta} = -0'1291,$$

faltando calcular

$$\left(X^t X\right)^{-1} R^t = \begin{pmatrix} 0'62 & -0'0688 & -0'0136 \\ -0'0688 & 0'0103 & -0'0033 \\ -0'0136 & -0'0033 & 0'0368 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} =$$

$$= \begin{pmatrix} -0'0824 \\ 0'007 \\ 0'0335 \end{pmatrix}.$$

Entonces, a partir de (8) se obtiene que:

$$\widehat{\beta}_R = \begin{pmatrix} 8'5189 \\ 5'5587 \\ -0'4296 \end{pmatrix} - \frac{0'1291}{0'0405} \cdot \begin{pmatrix} -0'0824 \\ 0'007 \\ 0'0335 \end{pmatrix} = \begin{pmatrix} 8'781563 \\ 5'536386 \\ -0'5363864 \end{pmatrix}.$$

A partir de esta estimación es fácil comprobar que se obtiene:

$$e_R^t e_R = 84'35455, \quad R_R^2 = 0'9731012, \quad \widehat{\sigma}_R^2 = 14'05909,$$

verificándose, como es sabido, que $e_R^t e_R > e^t e$ y $R_R^2 < R^2$.

Dado el modelo

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + u_t, \tag{9}$$

donde:

- Y es el consumo familiar mensual (medido en miles de euros).
- X_2 es la renta familiar mensual (medida en miles de euros).
- X_3 es una variable ficticia que toma el valor 1 si la familia correspondiente tiene una deuda en forma de un préstamo para la compra de una vivienda o coche, y el valor 0 en caso contrario.
- X_4 es el número de hijos de una familia.

Se pide analizar el modelo sabiendo que para 22 familias se ha obtenido que:

$$y^t y = 131'13, \quad X^t y = \begin{pmatrix} 48'5 \\ 204'45 \\ 37'9 \\ 69'3 \end{pmatrix},$$

$$(X^tX)^{-1} = \begin{pmatrix} 0'3342 & -0'0506 & -0'1626 & 0'0041 \\ -0'0506 & 0'0173 & 0'0051 & -0'0114 \\ -0'1626 & 0'0051 & 0'249 & -0'0317 \\ 0'0041 & -0'0114 & -0'0317 & 0'0514 \end{pmatrix}.$$

En primer lugar obtendremos la estimación de las cantidades constantes del modelo, es decir, de β y σ^2:

$$\widehat{\beta} = \begin{pmatrix} 0'3342 & -0'0506 & -0'1626 & 0'0041 \\ -0'0506 & 0'0173 & 0'0051 & -0'0114 \\ -0'1626 & 0'0051 & 0'249 & -0'0317 \\ 0'0041 & -0'0114 & -0'0317 & 0'0514 \end{pmatrix} \cdot \begin{pmatrix} 48'5 \\ 204'45 \\ 37'9 \\ 69'3 \end{pmatrix}$$

$$= \begin{pmatrix} -0'0149 \\ 0'4862 \\ 0'3969 \\ 0'2287 \end{pmatrix}, \quad (10)$$

$$\widehat{\sigma}^2 = \frac{y^ty - \widehat{\beta}^t X^t y}{n-k} = \frac{131'13 - 129'5643}{22-4} = \frac{1'5657}{18} = 0'087, \quad (11)$$

donde se ha usado que

$$\widehat{\beta}^t X^t y = (-0'0149\ 0'4862\ 0'3969\ 0'2287) \cdot \begin{pmatrix} 48'5 \\ 204'45 \\ 37'9 \\ 69'3 \end{pmatrix} = 129'5643,$$

y se deduce que $\widehat{\sigma} = 0'2949$ y

$$\widehat{Y}_t = -0'0149 + 0'4862 \cdot X_{2t} + 0'3969 \cdot X_{3t} + 0'2287 \cdot X_{4t}.$$

Además, a partir de la estimación de σ^2 se obtiene una estimación para la matriz de varianzas-covarianzas de $\widehat{\beta}$:

$$\widehat{Var\left(\widehat{\beta}\right)} = \widehat{\sigma}^2 \left(X^t X\right)^{-1} = \begin{pmatrix} 0'0291 & -0'0044 & -0'0141 & 0'0004 \\ -0'0044 & 0'0015 & 0'0004 & -0'001 \\ -0'0141 & 0'0004 & 0'0217 & -0'0028 \\ 0'0004 & -0'001 & -0'0028 & 0'0045 \end{pmatrix}.$$

Esta matriz tiene importancia de cara a los contrastes de significación individual ya que entonces se usaran sus elementos de la diagonal principal.
Pasamos a continuación a calcular la bondad del ajuste realizado, es decir, el coeficiente de determinación:

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

Como $SCR = 1'5657$ ya ha sido calculada en la estimación de la varianza de la perturbación aleatoria, tan sólo hay que calcular:

$$SCT = y^t y - n\overline{Y}^2 = 131'13 - 22 \cdot 2'2045^2 = 131'13 - 106'916 = 24'214,$$

donde se ha usado que a partir del primer elemento de $X^t y$, esto es, $\sum_{i=1}^{22} Y_t = 48'5$, se obtiene que $\overline{Y} = \frac{48'5}{22} = 2'2045$. En tal caso:

$$R^2 = 1 - \frac{1'5657}{24'214} = 1 - 0'0647 = 0'9353,$$

esto es, la estimación realizada explica un 93’53% de la variabilidad de Y.
Ahora bien, como es sabido, cuanto más cercano al 100% mejor será el coeficiente de determinación y, por tanto, la estimación realizada. ¿Está en este caso suficientemente cerca del 100% como para que la estimación realizada sea significativa?
Como respuesta afirmativa a esta pregunta, el coeficiente de determinación ha de ser superior a la siguiente cota:

$$\frac{\frac{k-1}{n-k} \cdot F_{k-1,n-k}(1-\alpha)}{1+\frac{k-1}{n-k} \cdot F_{k-1,n-k}(1-\alpha)} = \frac{\frac{3}{18} \cdot 3'15991}{1+\frac{3}{18} \cdot 3'15991} = \frac{0'5267}{1'5267} = 0'345,$$

donde se ha usado que $F_{3,18}(0'95) = 3'15991$. Puesto que el R^2 obtenido es superior a dicha cota inferior podemos afirmar que el coeficiente de determinación es significativo, es decir, valida al modelo.
Esta validación del modelo se puede establecer también a partir del contraste de significación conjunta. Bajo el supuesto de normalidad en el modelo rechazaremos $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ si

$$\frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}} > F_{k-1,n-k}(1-\alpha).$$

Para calcular la región de rechazo y tomar una decisión en este contraste planteamos la tabla ANOVA:

Fuentes de variación	Sumas de cuadrados	Grados de libertad	Medias
Explicada	$SCE = 22'6483$	$k-1=3$	$\frac{SCE}{k-1} = 7'5494$
No explicada	$SCR = 1'5657$	$n-k=18$	$\frac{SCR}{n-k} = 0'087$
Total	$SCT = 24'214$		

El único elemento no calculado hasta el momento de la tabla anterior es $SCE = SCT - SCR = 24'214 - 1'5657 = 22'6483$. En tal caso, se tiene para la región de rechazo que:

$$86'7747 > 3'15991,$$

de forma que es evidente que se rechaza la hipótesis nula de que todos los coeficientes pueden ser nulos de forma simultánea. Por tanto, se tiene que la relación existente entre las variables independientes y la dependiente no se debe al azar, validando el modelo.

Para finalizar estudiaremos los contrastes de significación individual. Como es sabido se rechazará la hipótesis $H_0 : \beta_i = 0$ si

$$\left| \frac{\widehat{\beta}_i}{\widehat{\sigma}\sqrt{w_{ii}}} \right| > t_{n-k}\left(1 - \frac{\alpha}{2}\right),$$

donde $\widehat{\sigma}\sqrt{w_{ii}}$ es la raíz cuadrada del elemento (i, i) de la matriz $\widehat{Var\left(\widehat{\beta}\right)}$ y $t_{n-k}\left(1 - \frac{\alpha}{2}\right) = t_{18}(0'975) = 2'10092.$

- $H_0 : \beta_2 = 0$

$$\left.\begin{array}{r} \widehat{\beta}_2 = 0'4862 \\ \widehat{\sigma}\sqrt{w_{22}} = \sqrt{0'0015} = 0'0387 \end{array}\right\} \Longrightarrow \frac{\widehat{\beta}_2}{\widehat{\sigma}\sqrt{w_{22}}} = 12'5633 > 2'10092.$$

- $H_0 : \beta_3 = 0$

$$\left.\begin{array}{r} \widehat{\beta}_3 = 0'3969 \\ \widehat{\sigma}\sqrt{w_{33}} = \sqrt{0'0217} = 0'1473 \end{array}\right\} \Longrightarrow \frac{\widehat{\beta}_3}{\widehat{\sigma}\sqrt{w_{33}}} = 2'6945 > 2'10092.$$

- $H_0 : \beta_4 = 0$

$$\left.\begin{array}{r}\widehat{\beta}_4 = 0'2287 \\ \widehat{\sigma}\sqrt{w_{44}} = \sqrt{0'0045} = 0'0671\end{array}\right\} \Longrightarrow \frac{\widehat{\beta}_4}{\widehat{\sigma}\sqrt{w_{44}}} = 3'4083 > 2'10092.$$

En todos los casos se rechaza la hipótesis nula, lo que se interpreta como que las variables X_2, X_3 y X_4 son significativas.
Como es sabido, para llegar a estas conclusiones también se podrían haber obtenido los intervalos de confianza de cada coeficiente:

$$\widehat{\beta}_i \pm t_{n-k}\left(1 - \frac{\alpha}{2}\right) \cdot \widehat{\sigma} \cdot \sqrt{w_{ii}}, \quad i = 1, 2, 3, 4.$$

Así por ejemplo, para el último coeficiente se tiene que el intervalo de confianza al 95% es:

$$0'2287 \pm 2'10092 \cdot 0'0671 = (0'08772827,\ 0'3696717).$$

Como el cero no pertenece a dicho intervalo se concluirá que el coeficiente correspondiente será distinto de cero.

El intervalo de confianza al 95% para el segundo coeficiente es:

$$0'4862 \pm 2'10092 \cdot 0'0387 = (0'4048944,\ 0'5675056).$$

Al igual que antes se concluirá que el coeficiente correspondiente será distinto de cero.
Para finalizar con el cálculo de intervalos de confianza, obtendremos a continuación el intervalo para la varianza de la perturbación aleatoria:

$$\left[\frac{(n-k)\cdot\widehat{\sigma}^2}{\chi^2_{n-k}\left(1-\frac{\alpha}{2}\right)}, \frac{(n-k)\cdot\widehat{\sigma}^2}{\chi^2_{n-k}\left(\frac{\alpha}{2}\right)}\right] = \left[\frac{SCR}{\chi^2_{n-k}\left(1-\frac{\alpha}{2}\right)}, \frac{SCR}{\chi^2_{n-k}\left(\frac{\alpha}{2}\right)}\right].$$

Puesto que $SCR = 1'56574$, $\chi^2_{n-k}\left(1-\frac{\alpha}{2}\right) = \chi^2_{18}(0'975) = 31'526$ y $\chi^2_{n-k}\left(\frac{\alpha}{2}\right) = \chi^2_{18}(0'025) = 8'231$ es claro que el intervalo para σ^2 es:

$$\left(\frac{1'56574}{31'526}, \frac{1'56574}{8'231}\right) = (0'04966504,\ 0'1902248)\,.$$

Por todo lo expuesto hasta ahora se tiene que el modelo estimado es válido y que las variables de renta familiar, deuda y número de hijos influyen positivamente en el consumo de las familias. Es decir, a mayor renta, deuda y número de hijos mayor consumo familiar. Además, al ser la variable correspondiente a la deuda una variable ficticia, habremos estimado la diferencia esperada en el consumo familiar entre familias con deuda y sin deuda con el mismo nivel de renta y número de hijos. En este caso se obtiene que dicha estimación es positiva, por lo que aquellas familias que tienen algún tipo de deuda consumen más que aquellas que no la tienen.

Supongamos que además del modelo del ejemplo anterior se tienen en cuenta estos otros dos modelos:

$$Y_t = \gamma_1 + \gamma_2 X_{t2} + \gamma_3 X_{t4} + v_t, \qquad (12)$$

$$Y_t = \alpha_1 + \alpha_2 X_{t2}, \qquad (13)$$

de forma que para el modelo (12) se tiene la siguiente información:

$$X^t X = \begin{pmatrix} 22 & 76'2 & 25 \\ 76'2 & 331'8 & 102 \\ 25 & 102 & 53 \end{pmatrix}, \quad X^t y = \begin{pmatrix} 48'5 \\ 204'45 \\ 69'3 \end{pmatrix},$$

mientras que para el modelo (13):

$$X^t X = \begin{pmatrix} 22 & 76'2 \\ 76'2 & 331'8 \end{pmatrix}, \quad X^t y = \begin{pmatrix} 48'5 \\ 204'45 \end{pmatrix}.$$

Se pide seleccionar el modelo más adecuado.

Para el modelo (9) se tiene que $n = 22$, $k = 4$ y $SCR = 1'5657$, por lo que:

$$\begin{aligned}\mathfrak{L} &= -\frac{n}{2} \cdot (1 + \ln(2 \cdot \pi) - ln(n)) - \frac{n}{2} \cdot \ln(SCR), \\ &= -11 \cdot (2'837877 - 3'091042) - 11 \cdot 0'448333 \\ &= -2'1468.\end{aligned}$$

Por tanto:

$$\begin{aligned}AIC &= -2 \cdot (-2'1468) + 8 = 12'2937, \\ BIC &= -2 \cdot (-2'1468) + 4 \cdot 3'091042 = 16'6579, \\ HQC &= -2 \cdot (-2'1468) + 8 \cdot 1'128508 = 13'3218.\end{aligned}$$

Para los modelos (12) y (13) seguirán siendo válidos los valores de $y^t \cdot y$, n y k, sin embargo, habrá que obtener la suma de cuadrados de los residuos de cada modelo.

Para el modelo (12) se verifica que:

$$
\begin{aligned}
SCR &= y^t y - \widehat{\beta}^t X^t y \\
&= 131'13 - (0'2456764,\ 0'4736592,\ 0'2800916) \cdot \begin{pmatrix} 48'5 \\ 204'45 \\ 69'3 \end{pmatrix} \\
&= 131'13 - 128'1653 = 2'9647.
\end{aligned}
$$

En tal caso:

$$
\begin{aligned}
\mathfrak{L} &= -11 \cdot (2'837877 - 3'091042) - 4 \cdot 1'0868 = -9'1697. \\
AIC &= -2 \cdot (-9'1697) + 6 = 24'3394, \\
BIC &= -2 \cdot (-9'1697) + 3 \cdot 3'091042 = 27'6126, \\
HQC &= -2 \cdot (-9'1697) + 6 \cdot 1'128508 = 25'1105.
\end{aligned}
$$

Para el modelo (13) se verifica que:

$$\begin{aligned} SCR &= y^t y - \widehat{\beta}^t X^t y \\ &= 131'13 - (0'3437073,\ 0'5372499) \cdot \begin{pmatrix} 48'5 \\ 204'45 \end{pmatrix} \\ &= 131'13 - 126'5105 = 4'6195. \end{aligned}$$

En tal caso:

$$\begin{aligned} \mathfrak{L} &= -11 \cdot (2'837877 - 3'091042) - 4 \cdot 1'5303 = -14'0483. \\ AIC &= -2 \cdot (-14'0483) + 4 = 32'0967, \\ BIC &= -2 \cdot (-14'0483) + 2 \cdot 3'091042 = 34'2788, \\ HQC &= -2 \cdot (-14'0483) + 4 \cdot 1'128508 = 32'6107. \end{aligned}$$

Atendiendo a la información anterior (resumida en la siguiente tabla), podemos observar que el modelo con menores valores para los criterios de selección es el (9). Por tanto, nos quedaríamos con este modelo a la hora de analizar el consumo familiar.

Criterios Selección	SCR	AIC	BIC	HQC
Modelo (9)	1’5657	12’2937	16’6579	13’3218
Modelo (12)	2’9647	24’3396	27’6127	25’1107
Modelo (13)	4’6194	32’0964	34’2785	32’6105

Lecturas recomendadas

[1] Salmerón, R. y García, C. (2010). Experiencia docente en la asignatura de Métodos Cuantitativos para la Economía y la Empresa para su adaptación al Espacio Europeo de Educación Superior. XXXII Congreso Nacional de Estadística e Investigación Operativa. A Coruña 14-17 septiembre 2010.

[2] Salmerón, R., López, M. y García, C. (2011). Uso de las TICs en la docencia de la Econometría. XXV Congreso Internacional de Economía Aplicada - ASEPELT, 8-11 junio 2011.

[3] Salmerón, R. y Tamayo, J. (2012). Incidencia de la calidad en la producción, explotación y exploración de las empresas. I International workshop on Diffusion Process and Multivariate Analysis. Granada 6 julio 2012.

[4] Salmerón, R. (2012). Entorno de programación R y la econometría: función MUM (online).

[5] Novales, A. (1993). *Econometría*. McGraw Hill. Capítulo 1 (repaso matrices).

Puedes encontrarlas en la dirección web:

http://www.ugr.es/local/romansg/material/WebEco/index.html

Bibliografia

[1] Esteban, M.V., Moral, M.P., Orbe, S., Regúlez, M., Zarraga, A. y Zubia, M. (2009). *Econometría básica aplicada con Gretl*. Sarriko-On, Universidad del País Vasco. Capítulos 2, 3, 4 y 7.

[2] Gujarati, D. (1997). *Econometría*. Ed. McGraw Hill. Capítulos 2, 3, 4, 5 y 7.

[3] Johnston, J. (1989). *Métodos de Econometría*. Ed. Vicens-Vives. Capítulos 2, 4, 5 y 6.

[4] Novales, A. (1993). *Econometría*. McGraw Hill. Capítulos 3 y 4.

[5] Uriel, E., Contreras, D., Moltó, M.L. y Peiró, A. (1990). *Econometría. El Modelo Lineal*. Editorial AC. Capítulos 2, 3, 4, 5, 6 y 8.

[6] Wooldridge, J.M. (2005). *Introducción a la Econometría: Un enfoque moderno*. Thomson. Capítulos 2, 3, 4 y 6, y Apéndices.