Regressió lineal múltiple

Josep Gibergans Bàguena

P08/05057/02312

Universitat Oberta de Catalunya
www.uoc.edu

Índex

El model de regressió múltiple

1. Introducció

La regressió lineal simple ens proporciona un model per explicar la relació entre dues variables: la variable *Y* que anomenem *variable dependent* o *explicada* i la variable *X* que rep el nom de *variable independent* o *explicativa*.

El pes no depèn només de l'alçada

Sabem que el pes (*Y*) està relacionat linealment amb l'alçada (X_1). Però també sabem que pot estar relacionat amb l'edat (X_2), el nombre setmanal d'hores d'esport (X_3), la quantitat de calories totals dels àpats (X_4), etc.

En aquest mòdul tindrem en compte que en la realitat gairebé sempre són més d'un els factors o variables que influeixen en els valors d'una altra variable i definirem un nou model.

Exemples de variables afectades per més d'una variable

El preu d'un ordinador depèn de la velocitat del processador, de la capacitat del disc dur, de la quantitat de memòria RAM, etc.

El sou d'un titulat per la UOC depèn de l'edat, dels anys que fa que va acabar els estudis, dels anys d'experiència a l'empresa, etc.

El preu de lloguer d'un pis depèn dels metres quadrats de superfície, de l'edat de la finca, de la proximitat al centre de la vila, etc.

El preu d'un cotxe depèn de la potència del motor, del nombre de portes i multitud d'accessoris que pot portar: coixí de seguretat (*air bag*), ordinador de viatge, equip d'alta fidelitat, volant esportiu, llantes especials, etc.

El plantejament del problema és idèntic al realitzat per a la regressió simple, amb la diferència que en aquest cas no tindrem una variable explicativa, sinó diverses. En aquest cas, serà molt útil la notació matricial.

2. El model de regressió lineal múltiple

El **model de regressió lineal múltiple** és una generalització del model de regressió lineal simple, en el qual relacionem la variable que volem explicar *Y* amb les *k* variables explicatives X_1, X_2, ..., X_k. El trobarem a partir dels valors (x_i, y_i) que prenen aquestes variables sobre els elements d'una mostra i mitjançant l'expressió següent:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

Notació

A la variable *Y* se l'anomena *variable dependent* o *explicada*. Les variables X_i reben el nom de *variables independents* o *explicatives*.

Expressant aquesta equació per a cada observació de la mostra, obtenim el sistema d'equacions següent:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + e_1 \\
y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + e_2 \\
&\ldots \quad \ldots \quad \ldots \quad \ldots \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \ldots \\
y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + e_n
\end{aligned}$$

Podem representar aquest sistema de forma matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$$

De manera que podem escriure el model de la forma següent:

$$y = X\beta + e$$

on:

- y: és el vector $(n \times 1)$ d'observacions de la variable Y.
- X: és la matriu $n \times (k + 1)$ d'observacions. A partir de la segona columna, cada columna x_i té les observacions corresponents a cadascuna de les variables que considerem.
- β : és el vector $(k + 1) \times 1$ dels coeficients de la regressió.
- e: és el vector $(n \times 1)$ dels residus o errors.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Suposem que estem interessats a explicar les despeses (en desenes d'euros/any) dels ordinadors d'un departament comercial a partir de la seva edat (en anys) i del nombre d'hores diàries que treballen (hores/dia).

Hem pres una mostra de cinc ordinadors i n'hem obtingut els resultats següents:

Despeses (Y) (desenes d'euros/any)	Antiguitat (X_1) (anys)	Hores de treball (X_2) (hores/dia)
24,6	1	11
33,0	3	13
36,6	4	13
39,8	4	14
28,6	2	12

Volem trobar un model de regressió de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Si desenvolupem aquesta equació en totes les observacions de la mostra, obtenim el sistema d'equacions següent:

$$\begin{cases} y_1 = \beta_0 + \beta_1 + 11\ \beta_2 + \dots + e_1 \\ y_2 = \beta_0 + 3\ \beta_1 + 13\ \beta_2 + \dots + e_2 \\ y_3 = \beta_0 + 4\ \beta_1 + 13\ \beta_2 + \dots + e_3 \\ y_4 = \beta_0 + 4\ \beta_1 + 14\ \beta_2 + \dots + e_4 \\ y_5 = \beta_0 + 2\ \beta_1 + 12\ \beta_2 + \dots + e_5 \end{cases}$$

Que podem escriure matricialment com a $y = X\,\beta + e$, on

$$y = \begin{bmatrix} 24{,}60 \\ 33{,}00 \\ 36{,}60 \\ 39{,}80 \\ 28{,}60 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

En el model de regressió lineal múltiple, que hem expressat matricialment com:

$$y = X\beta + e$$

- $X\,\beta$ és la part corresponent a la variació de y que queda explicada per les variables X_i.

- e és un terme que anomenem dels *residus* o *errors* i que d'alguna manera recull l'efecte de totes aquelles variables que també afecten y i que no es troben incloses en el model perquè són desconegudes o no se'n tenen dades. Sobre aquest terme farem dues suposicions importants:

1. Els errors es distribueixen segons una distribució normal de mitjana zero i una variància σ^2.

2. Els errors són independents.

Amb aquestes dues suposicions, tenim dues conseqüències importants:

1. Fixant uns valors $x_1, x_2, ..., x_k$ de les variables $X_1, X_2, ..., X_k$ i prenent valors esperats sobre l'equació del model, tenim que:

$$E(Y|x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Recordem que...

... en el model de regressió lineal simple la recta de regressió passa per $(x_i, E(y))$.

2. Així mateix, la variància de la distribució de Y és constant:

$$Var(Y|x_1, x_2, ..., x_k) = \sigma^2$$

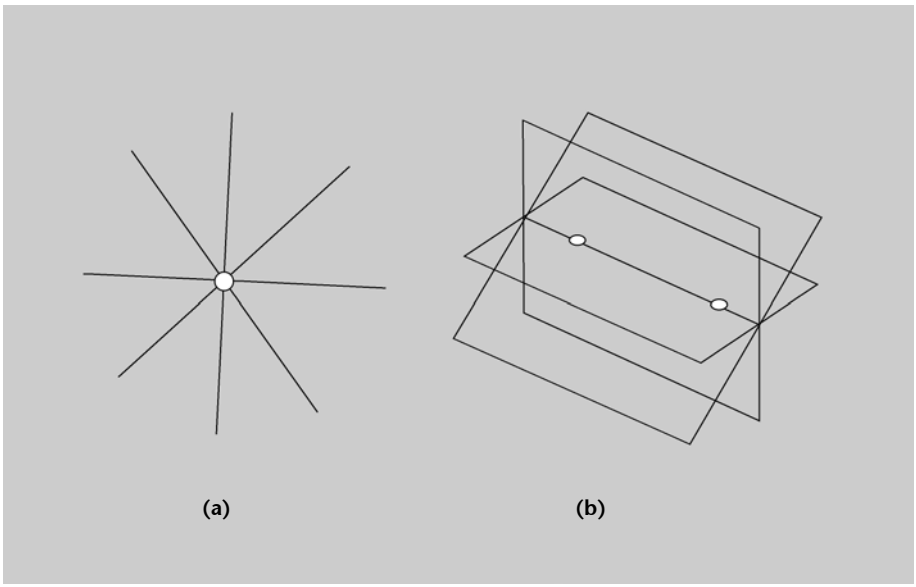
Hi afegirem un parell de suposicions addicionals sobre el model:

1. No podem tenir més paràmetres a estimar $(k + 1)$ que dades disponibles (n) i, per tant, $n > k + 1$.

En el cas del model lineal simple resulta clar que si tenim més paràmetres que dades, tenim una única dada. És impossible trobar quina és la recta que millor s'ajusta a un sol punt, ja que tenim infinites rectes que passen per aquest punt.

Es podria aplicar aquest mateix raonament si tinguéssim més variables explicatives, tot i que seria difícil de visualitzar.

En el cas del model lineal múltiple en què tenim dues variables explicatives, el nombre de paràmetres que cal estimar és tres. Si resulta que tenim dues o menys dades, és a dir, com a molt dos punts, tampoc no té sentit buscar un model de regressió, ja que tenim un nombre infinit de plans que passen per dos punts fixats.



Llegenda

a) Model de regressió lineal simple amb una observació

b) Model de regressió múltiple amb dues variables explicatives i dues observacions

2. Cap de les variables explicatives no pot ser combinació lineal de les altres, perquè en aquest cas no tindríem un model de k variables, sinó de k –1 variables (volem que les variables X_i siguin independents):

Per exemple, si: $X_2 = a + b\ X_1$, aleshores:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e = \\ &= \beta_0 + \beta_1 x_1 + \beta_2 (a + b x_1) + \ldots + \beta_k x_k + e = \\ &= (\beta_0 + a) + (\beta_1 + b) x_1 + \beta_3 x_3 + \ldots + \beta_k x_k + e = \\ &= \beta'_0 + \beta'_1 x_1 + \beta_3 x_3 + \ldots + \beta_k x_k + e \end{aligned}$$

Tenim només k – 1 variables.

3. Ajust del model: mètode dels mínims quadrats

Per a determinar els paràmetres de la recta de regressió en el model lineal simple, vam fer servir el mètode del mínims quadrats. Aquest mètode consisteix a trobar la recta que fa mínima la suma dels residus al quadrat.

Residu en el model de regressió lineal simple

En el model de regressió lineal simple el residu és la diferència entre els valor observat de la variable Y i el valor estimat sobre una recta.

En el cas que ara ens ocupa, procedirem d'una forma molt similar. Buscarem la suma dels residus al quadrat i després determinarem els paràmetres del model que fan que aquesta suma tingui un valor mínim.

Definirem els residus com la diferència entre els valors observats en la mostra (y_i) i els valors estimats pel model ($\hat{y}_i$):

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

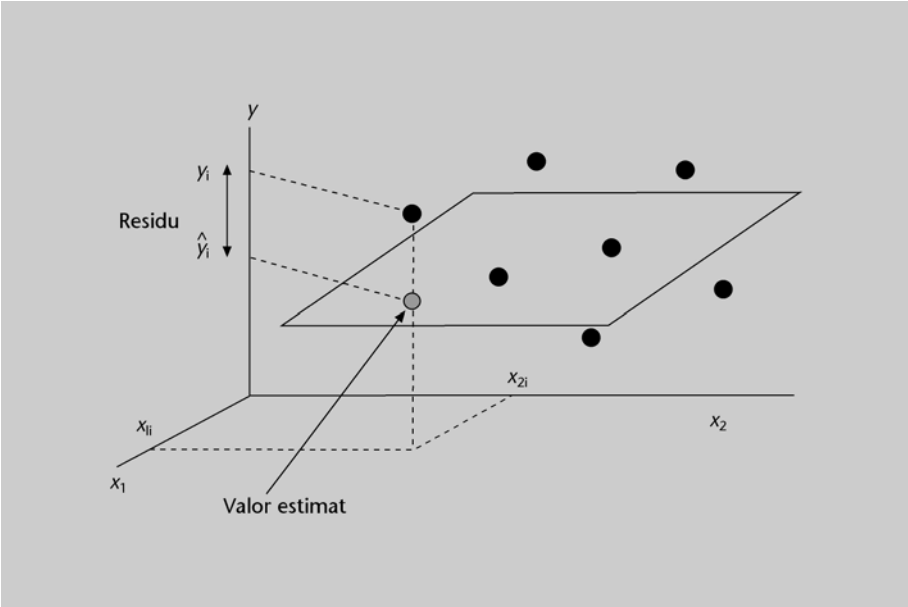
on x_{1i} i x_{2i} són dues observacions de les variables X_1 i X_2, respectivament.

Si considerem un model de regressió lineal múltiple amb dues variables explicatives X_1 i X_2, els residus vindran donats per:

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$

Geomètricament, el podem interpretar com la diferència entre el valor observat i el valor estimat sobre un pla. Els paràmetres del model es determinen trobant el pla que fa mínima la suma dels residus al quadrat. Aquest pla se l'anomena *pla de regressió per mínims quadrats*.

Representem el residu per a un model de regressió múltiple amb dues variables explicatives.



En un model de regressió múltiple amb k variables explicatives, tenim l'expressió següent per als residus:

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) \text{ per a } i = 1, 2, \ldots, n$$

que matricialment podem escriure

$$\begin{bmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \ldots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix}$$

$$e = y - \hat{y} = y - X\beta$$

on e és el vector dels residus, $\hat{y}$ és el vector de les estimacions de y i β és el vector dels paràmetres de la regressió.

Per a calcular la suma dels quadrats dels elements d'un vector, cal fer el producte escalar del vector per si mateix, o el que és el mateix, el producte matricial del vector transposat pel mateix vector.

Si ho fem amb el vector dels residus e:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - \hat{y}_i)^2 = (y - X\beta)^t (y - X\beta)$$

Fent ara els productes i utilitzant algunes propietats del càlcul matricial, obtenim la suma dels quadrats dels residus:

$$\sum_{i=1}^{n} e_1^2 = (y - X\beta)^t (y - X\beta) = y^t y - 2\beta X^t y + \beta^t X^t X\beta$$

Per a trobar els valor dels paràmetres que fan mínima aquesta suma, hem de derivar parcialment respecte dels paràmetres:

$$\frac{\partial}{\partial\beta}\left(\sum_{i=1}^{n} e_1^2 \right) = -2X^t y + 2X^t X\beta$$

I trobar aquells valors que fan nul·les aquestes derivades parcials:

$$\frac{\partial}{\partial\beta}\left(\sum_{i=1}^{n} e_1{}^2 \right) = 0 \Rightarrow -2\, X^t y + 2X^t X\beta = 0$$

Notació

Els estimadors dels paràmetres de la regressió que busquem són les solucions d'aquesta equació matricial, així que posem el "barret" que ens indica que es tracta d'estimadors.

Simplificant una mica, tenim $X^t X\hat{\beta} = X^t y$.

Podem aïllar el vector de paràmetres incògnita:

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

El vector $\hat{\beta}$ és el vector dels estimadors mínim quadràtics del paràmetres.

Finalment, només queda per comentar que, si a l'equació $X^tX\hat{\beta} = X^ty$ efectuem la multiplicació matricial, obtenim el sistema d'equacions següent, anomenat *sistema d'equacions normals de la regressió*:

$$\begin{vmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \dots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \dots & \sum_{i=1}^{n} x_{ik}x_{i1} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & \dots & \sum_{i=1}^{n} x_{ik}x_{i2} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \dots & \sum_{i=1}^{n} x_{ik}^2 \end{vmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{bmatrix} = \begin{vmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \dots \\ \sum_{i=1}^{n} x_{ik}y_i \end{vmatrix}$$

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Tenim:

$$y = \begin{bmatrix} 24{,}60 \\ 33{,}00 \\ 36{,}60 \\ 39{,}80 \\ 28{,}60 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix}$$

La matriu transposada de la matriu X és:

$$X^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 4 & 2 \\ 11 & 13 & 13 & 14 & 12 \end{bmatrix}$$

De manera que:

$$X^tX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 4 & 2 \\ 11 & 13 & 13 & 14 & 12 \end{bmatrix} \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix} = \begin{bmatrix} 5 & 14 & 63 \\ 14 & 46 & 182 \\ 63 & 182 & 799 \end{bmatrix}$$

Si calculem la inversa d'aquesta matriu:

$$(X^tX)^{-1} = \begin{bmatrix} 4 & 14 & 63 \\ 14 & 46 & 182 \\ 63 & 182 & 799 \end{bmatrix}^{-1} = \begin{bmatrix} 181{,}5 & 14 & -17{,}5 \\ 14 & 1{,}3 & -1{,}4 \\ -17{,}5 & -1{,}4 & 1{,}7 \end{bmatrix}$$

D'altra banda, tenim:

$$X^ty = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 4 & 2 \\ 11 & 13 & 13 & 14 & 12 \end{bmatrix} \begin{bmatrix} 24{,}60 \\ 33{,}00 \\ 36{,}6 \\ 39{,}8 \\ 28{,}6 \end{bmatrix} = \begin{bmatrix} 162{,}60 \\ 486{,}40 \\ 2075{,}80 \end{bmatrix}$$

I el vector dels paràmetres estimats de la regressió és:

$$\beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1}(X^t y) = \begin{bmatrix} 181,5 & 14 & -17,5 \\ 14 & 13 & -1,4 \\ -17,5 & -1,4 & 1,7 \end{bmatrix} \begin{bmatrix} 162,60 \\ 486,40 \\ 2075,80 \end{bmatrix} = \begin{bmatrix} -5 \\ 2,6 \\ 2,4 \end{bmatrix}$$

L'equació de regressió és, doncs

$$\hat{y} = -5,0 + 2,6x_1 + 2,4x_2$$

4. Interpretació dels paràmetres

De la mateixa manera que en la regressió lineal, una vegada obtingut el model de regressió lineal múltiple, és molt important fer una bona interpretació dels resultats obtinguts. De moment, només hem obtingut els paràmetres estimats del model de regressió:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

Per a interpretar-los correctament, hem de tenir present el fenomen que estudiem.

1. Interpretació de $\hat{\beta}_0$.:

Aquest paràmetre representa l'estimació del valor de Y quan totes les X_j prenen valor zero. No sempre té una interpretació lligada al context (geomètrica, física, econòmica...). Perquè sigui possible interpretar-lo, necessitem que:

a. Sigui realment possible que les $X_j = 0$.
b. S'han de tenir suficients observacions a prop dels valors $X_j = 0$.

2. Interpretació de $\hat{\beta}_j$:

Representa l'estimació de l'increment que experimenta la variable Y quan X_j augmenta el seu valor en una unitat i la resta de les variables es mantenen constants.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Continuant amb l'exemple dels ordinadors i a partir dels resultats obtinguts en l'ajust:

$$\hat{y} = -5,0 + 2,6x_1 + 2,4x_2$$

1. $\hat{\beta}_0 = -5$ (per desena d'euros)

Ens indica les despeses en desenes d'euros d'un ordinador amb zero anys d'antiguitat i zero hores setmanals de treball. És evident que en aquest exemple no té cap sentit.

2. $\hat{\beta}_1 = 2,6$ (per desena d'euros/any d'antiguitat)

Ens indica l'increment de les despeses en desenes d'euros per cada any d'antiguitat de l'ordinador, sense tenir en compte el nombre d'hores diàries d'ús. Així, doncs, per cada

any que passi tindrem $2,6 \cdot 10 = 26$ euros més en les despeses de manteniment d'un ordinador.

3. $\hat{\beta}_2 = 2,4$ (en desenes d'euros/hores diàries de treball)

Ens indica l'increment en les despeses en desenes d'euros per cada hora diària d'ús sense tenir en compte l'antiguitat de l'ordinador. Tenim que per cada hora de més de feina tindrem $2,4 \cdot 10 = 24$ euros més en les despeses anuals de manteniment d'un ordinador.

5. Resum

En aquesta sessió s'ha presentat el model de regressió lineal múltiple com una generalització del model de regressió lineal simple en aquells casos en què es té més d'una variable explicativa. Hem vist la manera de buscar els paràmetres del model pel mètode dels mínims quadrats i també la comoditat que pot suposar l'ús de la notació matricial a l'hora d'expressar i fer els càlculs.

Exercicis

1. Les dades següents s'han obtingut experimentalment per a determinar la relació entre el guany de corrent (y), el temps de difusió (x_1) i la resistència (x_2) en la fabricació d'un determinat tipus de transistor:

Y	5,3	7,8	7,4	9,8	10,8	9,1	8,1	7,2	6,5	12,6
X_1 (hores)	1,5	2,5	0,5	1,2	2,6	0,3	2,4	2,0	0,7	1,6
X_2 (ohm-cm)	66	87	69	141	93	105	111	78	66	123

Us demanem el següent:

a) Especifiqueu un model lineal múltiple per expressar el guany de corrent en termes del temps de difusió i de la resistència.

b) Estimeu els paràmetres del model de regressió lineal múltiple.

2. Es realitza un experiment per a veure si és possible determinar el pes d'un animal després d'un període de temps determinat a partir del seu pes inicial i de la quantitat d'aliment que li és subministrat. A partir les resultats obtinguts per a una mostra de $n = 10$:

$$\sum_{i=1}^{10} x_{i1} = 379\,,\ \sum_{i=1}^{10} x_{i2} = 2.417\,,\ \sum_{i=1}^{10} x_{i1}^2 = 14.533\,,\ \sum_{i=1}^{10} x_{i2}^2 = 601.365$$

$$\sum_{i=1}^{10} x_{i1}x_{i2} = 92.628\,,\ \sum_{i=1}^{10} y_i = 825\,,\ \sum_{i=1}^{10} x_{i1}y_i = 31.726\,,\ \sum_{i=1}^{10} x_{i2}y_i = 204.569$$

Trobeu l'equació del model de regressió lineal múltiple corresponent.

Solucionari

1.

a) Ara tenim:

Nombre d'observacions: $n = 10$

Nombre de variables independents: 2

Nombre de paràmetres: $k = 2 + 1 = 3$

El model lineal múltiple : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

$$\begin{bmatrix} 5,3 \\ 7,8 \\ 7,4 \\ 9,8 \\ 10,8 \\ 9,1 \\ 8,1 \\ 7,2 \\ 6,5 \\ 12,6 \end{bmatrix} = \begin{bmatrix} 1 & 1,5 & 66 \\ 1 & 2,5 & 87 \\ 1 & 0,5 & 69 \\ 1 & 1,2 & 141 \\ 1 & 2,6 & 93 \\ 1 & 0,3 & 105 \\ 1 & 2,4 & 111 \\ 1 & 2,0 & 78 \\ 1 & 0,7 & 66 \\ 1 & 1,6 & 123 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix}$$

$$Y = \quad X \quad \beta + e$$

b) Estimarem els paràmetres mitjançant el mètode dels mínims quadrats:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1} X^t Y$$

on $(X^t X)^{-1}$ és la matriu inversa de la matriu $(X^t X)$:

$$X^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,5 & 2,5 & 0,5 & 1,2 & 2,6 & 0,3 & 2,4 & 2,0 & 0,7 & 1,6 \\ 66 & 87 & 69 & 141 & 93 & 105 & 111 & 78 & 66 & 123 \end{bmatrix}$$

Atenció

Segons el nombre de xifres decimals que agafeu a partir d'aquí, els resultats poden ser una mica diferents, sense que això vulgui dir que siguin incorrectes.

$$X^t Y = \begin{bmatrix} 84,6 \\ 132,27 \\ 8.320,2 \end{bmatrix}; \qquad X^t X = \begin{bmatrix} 10 & 15,3 & 939 \\ 15,3 & 29,85 & 1.458,9 \\ 939 & 1.458,9 & 94.131 \end{bmatrix}$$

$$(X^t X)^{-1} \begin{bmatrix} 1,7985570396 & -1,855438825 & -0,01506576037 \\ -0,1855438825 & 0,1572804381 & -0,0005867432141 \\ -0,01506576037 & -0,0005867432141 & 0,0001700050851 \end{bmatrix}$$

Ja podem calcular els coeficients:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1} X^t Y = \begin{bmatrix} 2,2680510 \\ 0,2224947452 \\ 0,062325502 \end{bmatrix}$$

Obtenim:

$$\hat{\beta}_0 = 2,2680510\,,\ \hat{\beta}_1 = 0,224947452\,,\ \hat{\beta}_2 = 0,062325502$$

El model de regressió lineal múltiple obtingut és:

$$\hat{y} = 2{,}2680510 + 0{,}224947452x_1 + 0{,}062325502x_2 \ .$$

2. A partir de les equacions normals de la regressió múltiple:

$$\begin{vmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & .... & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & .... & \sum_{i=1}^{n} x_{ik}x_{i1} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & ... & \sum_{i=1}^{n} x_{ik}x_{i2} \\ ... & .... & .... & .... & .... \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & .... & \sum_{i=1}^{n} x_{ik}^2 \end{vmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_4 \end{bmatrix} = \begin{vmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ ... \\ \sum_{i=1}^{n} x_{ik}y_i \end{vmatrix}$$

$$\begin{bmatrix} 10 & 379 & 2417 \\ 379 & 14.533 & 92.628 \\ 2.417 & 92.628 & 601.365 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 825 \\ 31.726 \\ 204.569 \end{bmatrix}$$

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

Aïllant el vector de paràmetres estimats: $\hat{\beta} = (X^tX)^{-1}X^ty$.

Primer, hem de calcular la matriu inversa

$$(X^tX)^{-1} = \begin{bmatrix} 8{,}6176 & -0{,}21777 & -0{,}0010927 \\ -0{,}21777 & 0{,}0092689 & -0{,}00055243 \\ -0{,}0010927 & -0{,}00055243 & 0{,}000091145 \end{bmatrix}$$

Finalment, tenim que

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 8{,}6176 & -0{,}21777 & -0{,}0010927 \\ -0{,}21777 & 0{,}0092689 & -0{,}00055243 \\ -0{,}0010927 & -0{,}00055243 & 0{,}000091145 \end{bmatrix} \begin{bmatrix} 825 \\ 31.726 \\ 204.569 \end{bmatrix} =$$

$$= \begin{bmatrix} -22{,}984 \\ 1{,}395 \\ 0{,}218 \end{bmatrix}$$

El model de regressió lineal múltiple que obtenim és

$$\hat{y} = -22{,}984 + 1{,}395x_1 + 0{,}218x_2$$

La qualitat de l'ajust

1. Introducció

Una vegada trobat el model de regressió lineal múltiple a partir de les dades d'una mostra, el volem utilitzar per a fer inferències a tota la població. Abans, però, és necessari fer una comprovació de la idoneïtat del model obtingut.

En aquesta sessió estudiarem el coeficient de determinació per a la regressió múltiple com a indicador de la qualitat de l'ajust. També farem servir els gràfics dels residus com una important eina de diagnòstic del model.

2. Qualitat de l'ajust. El coeficient de determinació R^2

De la mateixa manera que en la regressió lineal simple, també podem definir ara el **coeficient de determinació R^2** com la proporció de variabilitat explicada pel model respecte a la variabilitat total, és a dir:

$$R^2 = \frac{\text{Variabilitat explicada pel?model}}{\text{Variabilitat total de la mostra}}$$

Terminologia

R també és coneix com a *coeficient de correlació múltiple*.

Si considerem que la variància total observada en la variable Y es descompon en dos termes, la variància explicada pel model de regressió lineal més la variància que no queda explicada pel model, és a dir, la variància dels residus:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

podem expressar el coeficient de determinació així:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

I a partir de les fórmules de les variàncies:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad s_{\hat{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}e_1^2$$

$$s_y^2 = \frac{SQT}{n-1} \qquad s_{\hat{y}}^2 = \frac{SQR}{n-1} \qquad s_e^2 = \frac{SQE}{n-1}$$

on:

$SQT = \sum_{i=1}^{n} (y_i - \bar{y}_2)$ **S**uma de **Q**uadrats **T**otals

$SQR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_2)^2$ **S**uma de **Q**uadrats de la **R**egressió

$SQE = \sum_{i=1}^{n} e_1^2$ **S**uma de **Q**uadrats dels **E**rrors

es pot demostrar que: $SQT = SQR + SQE$.

I tenint en compte que hem definit el coeficient de determinació com a $R^2 = s_{\hat{y}}^2 / s_y^2$, finalment, el podem escriure com a:

$$R^2 = \frac{SQR}{SQT} \qquad \text{o} \qquad R^2 = 1 - \frac{SQE}{SQT}$$

Per a calcular les sumes de quadrats, podem fer servir el càlcul matricial.

- **Suma dels quadrats totals**

Essent D el vector de desviacions de les y_i respecte de la mitjana $\bar{y}$:

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \dots \\ d_n \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{bmatrix}$$

Podem escriure la suma dels quadrats totals de la forma següent:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = D^t D$$

- **Suma dels quadrats de la regressió:**

A partir dels valors estimats:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{bmatrix}$$

podem calcular el vector de les desviacions dels valors estimats $\hat{y}_i$ respecte de la mitjana $\bar{y}$:

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \dots \\ \hat{y}_n - \bar{y} \end{bmatrix}$$

i, per tant, $SQR = \sum_{i=1}^{n} (\hat{y}_1 - \bar{y})^2 = (w)^t w$.

- **Suma dels quadrats dels errors**

A partir dels residus

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \dots \\ y_n - \hat{y}_n \end{bmatrix}$$

és fàcil calcular la suma dels seus quadrats:

$$SQE = \sum_{i=1}^{n} e_1^2 = e^t e$$

De la mateixa manera que en la regressió lineal simple, tenim que el valor del coeficient de determinació està sempre entre 0 i 1: $0 \leq R^2 \leq 1$.

1. $R^2 = 1$ es té quan $SQT = SQR$, és a dir, quan tota la variabilitat de Y s'explica pel model de regressió. En aquest cas, tenim que els valors estimats pel model són exactament iguals als observats.

2. $R^2 = 0$ es té quan $SQR = 0$, és a dir, quan el model no explica absolutament res de Y.

3. Com més gran sigui R^2, més gran serà la proporció de variabilitat de Y explicada pel model i, per tant, més gran serà la bondat de l'ajust.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Considerem una altra vegada l'exemple de les despeses anuals en el manteniment d'un ordinador. Teníem que $\bar{y} = 32{,}52$, de manera que la suma de quadrats totals val:

$$SQT = \sum_{i=1}^{n} (y_1 - \bar{y})^2 = D^t D = (y_1 - \bar{y}\ y_2 - \bar{y}\ \dots y_n - \bar{y}) \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{bmatrix} =$$

$$= (-7,92 \;\; 0,48 \;\; 4,08 \;\; 7,28 \;\; -3,92) \begin{bmatrix} -7,92 \\ 0,48 \\ 4,08 \\ 7,28 \\ -3,98 \end{bmatrix} = 147,97$$

Els valors estimats pel model de regressió múltiple són:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{bmatrix} = X\beta = \begin{bmatrix} 1 & 1 & 11 \\ 1 & 3 & 13 \\ 1 & 4 & 13 \\ 1 & 4 & 14 \\ 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} -5 \\ 2,6 \\ 2,4 \end{bmatrix} = \begin{bmatrix} 24 \\ 34 \\ 36,6 \\ 39 \\ 29 \end{bmatrix}$$

De manera que la suma de quadrats de la regressió és:

$$SQR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (w)^t w = (\hat{y}_1 - \bar{y} \;\; \hat{y}_2 - \bar{y} \;\; \ldots \hat{y}_n - \bar{y}) \begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \ldots \\ \hat{y}_n - \bar{y} \end{bmatrix} =$$

$$= (-8,52 \;\; 1,48 \;\; 4,08 \;\; 6,48 \;\; -3,52) \begin{bmatrix} -8,52 \\ 1,48 \\ 4,08 \\ 6,48 \\ -3,52 \end{bmatrix} = 145,81$$

La diferència entre els valors observats i els valors estimats ens permet d'obtenir els residus:

$$e = \begin{bmatrix} 24,6 \\ 33 \\ 36,6 \\ 39,8 \\ 28,6 \end{bmatrix} - \begin{bmatrix} 24 \\ 34 \\ 36,6 \\ 39 \\ 29 \end{bmatrix} = \begin{bmatrix} 0,6 \\ -1 \\ 0 \\ 0,8 \\ -0,4 \end{bmatrix}$$

Així, la suma dels quadrats dels residus és:

$$SQE = \sum_{i=1}^{n} e_1^2 = e^t e = (e_1 \;\; e_1 \;\; \ldots \;\; e_n) \begin{bmatrix} e_1 \\ e_1 \\ \ldots \\ e_n \end{bmatrix} = (0,6 \; -1 \; 0 \; 0,8 \; -0,4) \begin{bmatrix} 0,6 \\ -1 \\ 0 \\ 0,8 \\ -0,4 \end{bmatrix} = 2,16$$

El coeficient de determinació és:

$$R^2 = \frac{SQR}{SQT} = \frac{145,81}{147,97} = 0,985$$

També es pot calcular fent: $R^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{2.16}{147,97} = 1 - 0,0015 = 0,985$

Aquest resultat ens diu que el model de regressió múltiple obtingut explica el 98,5% de la variabilitat de les despeses dels ordinadors. Com que és molt proper al 100%, en principi és un bon model.

3. L'anàlisi dels residus

De la mateixa manera que en la regressió lineal simple, els residus del model de regressió lineal múltiple tenen un paper important a l'hora de determinar l'adequació del model.

En el cas de regressió lineal múltiple, és habitual construir dos tipus de gràfics:

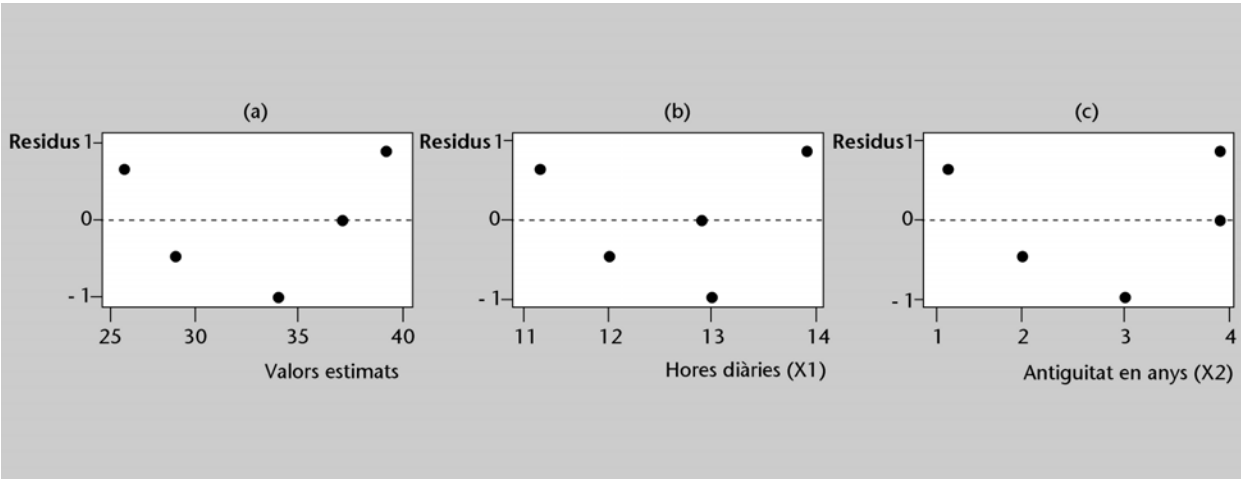
1. Gràfic de *residus enfront de valors estimats*: representem en l'eix d'ordenades els valors dels residus i en l'eix d'abscisses, els valors estimats, de manera que el núvol de punts $(\hat{y}_i, e_i)$ no ha de tenir cap tipus d'estructura i és proper a l'eix d'abscisses.

2. Gràfic de *residus enfront de variables explicatives*: representem sobre l'eix d'ordenades els valors dels residus i sobre l'eix d'abscisses, els valors observats de la variable explicativa. Tenim un gràfic d'aquests tipus per a cadascuna de les variables explicatives.

Sempre que el model sigui correcte, cap gràfic de residus no ha de mostrar cap tipus d'estructura. Els residus sempre han d'estar distribuïts a l'atzar al voltant del zero.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

En el cas dels ordinadors i les seves despeses en manteniment, tenim els gràfics de representació dels residus següents:



Els tres gràfics representen:

a) residus enfront de valors estimats pel model;

b) residus enfront valors de la variable X_1: hores diàries de feina;

c) residus enfront de valors de la variable X_2: antiguitat dels ordinadors en anys.

No observem cap mena d'estructura organitzada dels residus que ens faci pensar en una falta de linealitat del model. Tampoc no observem cap dada atípica.

4. Aplicacions a la predicció

L'aplicació bàsica d'un model de regressió lineal múltiple és predir (estimar) el valor de la variable Y a partir d'un conjunts de valors de les variables independents X_j .

Només cal substituir aquests valors x_i a l'equació de regressió obtinguda:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Considerant una vegada més el problema dels ordinadors, si volem calcular la despesa corresponent a un ordinador que té dos anys d'antiguitat i treballa catorze hores diàries, farem servir l'equació trobada:

$$\hat{y} = -5{,}0 + 2{,}6x_1 + 2{,}4x_2 ,$$

amb $x_1 = 2$ i $x_2 = 14$:

$$\hat{y} = -5{,}0 + 2{,}6 \cdot 2 + 2{,}4 \cdot 14 = -5{,}0 + 5{,}4 + 33{,}6 = 34$$

Per tant, podem esperar una despesa de manteniment de 340 euros anuals per a aquest ordinador.

A l'hora d'aplicar l'equació de regressió trobada, sempre hem de mirar si els valors de les variables X_i per als quals volem estimar el valor de la variable Y, es troben dintre el conjunt de valors que hem fet servir per a construir el model. Si no és així, hem d'anar amb molta cautela, ja que el resultat que ens doni el model pot ser que no tingui cap mena de sentit. El perill de l'extrapolació també el tenim present en la regressió lineal múltiple.

Exemple de resultat irreal

Si volem fer servir el nostre model per a calcular la despesa de manteniment del nostre ordinador quan tingui una antiguitat de cinquanta anys, és evident que no té cap sentit fer servir l'equació trobada: ni l'ordinador existirà d'aquí a cinquanta anys (i si existeix estarà en un museu), ni els preus de manteniment tindran res a veure amb els d'ara, etc.

5. Resum

En aquesta sessió hem estudiat el coeficient de determinació com una mesura de la bondat de l'ajust del model a les dades de la mostra. A continuació s'ha comentat la importància de fer una anàlisi dels residus per tenir un diagnòstic del model lineal obtingut. Hem acabat la sessió amb l'aplicació de la regressió a la predicció, que posa de manifest el perill de l'extrapolació.

Exercicis

1. Les dades següents s'han obtingut experimentalment per a determinar la relació entre el guany de corrent (Y), el temps de difusió (X_1) i la resistència (X_2) en la fabricació d'un determinat tipus de transistor:

Y	5,3	7,8	7,4	9,8	10,8	9,1	8,1	7,2	6,5	12,6
X_1 (hores)	1,5	2,5	0,5	1,2	2,6	0,3	2,4	2,0	0,7	1,6
X_2 (ohm-cm)	66	87	69	141	93	105	111	78	66	123

Si el model de regressió obtingut a partir d'aquestes dades és:

$$\hat{y} = 2{,}268 + 0{,}225x_1 + 0{,}062x_2$$

Feu una anàlisi dels residus i comenteu els resultats obtinguts.

2. Es realitza un experiment per a veure si és possible determinar el pes d'un animal després d'un període de temps determinat a partir del seu pes inicial i de la quantitat d'aliment que li és subministrat. A partir dels resultats obtinguts per a una mostra de $n = 10$:

Pes final (kg)	95	77	80	100	97	70	50	80	92	84
Pes inicial (kg)	42	33	33	45	39	36	32	41	40	38
Aliment (kg)	272	226	259	292	311	183	173	236	230	235

s'ha obtingut el model de regressió lineal:

$$\hat{y} = -22{,}984 + 1{,}395x_1 + 0{,}218x_2$$

Calculeu el coeficient de determinació i interpreteu-lo.

Solucionari

1. Per a fer una anàlisi de residus, hem de construir dos tipus de gràfics:

- Gràfic de *residus enfront de valors estimats*: representarem en el pla el núvol de punts: $(\hat{y}_1,\ e_i)$.

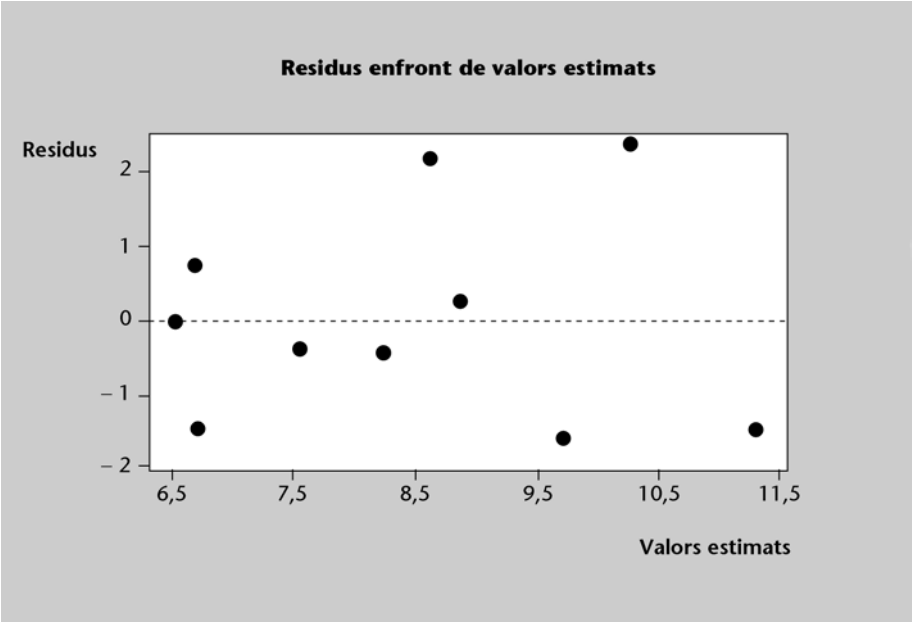
Abans haurem de calcular els valors estimats: $\hat{y} = X\beta$

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\\ \hat{y}_6\\ \hat{y}_7\\ \hat{y}_8\\ \hat{y}_9\\ \hat{y}_{10}\end{bmatrix} = \begin{bmatrix}1 & 1{,}5 & 66\\ 1 & 2{,}5 & 87\\ 1 & 0{,}5 & 69\\ 1 & 1{,}2 & 141\\ 1 & 2{,}6 & 93\\ 1 & 0{,}3 & 105\\ 1 & 2{,}4 & 111\\ 1 & 2{,}0 & 78\\ 1 & 0{,}7 & 66\\ 1 & 1{,}6 & 123\end{bmatrix} \begin{bmatrix}2{,}268\\ 0{,}225\\ 0{,}062\end{bmatrix} = \begin{bmatrix}6{,}71\\ 8{,}25\\ 6{,}68\\ 11{,}32\\ 8{,}64\\ 8{,}87\\ 9{,}72\\ 7{,}57\\ 6{,}53\\ 10{,}29\end{bmatrix}$$

I els residus: $e = \hat{y} - y$

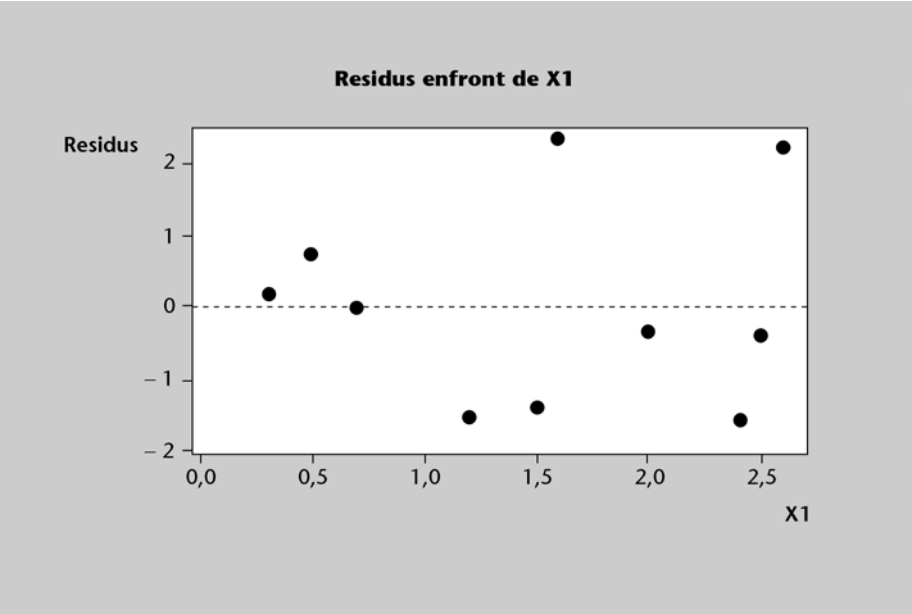
$$\begin{bmatrix}e_1\\ e_2\\ e_3\\ e_4\\ e_5\\ e_6\\ e_7\\ e_8\\ e_9\\ e_{10}\end{bmatrix} = \begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\\ y_9\\ y_{10}\end{bmatrix} - \begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\\ \hat{y}_6\\ \hat{y}_7\\ \hat{y}_8\\ \hat{y}_9\\ \hat{y}_{10}\end{bmatrix} = \begin{bmatrix}5{,}3\\ 7{,}8\\ 7{,}4\\ 9{,}8\\ 10{,}8\\ 9{,}1\\ 8{,}1\\ 7{,}2\\ 6{,}5\\ 12{,}6\end{bmatrix} - \begin{bmatrix}6{,}71\\ 8{,}25\\ 6{,}68\\ 11{,}32\\ 8{,}64\\ 8{,}87\\ 9{,}72\\ 7{,}57\\ 6{,}53\\ 10{,}29\end{bmatrix} = \begin{bmatrix}-1{,}41\\ -0{,}45\\ 0{,}72\\ -1{,}52\\ 2{,}16\\ 0{,}22\\ -1{,}62\\ -0{,}37\\ -0{,}03\\ 2{,}31\end{bmatrix}$$

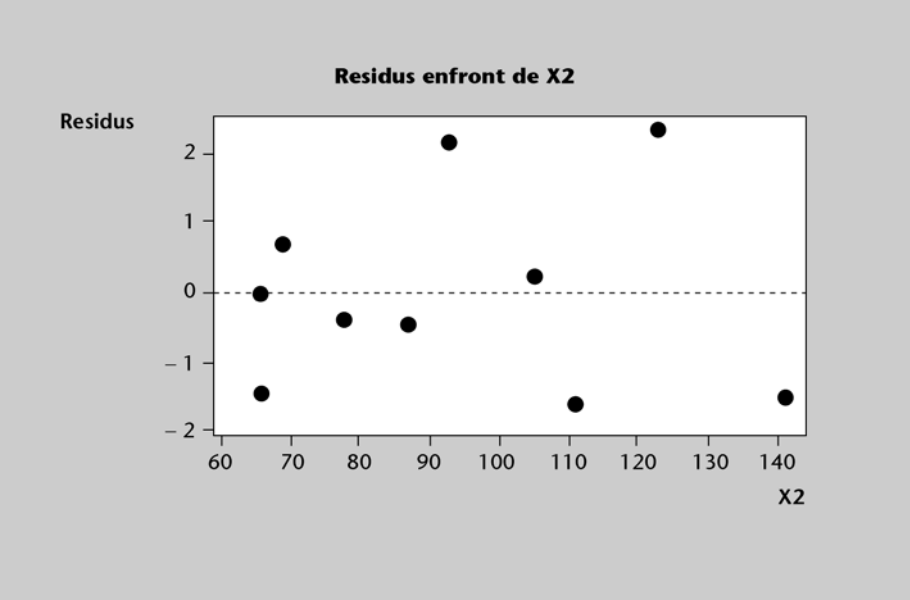
El gràfic resultant és:



No observem cap mena d'estructura en el núvol de punts.

- Gràfics de residus enfront de variables explicatives: ara per cada variable explicativa tenim un gràfic. En aquest gràfic representem (x_{ij}, e_i).





En cap d'aquestes dues representacions no podem veure cap mena d'estructura en els núvols de punts.

2. El podem calcular a partir de qualsevol de les expressions:

$$R^2 = \frac{SQR}{SQT} \qquad R^2 = 1 - \frac{SQE}{SQT} \qquad 0 \le R^2 \le 1$$

Haurem de tenir en compte que, si el calculem de les dues formes, els resultats seran lleugerament diferents a causa de l'error d'arrodoniment associat als càlculs.

Per a calcular la suma de quadrats de la regressió (SQR), ens fa falta conèixer la mitjana de y:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i = 82{,}5$$

I els valors estimats de y_i, $\hat{y}_i$:

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \dots\\ \hat{y}_n\end{bmatrix} = \begin{bmatrix}1 & x_{11} & x_{21} & \dots & x_{k1}\\ 1 & x_{12} & x_{22} & \dots & x_{k2}\\ \dots & \dots & \dots & \dots & \dots\\ 1 & x_{1n} & x_{2n} & \dots & x_{kn}\end{bmatrix} \begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \dots\\ \hat{\beta}_k\end{bmatrix}$$

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\\ \hat{y}_6\\ \hat{y}_7\\ \hat{y}_8\\ \hat{y}_9\\ \hat{y}_{10}\end{bmatrix} = \begin{bmatrix}1 & 42 & 272\\ 1 & 33 & 226\\ 1 & 33 & 259\\ 1 & 45 & 292\\ 1 & 39 & 311\\ 1 & 36 & 183\\ 1 & 32 & 173\\ 1 & 41 & 236\\ 1 & 40 & 230\\ 1 & 38 & 235\end{bmatrix} \begin{bmatrix}-22{,}984\\ 1{,}395\\ 0{,}218\end{bmatrix} = \begin{bmatrix}94{,}778\\ 72{,}216\\ 79{,}396\\ 103{,}314\\ 99{,}079\\ 67{,}0.45\\ 59{,}290\\ 85{,}550\\ 82{,}850\\ 81{,}148\end{bmatrix}$$

I per calcular la suma de quadrats dels errors (*SQE*), ens fa falta el vector d'errors:

$$e = \begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\\ y_9\\ y_{10}\end{bmatrix} - \begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\\ \hat{y}_6\\ \hat{y}_7\\ \hat{y}_8\\ \hat{y}_9\\ \hat{y}_{10}\end{bmatrix} = \begin{bmatrix}0{,}222\\ 4{,}784\\ 0{,}604\\ -3{,}314\\ -2{,}079\\ 2{,}955\\ -9{,}290\\ -5{,}550\\ 9{,}150\\ 2{,}852\end{bmatrix}$$

Les sumes de quadrats són:

- $SQT = \sum(y_i - \bar{y})^2 = 2.020{,}50$
- $SQR = \sum(\hat{y}_i - \bar{y})^2 = 1{,}762{,}99$
- $SQE = \sum(y_i - \hat{y})^2 = \sum e_i^2 = 256{,}30$

Per tant, el coeficient de determinació és $R^2 = \frac{SQR}{SQT} \approx 0{,}873$.

Com que el coeficient de determinació és la relació entre la variància explicada i la variància total, tenim que és bastant proper a 1; per tant, vol dir que tenim bondat en l'ajust. El model de regressió explica el 87,3% de la variabilitat del pes dels animals a partir del seu pes inicial i la quantitat d'aliment.

Inferència en la regressió lineal múltiple

1. Introducció

Una vegada estimat el model de regressió, estem interessats a poder aplicar-lo a la població d'on hem tret la mostra. Ara determinarem intervals de confiança per als paràmetres del model i en farem contrastos d'hipòtesis, per així poder detectar quines són les variables realment significatives. Finalment, comentarem com podem detectar i evitar el problema de la duplicació d'informació que sorgeix quan s'utilitzen variables correlacionades, conegut amb el nom de *multicol·linealitat*.

2. Estimació de la variància dels errors

Donada una mostra d'observacions, el model estarà totalment determinat un cop s'especifiquin els valores estimats dels coeficients β_0, β_1, ... β_k i s'estimi la variància dels errors σ^2. Encara ens falta determinar aquesta última.

Considerant els residus com a estimacions del valors del terme d'error, aleshores podem estimar la variància d'aquest terme a partir de la variància dels residus:

$$s^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki})^2$$

Si tenim en compte que aquest sumatori és la suma dels quadrats dels errors, el podem escriure d'aquesta manera:

$$s^2 = \frac{SQE}{n-k-1}$$

Residus no independents

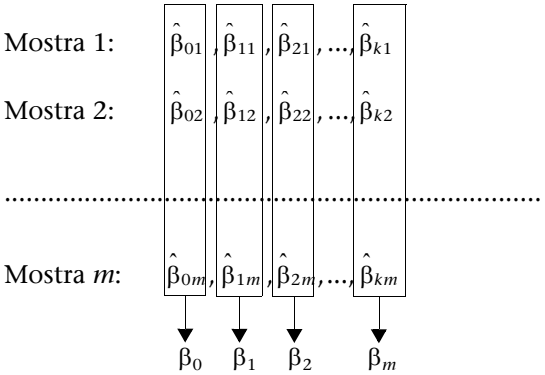
Es divideix per

$n - (k + 1) = n - k - 1$

perquè els n residus no són independents (estan lligats per les $(k + 1)$ equacions normals de la regressió).

3. Distribucions probabilístiques dels paràmetres de la regressió

En primer lloc, ha de quedar molt clar que cada mostra determina una regressió lineal múltiple i, per tant, un conjunt de coeficients:



De manera que tindríem per a cada coeficient de la regressió una col·lecció de valors estimats dels paràmetres:

$\hat{\beta}_{01}, \hat{\beta}_{02}, \hat{\beta}_{03},..., \hat{\beta}_{0m} \longrightarrow \beta_0$

$\hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13},..., \hat{\beta}_{1m} \longrightarrow \beta_1$

....

$\hat{\beta}_{k1}, \hat{\beta}_{k2}, \hat{\beta}_{k3},..., \hat{\beta}_{km} \longrightarrow \beta_k$

Notació

El primer subíndex ens indica el paràmetre i el segon ens indica que és una observació d'aquest obtinguda a partir de la mostra.

Així, $\beta_0, \beta_1,... \beta_k$ són unes variables aleatòries que caldrà estudiar per a poder inferir els nostres resultats a la població d'on hem extret les mostres. Primer, les caracteritzarem calculant-ne els valors esperats i les desviacions estàndard:

a) Valor esperat de $\hat{\beta}_j$: $E(\beta_j) = \beta_j$; per a $j = 1, ..., k$. Observem que els valors esperats d'aquests paràmetres són iguals als valors poblacionals d'aquests. Encara que aquests valors siguin desconeguts, aquest resultat ens serà de gran utilitat a l'hora de fer inferència estadística.

Aquests càlculs es mostren detalladament en els annexos 3.1 i 3.2.

b) Variància de $\hat{\beta}_j$. Les variàncies de les $\hat{\beta}_j$ són els elements de la diagonal de la matriu $\sigma^2(X^tX)^{-1}$, és a dir:

$$\begin{bmatrix} \text{var}(\hat{\beta}_0) & & & \\ & \text{var}(\hat{\beta}_1) & & \\ & & \ldots & \\ & & & \text{var}(\hat{\beta}_k) \end{bmatrix} = \sigma^2 \; diag \; (X^tX)^{-1}$$

Ja hem calculat la mitjana i la variància dels estimadors $\hat{\beta}_j$. Com que la variable Y es distribueix normalment i les $\hat{\beta}_j$ són combinació lineal de les observacions y_j, es pot assegurar que les $\hat{\beta}_j$ es distribuiran normalment:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 \; q_{jj})$$

on q_{jj} és l'element de la fila j i columna j de la matriu $(X^tX)^{-1}$. Com que la variància σ^2 és desconeguda, n'haurem de fer servir el valor estimat a partir de les dades de la mostra, cosa que ja hem fet en l'apartat 1 d'aquesta sessió:

$$s^2 = \frac{SQE}{n-k-1}$$

De manera que:

$$\begin{bmatrix} \text{var}(\hat{\beta}_0) & & & \\ & \text{var}(\hat{\beta}_1) & & \\ & & \ldots & \\ & & & \text{var}(\hat{\beta}_k) \end{bmatrix} = s^2 \; diag \; (X^tX)^{-1}$$

I les desviacions estàndard dels estimadors seran:

$$s_{\hat{\beta}_j} = \sqrt{\text{var}(\hat{\beta}_j)} \text{ , per } j = 1, 2, ..., k$$

Una vegada conegudes les estimacions dels paràmetres, $\hat{\beta}_j$, i de les seves desviacions estàndard, $s_{\hat{\beta}_j}$, escriurem el resultat de la regressió de la forma següent:

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

$$s_{\hat{\beta}_0} \quad s_{\hat{\beta}_1} \quad s_{\hat{\beta}_2} \quad s_{\hat{\beta}_k}$$

$$s^2 \quad R^2$$

És a dir:

1) Escrivim el model de regressió lineal obtingut.

2) Sota cadascun dels paràmetres estimats escrivim la seva desviació típica.

3) Per últim, a la ratlla següent escrivim l'estimació de la variància dels residus i el coeficient de determinació.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Continuem amb el cas del qual volíem explicar les despeses (en desenes d'euros/any) dels ordinadors d'un departament comercial a partir de la seva edat (en anys) i dels nombre d'hores diàries que treballen (hores/dia). Amb aquesta finalitat, s'havia pres una mostra de cinc ordinadors i s'havien obtingut els resultats següents:

Despeses (Y) (desenes d'euros/any)	Antiguitat (X_1) (anys)	Hores de treball (X_2) (hores/dia)
24,6	1	11
33,0	3	13
36,6	4	13
39,8	4	14
28,6	2	12

El model de regressió obtingut era el següent: $\hat{y} = -5{,}0 + 2{,}6x_1 + 2{,}4x_2$. Havíem trobat:

$$(X^tX)^{-1} = \begin{bmatrix} 181{,}4 & 14 & -17{,}5 \\ 14 & 1{,}3 & -1{,}4 \\ -17{,}5 & -1{,}4 & 1{,}7 \end{bmatrix}, \text{ i també } s^2 = \frac{SQE}{n-k-1} = \frac{2{,}16}{5-2-1} = \frac{2{,}16}{2} = 1{,}08$$

De manera que:

- $\text{var}(\hat{\beta}_0) = 1{,}08 \cdot 181{,}4 = 195{,}91 \Rightarrow s_{\hat{\beta}_0} = 14{,}0$
- $\text{var}(\hat{\beta}_1) = 1{,}08 \cdot 1{,}3 = 1{,}404 \Rightarrow s_{\hat{\beta}_1} = 1{,}18$
- $\text{var}(\hat{\beta}_2) = 1{,}08 \cdot 1{,}7 = 1{,}836 \Rightarrow s_{\hat{\beta}_2} = 1{,}35$

Podem escriure els resultats de la manera següent:

$$\begin{array}{ccccc} y = -5{,}0 & + & 2{,}6x_1 & + & 2{,}4x_2 \\ (14{,}0) & & (1{,}18) & & (1{,}35) \end{array}$$

$$S^2 = 1{,}08 \qquad R^2 = 0{,}985$$

4. Intervals de confiança dels paràmetres del model

En els models de regressió lineal múltiple és d'utilitat construir estimacions d'intervals de confiança per als coeficients de la regressió $\hat{\beta}_j$. Com hem vist en l'apartat anterior, els estimadors $\hat{\beta}_j$ segueixen distribucions $N(\beta_j, s^2_{\hat{\beta}_j})$. Per tant, es pot demostrar que la variable tipificada:

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}}$$

segueix una **distribució *t* de Student amb *n* – *k* – 1 graus de llibertat**. Com que:

$$P\left(-t_{\alpha/2,n-k-1} \leq \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \leq t_{\alpha/2,n-k-1}\right) = 1 - \alpha$$

Un **interval de confiança** amb un nivell de confiança de $100(1 - \alpha)\%$ per al coeficient $\hat{\beta}_j$ de la regressió ve donat per:

$$[\hat{\beta}_j - t_{\alpha/2,n-k-1} s_{\hat{\beta}_j},\ \hat{\beta}_j + t_{\alpha/2,n-k-1} s_{\hat{\beta}_j}]$$

On $\hat{\beta}_j$ és el valor estimat del paràmetre a partir de la mostra.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Calculem ara els intervals de confiança per als paràmetres $\hat{\beta}_1$ i $\hat{\beta}_2$ del nostre exemple:

- Interval de confiança per a $\hat{\beta}_1$ amb un nivell de confiança del 95%. Mirant la taula de la distribució *t* de Student amb $n - k - 1 = 5 - 2 - 1 = 2$ graus de llibertat, el valor crític corresponent per a $\alpha/2 = 0{,}025$ és: $t_{0{,}025;2} = 4{,}3027$. L'interval de confiança serà:

$$[2{,}6 - 4{,}3027 \cdot 1{,}18;\ 2{,}6 + 4{,}3027 \cdot 1{,}18] = [-2{,}50;\ 7{,}70]$$

- Interval de confiança per a $\hat{\beta}_2$ amb un nivell de confiança del 95%. Ara l'interval de confiança serà:

$$[2{,}4 - 4{,}3027 \cdot 1{,}35;\ 2{,}4 + 4{,}3027 \cdot 1{,}35] = [-3{,}43;\ 8{,}23]$$

5. Contrast d'hipòtesi sobre els paràmetres del model

Moltes vegades és interessant fer testos d'hipòtesi sobre els coeficients de la regressió. Gairebé sempre ens interessarà de saber si un coeficient β_i és igual a zero, ja que això voldria dir que la variable X_i corresponent no figura en el model de regressió i, per tant, no és una variable explicativa del comportament de la variable Y.

Per a fer aquest contrast d'hipòtesis, seguim el procediment següent:

1) Establim les hipòtesis. Per cada β_j:

- Hipòtesi nul·la: H_0: $\beta_j = 0$ (la variable X_j no és explicativa)
- Hipòtesi alternativa: H_1: $\beta_j \neq 0$

En cas que no rebutgem la hipòtesi nul·la, voldrà dir que la variable X_j no és una variable explicativa i que, per tant, la podem eliminar del model.

2) Calculem l'estadístic de contrast: si la hipòtesi nul·la és certa ($\beta_j = 0$), aleshores obtenim l'estadístic de contrast:

$$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

que és una observació d'una distribució t de Student amb $n - k - 1$ graus de llibertat.

3) Finalment, a partir d'un nivell de significació (α) establirem un criteri de decisió. Per a fer això, tenim dues opcions:

a) A partir del p-valor. El p-valor és la probabilitat del resultat observat o d'un de més allunyat si la hipòtesi nul·la és certa. És a dir:

$$p = 2P(t_{n-k-1} > |t|)$$

- Si $p \leq \alpha$, es rebutja la hipòtesi nul·la H_0
- Si $p > \alpha$, no es rebutja la hipòtesi nul·la H_0

b) A partir dels valors crítics $\pm t_{\alpha/2;n-k-1}$, de manera que:

- Si $|t| > t_{\alpha/2;n-k-1}$, es rebutja la hipòtesi nul·la H_0; per tant, la variable X_j és una variable explicativa de la variable Y i, per tant, no la podem eliminar del model.

- Si $|t| \leq t_{\alpha/2,n-k-1}$, no es rebutja la hipòtesi nul·la H_0; per tant, la variable X_j no és una variable explicativa de la variable Y i, per tant, la podem eliminar del model.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Tornem al nostre exemple per fer un contrast d'hipòtesi sobre els paràmetres de la regressió i assabentar-nos de si les variables són explicatives de les despeses anuals de manteniment dels ordinadors o no. Farem servir un nivell de significació $\alpha = 0{,}05$.

- Contrast per β_1

1. Establim les hipòtesis nul·la i alternativa:

– Hipòtesi nul·la: H_0: $\beta_1 = 0$
– Hipòtesi alternativa: H_1: $\beta_1 \neq 0$

2. Calculem l'estadístic de contrast: $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{2{,}6}{1{,}18} = 2{,}20$.

3. Calculem el *p*-valor corresponent a aquest estadístic de contrast:

$$p = 2\,P(t_{n-k-1} > |t|) = 2\,P(t_2 > 2{,}20) = 2 \cdot 0{,}0794 = 0{,}1588$$

Com que 0,1588 > 0,05, no rebutgem H_0. Per tant, la variable X_1 no és una variable explicativa i, per tant, la podem eliminar del model.

- Contrast per β_2

1. Establim les hipòtesis:

– Hipòtesi nul·la: H_0: $\beta_2 = 0$
– Hipòtesi alternativa: H_1: $\beta_2 \neq 0$

2. Calculem l'estadístic de contrast: $t = \dfrac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \dfrac{2{,}4}{1{,}35} = 1{,}77$.

3. Calculem el *p*-valor corresponent a aquest estadístic de contrast:

$$p = 2P(t_{n-k-1} > |t|) = 2P(t_2 > 1{,}77) = 2 \cdot 0{,}1094 = 0{,}2188$$

Com que 0,2188 > 0,05, no rebutgem H_0. Per tant, la variable X_2, tampoc no és una variable explicativa i, per tant, la podem eliminar del model.

En aquest model de regressió lineal múltiple, cap de les dues variables no ens explica la variable "despesa en manteniment".

6. Contrastació conjunta del model

Hem vist com cal fer el contrast d'hipòtesi per a veure si cadascuna de les variables X_i, individualment contribueixen a explicar la variable Y.

Ara volem contrastar el model de forma global, tenint en compte totes les variables X_i que hem fet servir per a trobar-lo.

1) Establim les hipòtesis:

- Hipòtesi nul·la: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. Ens indica que no hi ha relació lineal entre la variable Y i cap de les variables X_i.

- Hipòtesi alternativa: H_1: al menys una $\beta_0 \neq 0$

Altres maneres d'expressar les hipòtesis

Una altra forma d'expressar aquestes hipòtesis és la següent:

Hipòtesi nul·la:

$$H_0: R^2 = 0$$

Ens indica que la part de la variació explicada pel model és zero, és a dir, que no hi ha cap relació lineal entre la variable Y i qualsevol de les variables X_i.

Hipòtesi alternativa:

$$H_1: R^2 > 0$$

2) Calculem l'estadístic de contrast.

Aquesta prova es basa en un estadístic de contrast que és una observació d'una distribució F quan H_0 es certa.

Buscarem una relació entre la variació explicada pel model de regressió múltiple i la no explicada pel mateix model. Si la proporció de variació explicada en relació amb la no explicada és gran, aleshores es confirmarà la utilitat del model i no rebutjarem la hipòtesi nul·la H_0.

Recordem que...

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

$$SQE = \sum_{i=1}^{n} e_i^2$$

A partir de la descomposició de la suma de quadrats totals segons la suma de quadrats de la regressió més la suma dels quadrats dels errors:

$$SQT = SQR + SQE$$

- *SQT*: és la suma de quadrats que, dividida per $(n - 1)$, ens dóna la variància mostral de la variable Y. Aquesta suma te $n - 1$ graus de llibertat.
- *SQE*: és la suma dels quadrats dels errors, que com ja hem comentat en més d'una ocasió, té $(n - k + 1)$ graus de llibertat.
- *SQR*: és la suma dels quadrats de la regressió. Aquesta quantitat té k graus de llibertat.

Sota la hipòtesi nul·la, H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$:

- *SQR* té una distribució χ^2 amb k graus de llibertat
- *SQE* té una distribució χ^2 amb $n - k - 1$ graus de llibertat
- *SQR* i *SQE* són independents

El quocient de dues variables χ^2 dividides pels seus graus de llibertat dóna una variable F de Snedecor amb els graus de llibertat corresponent al numerador i denominador del quocient.

Podem definir, doncs, l'**estadístic de contrast**:

$$f = \frac{SQR/k}{SQE/(n-k-1)}$$

És una observació d'una distribució F de Snedecor amb k i $(n - k - 1)$ graus de llibertat.

Si la hipòtesi nul·la és certa i, per tant, no hi ha cap relació lineal entre Y i les variables X_i, l'estadístic tindrà un valor prop d'u. Però quan hi ha una certa relació, la suma dels quadrats de la regressió (numerador) augmenta i la suma dels quadrats dels errors (denominador) disminueix, de manera que el valor de l'estadístic de contrast augmenta. Si aquest valor supera un valor crític de la distribució F, aleshores rebutgem la hipòtesi nul·la.

3) Establim un criteri de decisió a partir d'un nivell de significació α:

A partir d'aquest valor crític de la distribució F de Snedecor:

- Si $f > F_{\alpha;k;n-k-1}$, rebutgem H_0; per tant, el model explica significativament la variable Y. És a dir, el model sí que contribueix amb informació a explicar la variable Y.

- Si $f < F_{\alpha;k;n-k-1}$, no rebutgem H_0; per tant, el model no explica de forma significativa la variable Y.

També ho podem fer a partir del p-valor: $p = P(F_{\alpha;k;n-k-1} > f)$.

- Si $p \leq \alpha$, es rebutja la hipòtesi nul·la H_0.
- Si $p > \alpha$, no es rebutja la hipòtesi nul·la H_0.

Els càlculs necessaris es poden resumir en la taula següent, coneguda com a **taula d'anàlisi de la variància**:

Font de la variació	Suma de quadrats	Graus de llibertat	Mitjana de quadrats
$x_1, x_2, \ldots, x_k$	*SQR*	k	SQR/k
e	*SQE*	$n-k-1$	$SQE/(n-k-1)$
y	*SQT*	$n-1$	

Taula d'anàlisi de la variància

En la primera columna es posa la **font de la variació**, és a dir, els elements del model responsables de variació.

En la segona columna posem les **sumes de quadrats** corresponents.

En la tercera columna posem els **graus de llibertat** corresponents a les sumes de quadrats.

En la quarta columna i sota el nom de **mitjana de quadrats** es posen les sumes de quadrats dividides pels graus de llibertat corresponents. Només per *SQR* i *SQE*.

És molt important tenir present el fet següent: que el model lineal expliqui significativament la variable Y no implica que totes les variables siguin explicatives; per a saber-ho, haurem de contrastar-les d'una a una tal com s'ha explicat en l'apartat anterior.

Exemple de les despeses dels ordinadors segons la seva antiguitat i les hores diàries de treball

Farem un contrast conjunt del model obtingut anteriorment per als ordinadors. Prendrem $\alpha = 0{,}05$.

1. Establim les hipòtesis nul·la i alternativa:

- Hipòtesi nul·la: H_0: $\beta_1 = \beta_2 = 0$
- Hipòtesi alternativa: H_1: al menys una $\beta_i \neq 0$, $i = 1, 2$

2. Calculem l'estadístic de contrast:

Font de la variació	Suma de quadrats	Graus de llibertat	Mitjana de quadrats
x_1, x_2	*SQR*	2	145,81/2 = 72,9
E	*SQE*	5 – 2 – 1 = 2	2,16/2 = 1,08
y	*SQT*	5 – 1 = 4	

Tenim que: $f = \dfrac{SQR/k}{SQE/(n-k-1)} = \dfrac{72{,}9}{1{,}08} = 67{,}5$.

3. Establim un criteri de decisió a partir d'un nivell de significació $\alpha = 0{,}05$. Mirant les taules de la distribució F de Snedecor, tenim que el valor crític per a $\alpha = 0{,}05$ i 2 graus de llibertat al numerador i 2 al denominador és $F_{0{,}05;2;2} = 19{,}0$.

Com que $67{,}5 > 19{,}0$, aleshores rebutgem la hipòtesi nul·la de manera que el model en conjunt és bo per a explicar la variable Y.

Amb el p-valor tenim que: $p = P(F_{2;2} > 67{,}5) = 0{,}0146 < 0{,}05$; per tant, rebutgem la hipòtesi nul·la.

Arribats a aquest punt, ens fem la pregunta següent: Com pot ser que el model en conjunt sigui bo per a explicar la variable Y, i en canvi, el contrast per separat per a cadascuna de les variables X_1 i X_2 ens ha donat que cap de les dues no era explicativa de la variable Y? A primera vista sembla que siguin resultats contradictoris. Això és degut a la presència de multicol·linealitat en el nostre problema. Ho tractarem en l'apartat següent.

7. El problema de la multicol·linealitat

En els problemes de regressió lineal múltiple esperem trobar dependència entre la variable Y i les variables explicatives X_1, X_2, ..., X_k. Però en alguns problemes de regressió podem tenir també algun tipus de dependència entre algunes de les variables X_j. En aquest cas, tenim informació redundant en el model.

Exemple de model que pot presentar multicol·linealitat

Si volem construir un model per predir el preu (Y) d'un ordinador segons la velocitat del processador (X_1), la capacitat del disc dur (X_2) i la quantitat de memòria RAM (X_3), és possible que les variables X_1 i X_3 estiguin relacionades: seria el cas en què el processador necessités un mínim de memòria RAM per a funcionar de manera òptima.

En cas que hi hagi algun tipus de dependència entre les variables, direm que hi ha **multicol·linealitat**. La multicol·linealitat pot tenir efectes molt importants en les estimacions dels coeficients de la regressió i, per tant, sobre les posteriors aplicacions del model estimat.

Variables explicatives independents

A les hipòtesis estructurals bàsiques del model de regressió lineal múltiple ja hem demanat que les variables X_1, X_2, ..., X_k siguin independents.

Com ja s'ha comentat abans, un efecte de la multicol·linealitat l'hem sofert durant aquesta sessió en el nostre exemple dels ordinadors.

Hem fet contrast sobre els paràmetres de la regressió i sobre el model conjunt i hem obtingut resultats aparentment contradictoris, però que realment no ho són.

Els contrastos individuals sobre els paràmetres indiquen que la contribució d'una variable, per exemple, l'antiguitat dels ordinadors, no té significació després d'haver descomptat l'efecte de la variable "nombre d'hores de funcionament".

D'altra banda, el contrast conjunt indica que almenys una de les dues variables contribueix a la predicció de Y (és a dir, un dels paràmetres o tots dos són

diferents de zero). De fet, és molt probable que totes dues variables hi contribueixin, però la contribució de l'una encobreix la de l'altra.

Així, doncs, ens aquests casos en què tenim variables independents molt correlacionades en un model de regressió, els resultats poden ser confusos. Habitualment, el que es fa és incloure només una d'aquestes variables en el model.

8. Resum

En aquesta darrera sessió dedicada a la regressió lineal múltiple, hem vist com hem de fer inferència sobre els coeficients de la regressió obtinguts a partir de la mostra, en particular, com hem de calcular un interval de confiança i com hem de fer un contrast d'hipòtesi per a cadascun dels coeficients obtinguts per a decidir si les variables X_j ens expliquen realment el comportament de la variable Y o en podem prescindir d'algunes. També hem vist com hem de fer un contrast conjunt del model. Finalment, hem presentat els possibles problemes de multicol·linealitat que podem tenir i que són deguts a la relació entre algunes de les variables explicatives que suposadament són independents.

Exercicis

1. Es realitza un experiment per a veure si és possible determinar el pes d'un animal després d'un període de temps determinat a partir del seu pes inicial i de la quantitat d'aliment que li és subministrat. A partir dels resultats obtinguts per a una mostra de $n = 10$:

Pes final (kg)	95	77	80	100	97	70	50	80	92	84
Pes inicial (kg)	42	33	33	45	39	36	32	41	40	38
Aliment (kg)	272	226	259	292	311	183	173	236	230	235

s'ha obtingut el model de regressió lineal:

$$\hat{y} = -22{,}984 + 1{,}395x_1 + 0{,}218x_2$$

i les sumes de quadrats següents:

$$SQR = 1.762{,}99 \qquad \text{SQE} = 256{,}30 \qquad SQT = 2.020{,}50$$

a) Podeu afirmar que les variables "pes inicial" i "quantitat d'aliment subministrat" són explicatives del "pes final" de l'animal?

b) Penseu que aquest model lineal múltiple explica significativament el pes final dels animals?

2. Considerem una mostra aleatòria de cinc famílies amb les característiques següents:

Família	**Estalvis (euros) Y**	**Ingressos (euros) X_1**	**Capital (euros) X_2**
A	600	8.000	12.000
B	1.200	11.000	6.000
C	1.000	9.000	6.000
D	700	6.000	3.000
E	300	6.000	18.000

a) Especifiqueu un model lineal múltiple per expressar l'estalvi d'acord amb els ingressos i dels capitals.

b) Estimeu els paràmetres del model de regressió lineal múltiple.

c) Podeu afirmar que les variables x_1 i x_2 són explicatives?

d) Penseu que aquest model lineal múltiple explica significativament els estalvis?

Solucionari

1.

a) Per a saber si les variables del model de regressió són explicatives, haurem de fer un contrast d'hipòtesis sobre els paràmetres obtinguts.

- **Variable X_1**:

1) Establim les hipòtesis nul·la i alternativa:

- Hipòtesi nul·la: $\beta_1 = 0$. Si aquest coeficient és nul, aleshores la variable X_1 no participaria en el model i, per tant, no seria explicativa del pes final dels animals.

- Hipòtesi alternativa: $\beta_1 \neq 0$. En aquest cas, la variable X_1 aporta informació al model; per tant, sí és explicativa del pes final.

2) Determinem un nivell significatiu $\alpha = 0{,}05$.

3) Calculem l'estadístic de contrast: $t = \dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = 2{,}3943$.

4) L'estadístic de contrast calculat és una observació d'una distribució t de Student amb $10 - 2 - 1 = 7$ graus de llibertat. Buscant en les taules trobem el valor crític corresponent:

$$t_{0{,}025;7} = 2{,}3646$$

Com que $2{,}3943 > 2{,}3646$, rebutgem H_0. La variable X_1 és significativa, encara que per molt poc.

- **Variable X_2**:

1) Establim les hipòtesis:

- Hipòtesi nul·la: $\beta_2 = 0$
- Hipòtesi alternativa: $\beta_2 \neq 0$

2) Determinem un nivell de significació: $\alpha = 0{,}05$.

3) Calculem l'estadístic de contrast: $t = \dfrac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = 3{,}7663$.

4) Com que $3{,}7663 > 2{,}3646$, rebutgem H_0. La variable X_2 (quantitat d'aliment) és significativa del pes final dels animals.

b) Farem una contrastació conjunta del model:

1) Establim les hipòtesis:

- Hipòtesi nul·la: H_0: $\beta_1 = \beta_2 = 0$
- Hipòtesi alternativa: H_1: hi ha un $\beta_j \neq 0$

2) Fixem el nivell de significació: $\alpha = 0{,}05$.

3) Calculem l'estadístic de contrast. Primer, però, construïm la taula d'anàlisi de la variància:

Font de la variació	Suma de quadrats	Graus de llibertat	Mitjana de quadrats
X_1, X_2	SQR = 1.762,99	$k = 2$	SQR/k = 881,50
e	SQE = 256,30	$n - k - 1 = 7$	SQE / ($n - k - 1$) = 36,61
Y	SQT = 2.020,50	$n - 1 = 9$	–

Estadístic de contrast: $f = \dfrac{SQR/k}{SQE/(n-k-1)} = 24{,}07$

És una observació d'una distribució F de Snedecor amb $k = 2$ i $n - k - 1 = 7$ graus de llibertat.

4) De les taules tenim un valor crític de $F_{0,05;2;7} = 4{,}74$. Com que $24{,}07 > 4{,}74$, rebutgem H_0 amb una confiança del 95%. Aleshores, el model explica significativament el pes final dels animals.

2.

a) En aquest problema tenim que el nombre d'observacions és $n = 5$ i que el nombre de variables independents és $k = 2$.

Model lineal múltiple: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

Matricialment:

$$\begin{bmatrix} 600 \\ 1.200 \\ 1.000 \\ 700 \\ 300 \end{bmatrix} = \begin{bmatrix} 1 & 8.000 & 12.000 \\ 1 & 11.000 & 6.000 \\ 1 & 9.000 & 6.000 \\ 1 & 6.000 & 3.000 \\ 1 & 6.000 & 18.000 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

b) Els valors estimats del model de regressió venen donats per:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{bmatrix} = (X^t X)^{-1} X^t Y$$

on $(X^t X)^{-1}$ és la matriu inversa de la matriu $(X^t X)$

Ara tenim:

$$X^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 8.000 & 11.000 & 9.000 & 6.000 & 6.000 \\ 12.000 & 6.000 & 6.000 & 3.000 & 18.000 \end{bmatrix}$$

$$X^t Y = \begin{bmatrix} 3.800 \\ 33.000.000 \\ 27.900.000 \end{bmatrix}$$

$$X^t X = \begin{bmatrix} 5 & 4.000 & 45.000 \\ 40.000 & 338.000.000 & 342.000.000 \\ 45.000 & 342.000.000 & 549.000.000 \end{bmatrix}$$

Atenció

Segons el nombre de xifres decimals que agafeu a partir d'aquí, els resultats poden ser una mica diferents, sense que això vulgui dir que siguin incorrectes.

$$(X^t X)^{-1} = \begin{bmatrix} 6,0492063 & -0,00057936 & -0,00013492 \\ -0,00057936 & 0,6349206 \cdot 10^{-7} & 0,79365079 \cdot 10^{-8} \\ -0,00013492 & 0,79365079 \cdot 10^{-8} & 0,79365079 \cdot 10^{-8} \end{bmatrix}$$

Ja podem calcular els paràmetres:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^t X)^{-1} X^t Y = \begin{bmatrix} 103,6507937 \\ 0,1150793651 \\ -0,02936507937 \end{bmatrix}$$

Tenim: $\hat{\beta}_0 = 103,6507937$ $\hat{\beta}_1 = 0,1150793651$ $\hat{\beta}_2 = -0,02936507937$

El model de regressió obtingut és:

$$\hat{y} = 103,6507937 + 0,1150793651 x_1 - 0,02936507937 x_2$$

c) Per a determinar si les variables són explicatives, hem de fer inferència estadística sobre els paràmetres del model.

Abans, però, hem de fer alguns càlculs més. Primer calcularem les variàncies dels paràmetres estimats. Vénen donades pels termes de la diagonal de la matriu:

$$s^2 diag(X^t X)^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_0) & & & \\ & \text{var}(\hat{\beta}_1) & & \\ & & \ldots & \\ & & & \text{var}(\hat{\beta}_k) \end{bmatrix}$$

on s^2 és la variància dels errors:

$$s^2 = \frac{SQE}{n-(k+1)} = 3.896{,}825$$

En aquests cas, tenim les variàncies i desviacions típiques dels estimadors següents:

$$s_{\hat{\beta}_0} = \sqrt{\text{var}(\hat{\beta}_0)} = 125{,}3600174$$

$$s_{\hat{\beta}_1} = \sqrt{\text{var}(\hat{\beta}_1)} = 0{,}012843081$$

$$s_{\hat{\beta}_2} = \sqrt{\text{var}(\hat{\beta}_2)} = 0{,}00454071$$

Ara ja estem en condicions de fer contrastos d'hipòtesis sobre els paràmetres del model.

- **Variable X_1**:

1) Establim les hipòtesis:

- Hipòtesi nul·la: $\beta_1 = 0$. Si el coeficient β_1 que vincula la relació entre X_1 i Y pot ser zero, vol dir que X_1 pot no tenir cap efecte sobre Y; aleshores, direm que x_1 no és una variable explicativa.

- Hipòtesi alternativa: $\beta_1 \neq 0$. En aquest cas, direm que X_1 és una variable explicativa.

2) Determinen un nivell de significació: $\alpha = 0{,}05$.

3) Calculem l'estadístic de contrast:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 8{,}96041$$

És una observació d'una distribució t de Student amb $n - k - 1 = 2$ graus de llibertat.

4) Si mirem les taules, tenim per a un valor crític: $t_{0,025;2} = 4,3027$. Com que $8,96041 > 4,3027$, rebutgem H_0. La variable X_1 (ingressos) és explicativa dels estalvis.

- **Variable X_2**: Farem el mateix per a la variable X_2 (capital).

1) Establim les hipòtesis:

- Hipòtesi nul·la: $\beta_2 = 0$
- Hipòtesi alternativa: $\beta_2 \neq 0$

2) Determinen un nivell de significació: $\alpha = 0,05$.

3) Calculem l'estadístic de contrast:

$$t = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = -6,46705$$

és una observació d'una distribució t de Student amb $n - k - 1 = 2$ graus de llibertat.

4) De les taules, teníem un valor crític: $t_{0,025;2} = 4,3027$. Com que $6,46705 > 4,3027$, rebutgem H_0. La variable X_2 (capital) també és explicativa dels estalvis.

d) Per a determinar si aquest model lineal múltiple explica significativament els estalvis de les famílies, haurem de fer una contrastació conjunta del model.

1) Establim les hipòtesis nul·la i alternativa:

- Hipòtesi nul·la: H_0: $\beta_1 = \beta_2 = 0$
- Hipòtesi alternativa: H_1: hi ha al menys un $\beta_j \neq 0$

2) Determinen un nivell significatiu, per exemple $\alpha = 0,05$.

3) Calcularem l'estadístic de contrast. Abans, però, haurem de calcular les sumes de quadrats i construir la taula de l'anàlisi de la variància. Per a calcular la suma de quadrats de la regressió (*SQR*) ens fa falta conèixer:

– la mitjana de les $y_i = \bar{y} = 760,0$.

– I els valor estimats de y_i y :

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \dots\\ \hat{y}_n\end{bmatrix} = \begin{bmatrix}1 & x_{11} & x_{21} & \dots & x_{k1}\\ 1 & x_{12} & x_{22} & \dots & x_{k2}\\ \dots & \dots & \dots & \dots & \dots\\ 1 & x_{1n} & x_{2n} & \dots & x_{kn}\end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \dots\\ \hat{\beta}_k\end{bmatrix}, \text{ ara tenim: } \begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\end{bmatrix} = \begin{bmatrix}671{,}9047619\\ 1.193{,}33333\\ 963{,}1746032\\ 706{,}0317460\\ 265{,}5555556\end{bmatrix}$$

Per a *SQE*, abans hem de calcular el vector dels errors:

$$e = \begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\end{bmatrix} - \begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\\ \hat{y}_4\\ \hat{y}_5\end{bmatrix} = \begin{bmatrix}-71{,}90476190\\ 6{,}666666667\\ 36{,}82539683\\ -6{,}031746032\\ 34{,}44444444\end{bmatrix}$$

Així, doncs, les sumes de quadrats són:

$$SQT = \sum(y_i - \bar{y})^2 = 492.000$$

$$SQR = \sum(\hat{y}_i - y)^2 = 484.206{,}34$$

$$SQE = \sum(y_i - \hat{y})^2 = \sum e_i^2 = 7.793{,}65$$

Podem construir la taula d'anàlisi de la variància:

Font de la variació	Suma de quadrats	Graus de llibertat	Mitjana de quadrats
X_1, X_2	SQR = 484.206,34	k = 2	SQR/k = 242.103,17
e	SQE = 7.793,65	$n - k - 1$ = 2	$SQE/(n - k - 1)$ = 3.896,825
Y	SQT = 492.000	$n - 1$ = 4	–

Estadístic de contrast: $f = \dfrac{SQR/k}{SQE/(n-k-1)} = 62{,}12.$

És una observació d'una distribució F de Snedecor amb $k = 2$i$n - k - 1 = 2$ graus de llibertat.

4) De les taules tenim un valor crític de $F_{0,05;2;2} = 1{,}90$. Com que $62{,}12 > 19{,}0$, rebutgem H_0. Així, doncs, aquest model de regressió múltiple explica significativament els estalvis de les famílies a partir dels ingressos i del capital.

Annex 3.1

Valor esperat de $\hat{\beta}_j$:

Per a cercar els valors esperats de $\hat{\beta}_j$, farem servir la notació matricial que ja vam introduir en el mòdul anterior i que ens permetrà certa comoditat a l'hora d'escriure totes les equacions. A partir de l'equació matricial que ens permetia de trobar els estimadors dels coeficients de la regressió:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Per a simplificar encara més els càlculs, anomenarem $C = (X^t X)^{-1} X^t$ i així podrem escriure la darrera equació de la forma: $\hat{\beta} = CY$. D'altra banda, el model de regressió lineal múltiple:

Obseveu que...

$CX((X^t X)^{-1} X^t X) = I$
és la matriu identitat.

$$y_i = \beta_0 = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i$$

escrit matricialment: $Y = X\beta + e$. De manera que:

$$\hat{\beta} = CY = C(X\beta + e) = CX\beta + Ce = \beta + Ce$$

Si calculem ara el valor esperat:

$$E(\hat{\beta}) = E(\beta) + E(Ce) = \beta + CE(e) = \beta$$

on hem considerat que $E(e) = 0$, tal com vam suposar en la sessió anterior en les hipòtesis estructurals bàsiques del model de regressió lineal múltiple.

Linealitat

Hem fet servir la propietat de linealitat de l'esperança matemàtica:
$E(aX) = aE(X)$

En resum, hem obtingut que: $E(\hat{\beta}) = \beta$, és a dir:

$$E(\hat{\beta}_1) = \beta_1$$
$$E(\hat{\beta}_2) = \beta_2$$
$$.....$$
$$E(\hat{\beta}_k) = \beta_k$$

Annex 3.2

Variància de $\hat{\beta}_j$:

Per a una $\hat{\beta}_j$, la seva variància vindrà donada com sempre per:

Aquí ja hem utilitzat el resultat anterior:
$E(\hat{\beta}_j) = \beta_j$

$$Var(\hat{\beta}_j) = E[(\hat{\beta}_j - \beta_j)^2]$$

Per a calcular aquesta variància, farem servir una vegada més la notació i el càlcul matricial.

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t] = E\left[\begin{bmatrix}(\hat{\beta}_0-\beta_0)\\(\hat{\beta}_1-\beta_1)\\\dots\\(\hat{\beta}_k-\beta_k)\end{bmatrix}((\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\dots(\hat{\beta}_k-\beta_k))\right] =$$

$$= \begin{bmatrix} E[(\hat{\beta}_0-\beta_0)^2] & E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)] & \dots & E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_k-\beta_k)] \\ E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)] & E[(\hat{\beta}_1-\beta_1)^2] & \dots & E[(\hat{\beta}_1-\beta_1)(\hat{\beta}_k-\beta_k)] \\ \dots & \dots & \dots & \dots \\ E[(\hat{\beta}_0-\beta_0)(\hat{\beta}_k-\beta_k)] & E[(\hat{\beta}_1-\beta_1)(\hat{\beta}_k-\beta_k)] & \dots & E[(\hat{\beta}_k-\beta_k)^2] \end{bmatrix}$$

La matriu anterior rep el nom de **matriu de variàncies-covariàncies**, ja que els seus elements de la diagonal són les variàncies de les $\hat{\beta}_j$ i els elements de fora de la diagonal són les covariàncies dels parells de variables $\hat{\beta}_j$ i $\hat{\beta}_m$. A nosaltres ens interessen les variàncies de les $\hat{\beta}_j$, és a dir, els valors esperats dels elements de la diagonal de la matriu:

L'esperança d'una matriu

Hem fet servir que la esperança d'una matriu és la matriu de les esperances dels seus elements.

$$(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t$$

D'altra banda, hem vist abans que $\hat{\beta} = \beta + Ce$, de manera que podem escriure: $\hat{\beta}-\beta = Ce$ i, per tant:

$$(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t = (Ce)(Ce)^t$$

Combinant aquests resultats, tenim que les variàncies de les $\hat{\beta}_j$ són els valors esperats dels elements de la diagonal de la matriu $(Ce)(Ce)^t$, és a dir:

Producte de matrius

Recordem la important propietat del producte de matrius:
$(AB)^t = B^tA^t$

$$E[Cee^tC^t] = CE[ee^t]C^t = \sigma^2CC^t = \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}$$

on hem tingut en compte que $E[ee^t] = \sigma^2 I$ per a les hipòtesis estructurals bàsiques del model de regressió lineal múltiple que vam suposar en la sessió anterior.

Finalment, tenim que les variàncies de les $\hat{\beta}_j$ són els elements de la diagonal de la matriu: $\sigma^2(X^tX)^{-1}$, és a dir:

$$\begin{bmatrix}\text{var}\,(\hat{\beta}_0) & & & \\ & \text{var}\,(\hat{\beta}_1) & & \\ & & \dots & \\ & & & \text{var}\,(\hat{\beta}_k)\end{bmatrix} = \sigma^2 diag(X^tX)^{-1}$$