Relació entre variables: causalitat, correlació i regressió

Correlació entre variables. Models de regressió simple (lineal, quadràtica, cúbica). Models de regressió múltiple

Blanca de la Fuente i Patricia Carracedo

PID_00242446

Temps de lectura i comprensió: **5 hores**

UOC
Universitat Oberta de Catalunya

Índex

Introducció

En aquest mòdul, estudiarem les relacions que es poden presentar entre diferents variables. En concret estudiaran possibles relacions de dependència entre les variables intentant trobar una expressió que ens permetrà estimar una variable en funció d'altres. Per aprofundir en l'anàlisi és necessari determinar la *forma* concreta com es relacionen i mesurar el seu *grau* d'associació.

Així, per exemple, podem aplicar l'estudi de les relacions entre variables per a donar respostes a preguntes i a casos com ara:

- Hi ha relació entre l'edat dels lectors i el nombre de préstecs de llibres?

- En un altre cas, una editorial podria usar la relació entre el nombre de pàgines d'un treball i el temps d'impressió, per a predir el temps emprat en la impressió.

- Es vol estudiar el temps de resposta d'uns certs programes de recerca bibliogràfica en funció del nombre d'instruccions en què estan programats.

- En una determinada empresa de venda de llibres en línia, com representem que l'augment de la quantitat gastada en publicitat provoca un increment de les vendes?

Aquest mòdul examina la relació entre dues variables, la independent i la dependent, per mitjà de la regressió simple i la correlació. També es considera el model de regressió múltiple en què apareixen dues variables independents o més.

Objectius

Els objectius acadèmics d'aquest mòdul es descriuen a continuació:

1. Comprendre la relació entre correlació i regressió simple.

2. Usar gràfics per a ajudar a comprendre una relació de regressió.

3. Ajustar una recta de regressió i interpretar els coeficients.

4. Obtenir i interpretar les correlacions i la seva significació estadística.

5. Utilitzar els residus d'una regressió per a comprovar la validesa de les suposicions necessàries per a la inferència estadística.

6. Aplicar contrastos d'hipòtesi.

7. Ajustar una equació de regressió múltiple i interpretar-ne els resultats.

1. Relació entre variables

Quan estudiem conjuntament dues variables o més que no són independents, la relació entre elles pot ser **funcional** (per exemple, relació matemàtica exacta entre dues variables, espai recorregut per un vehicle que circula a velocitat constant i el temps emprat a recórrer-lo) o **estadística** (per exemple, no hi ha una expressió matemàtica exacta que relacioni ambdues variables, hi ha una relació aproximada entre les dues variables, increment de les vendes de llibres en funció de la quantitat gastada en publicitat). En aquest últim cas ens interessa estudiar el grau de dependència existent entre ambdues variables. Ho farem mitjançant **l'anàlisi de correlació** i, finalment, desenvoluparem un model matemàtic per a estimar el valor d'una variable basant-nos en el valor d'una altra, de la qual cosa en direm **anàlisi de regressió**.

L'anàlisi de regressió no **es pot** interpretar com un procediment per a establir una relació **causa-efecte o causalitat** entre variables. La regressió només pot indicar com estan **associades** les variables entre elles i ens permet construir un model per a explicar-ne la relació. La correlació indica el grau de la relació entre dues variables sense suposar que una alteració en l'una causi un canvi en l'altra variable.

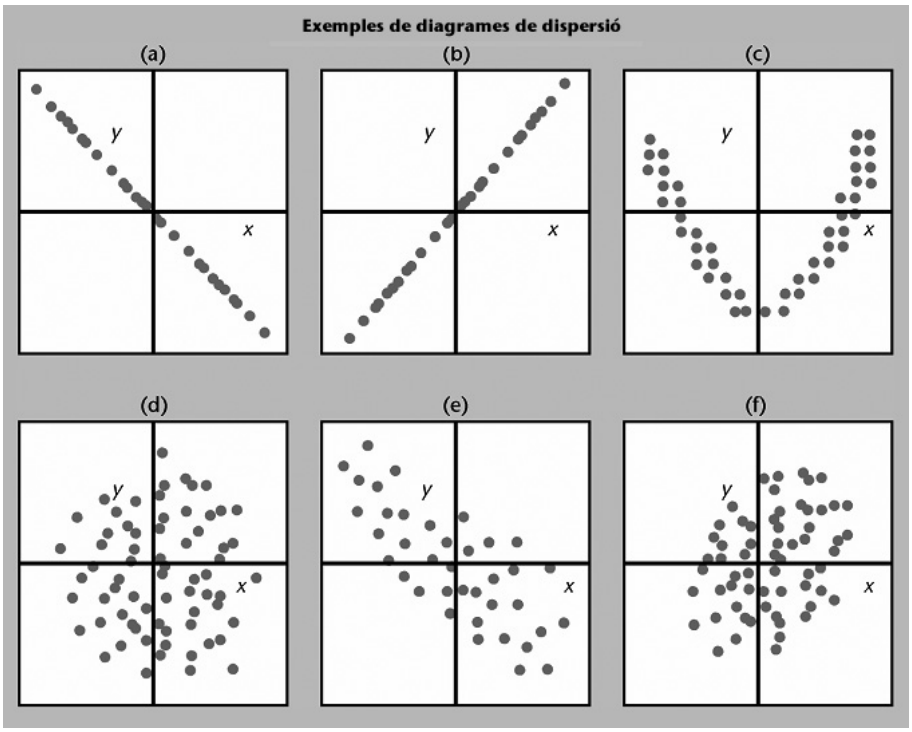
L'objectiu principal de l'anàlisi de regressió és explicar el comportament d'una **variable dependent *Y*** (endògena o explicada) a partir d'una **variable independent** o diverses (exògenes o explicatives). El tipus més senzill de regressió és la **regressió simple**. La regressió lineal simple estima una equació lineal que descriu la relació, mentre que la correlació mesura la força de la relació lineal. A part dels models lineals, podem establir altres models de regressió no lineals. L'anàlisi de regressió on intervenen dues variables independents o més en diem anàlisi de regressió múltiple, en la qual una variable és explicada per l'acció simultània d'altres variables.

Diagrama de dispersió

Abans d'abordar el problema, podem intuir si hi ha relació entre les variables a través de la representació gràfica anomenada **diagrama de dispersió** o **núvol de punts**.

A partir d'un conjunt d'observacions (x_i, y_i) de dues variables X i Y sobre una mostra d'individus es representen aquestes dades sobre un eix de coordenades $x - y$. En la figura 1 s'inclouen diverses gràfiques de dispersió que il□lustren alguns tipus de relació entre variables.

Figura 1. Diagrames de dispersió



En els casos (a) i (b) tenim que les observacions estan sobre una recta. En el primer cas, amb pendent negatiu, indica una relació inversa entre les variables (a mesura que X augmenta, la Y és cada vegada menor) i el contrari en el segon cas, en què el pendent és positiu indica una relació directa entre les variables (a mesura que augmenta X, la Y també augmenta). En aquests dos casos els punts s'ajusten perfectament sobre la recta, de manera que tenim una relació funcional entre totes dues variables donada per l'equació de la recta.

En el cas (c) els punts estan situats en una franja bastant estreta que té una forma ben determinada. No serà una relació funcional, ja que els punts no se situen sobre una corba, però sí que és possible assegurar l'existència d'una forta relació entre totes dues variables. De tota manera, veiem que no es tracta d'una relació lineal (el núvol de punts té forma de paràbola).

En el cas (d) no tenim cap tipus de relació entre les variables. El núvol de punts no presenta una forma ben determinada; els punts estan absolutament dispersos.

En els casos (e) i (f) podem observar que sí que hi ha algun tipus de relació entre les dues variables. En el cas (e) podem veure un tipus de dependència lineal amb pendent negatiu, ja que a mesura que el valor de X augmenta, el valor de Y disminueix. Els punts no estan sobre una línia recta, però s'apropen bastant, de manera que podem pensar en una relació lineal. En el cas (f) observem una relació lineal amb pendent positiu, però no tan forta com l'anterior.

Després d'estudiar el diagrama de dispersió, el pas següent és comprovar analíticament la dependència o independència d'ambdues variables.

2. Anàlisi de la correlació

L'anàlisi de correlació mesura el grau de relació entre les variables. En aquest apartat veurem l'anàlisi de correlació simple, que mesura la relació entre només una variable independent (X) i la variable dependent (Y). En l'apartat 4 d'aquest mòdul es descriu l'anàlisi de correlació múltiple que mostra el grau d'associació entre dues variables independents o més i la variable dependent.

La correlació simple determina la quantitat de variació conjunta que presenten dues variables aleatòries d'una distribució bidimensional. En concret, quantifica la dependència lineal, i en aquest cas hi haurà correlació lineal. El coeficient de correlació lineal es diu coeficient de correlació de Pearson designat ***r***, el valor del qual oscil□la entre –1 i +1. La seva expressió és el quocient entre la covariància mostral entre les variables i el producte de les seves respectives desviacions típiques:

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

El valor de r s'aproxima a +1 quan la correlació tendeix a ser lineal directa (majors valors de X signifiquen majors valors de Y), i s'aproxima a –1 quan la correlació tendeix a ser lineal inversa. Podem formular la pregunta: a partir de quin valor de r podem dir que la relació entre les variables és forta? Una regla raonable és dir que la relació és feble si $0 \leq |r| \leq 0{,}5$; forta si $0{,}8 \leq |r| \leq 1$, i moderada si té un altre valor.

Donada una variable X amb $x_1, x_2, \ldots, x_n$ valors mostrals i una altra variable Y amb $y_1, y_2, \ldots, y_n$ valors mostrals, essent n el nombre total d'observacions i essent la mitjana de X: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ i la mitjana de Y: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

La covariància mostral entre dues variables X i Y ens permet mesurar aquestes relacions positives i negatives entre les variables X i Y:

$$Cov(X,Y) = S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})$$

La covariància mostral podem calcular-la mitjançant una altra expressió equivalent:

$$S_{XY} = \frac{\left[\sum_{i,j=1}^{n} x_i y_j\right] - n \cdot \bar{x} \cdot \bar{y}}{n-1}$$

Exemple 1. Estudi dels serveis oferits per un centre de documentació.

Estem realitzant un procés d'avaluació dels serveis oferits per un centre de documentació. Per conèixer l'opinió dels usuaris els hem demanat que emplenin un qüestionari d'avaluació del servei. Fem dues preguntes, una perquè valorin de 0 a 10 la impressió que tenen sobre el funcionament global del centre i una altra pregunta que valora específicament l'atenció als usuaris. Volem conèixer si hi ha relació entre ambdues valoracions, per determinar si les valoracions pel que fa a l'atenció a l'usuari (representades per la variable dependent *Y*) estan relacionades amb les valoracions obtingudes respecte al funcionament global del centre (variable independent *X*).

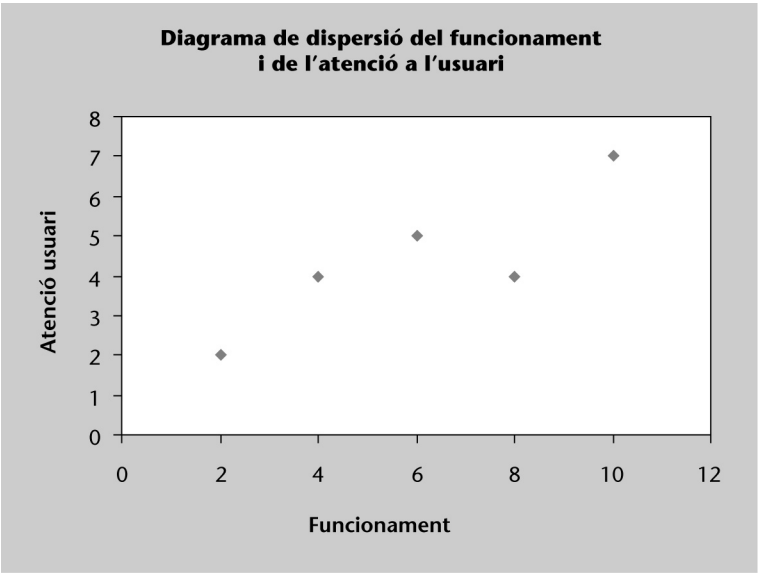
Per a això, un investigador ha seleccionat a l'atzar cinc persones entrevistades i donen les valoracions següents:

Taula 1. Dades obtingudes de respostes a 5 entrevistes realitzades sobre valoracions de funcionament i atenció a usuaris d'un centre de documentació

Entrevista (*i*)	Funcionament (*X*)	Atenció (*Y*)
1	2	2
2	4	4
3	6	5
4	8	4
5	10	7

El diagrama de dispersió (figura 2) ens permet observar gràficament les dades i treuren conclusions. Sembla que les valoracions d'atenció a l'usuari són millors per a valoracions elevades del funcionament global del centre. A més, per a aquestes dades la relació entre l'atenció a l'usuari i el funcionament sembla poder aproximar-se a una línia recta; realment sembla haver-hi una relació lineal positiva entre *X* i *Y*.

Figura 2. Diagrama de dispersió del funcionament del centre i de l'atenció a l'usuari



Per a determinar si hi ha correlació lineal entre les dues variables, calculem el coeficient de correlació ***r***.

En la taula 2 fem els càlculs necessaris per a determinar els valors de les variàncies, les desviacions típiques mostrals i la covariància mostral.

Taula 2. Càlcul de les sumes de quadrats per a l'equació estimada de regressió de mínims quadrats

Funcionament (*X*)	Atenció (*Y*)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
2	2	–4	–2,4	9,6	16	5,76
4	4	–2	–0,4	0,8	4	0,16
6	5	0	0,6	0	0	0,36
8	4	2	–0,4	–0,8	4	0,16
10	7	4	2,6	10,4	16	6,76

y_i representa les valoracions observades (reals) del funcionament global obtingudes en l'entrevista i,

$$n = 5 \quad \sum_{i=1}^{5} x_i = 30 \quad \sum_{i=1}^{5} y_i = 4,4 \quad \sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = 20 \quad \sum_{i=1}^{5} (x_i - \bar{x})^2 = 40 \quad \sum_{i=1}^{5} (y_i - \bar{y})^2 = 13,2$$

fent les operacions següents obtindrem el coeficient de correlació lineal.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{30}{5} = 6 \; ; S_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{40}{5-1}} = 3,16$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n} = \frac{22}{5} = 4,4 \; ; S_Y = \sqrt{\frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n-1}} = \sqrt{\frac{13,2}{5-1}} = 1,82$$

$$Cov(X,Y) = \frac{1}{n-1} \sum_{i,j=1}^{n} (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{5-1} \cdot 20 = 5$$

El coeficient de correlació lineal és:

$$r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{5}{3,16 \cdot 1,82} = 0,87$$

Com que el valor del coeficient de correlació lineal és pròxim a 1, es pot afirmar que hi ha una correlació lineal positiva entre les valoracions obtingudes d'atenció a l'usuari i les valoracions del funcionament global del centre. És a dir, el funcionament global està associat positivament a l'atenció a l'usuari.

3. Models de regressió simple

3.1. Models de regressió lineal simple

Una vegada que hem obtingut el diagrama de dispersió i després d'observar una possible relació lineal entre les dues variables, el pas següent seria trobar l'equació de la recta que millor s'ajusti al núvol de punts. Aquesta recta es denomina **recta de regressió.** Una recta queda ben determinada si el valor del seu pendent (b) i de l'ordenada en l'origen (a) són conegudes. D'aquesta manera l'equació de la recta és donada per:

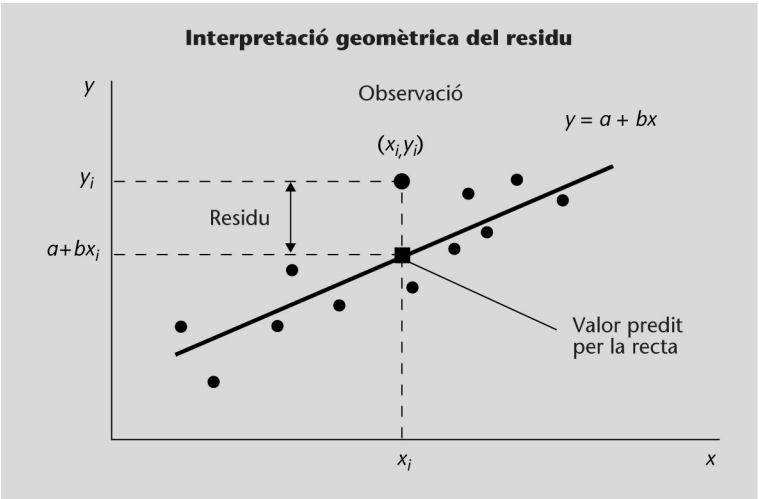
$$Y = a + bx$$

A partir de la fórmula anterior definim per a cada observació (x_i, y_i) l'*error* o *residu* com la distància vertical entre el punt (x_i, y_i) i la recta, és a dir:

$$y_i - (a + bx_i)$$

Per cada recta considerada, hi ha una col·lecció diferent de residus. Cal buscar la recta que minimitzi la suma dels quadrats dels residus. Aquest és el **mètode dels mínims quadrats**, un procediment per a trobar l'equació de regressió que consisteix a buscar els valors dels coeficients a i b de manera que la suma dels quadrats dels residus sigui mínima, i així obtenim la **recta de regressió per mínims quadrats** (figura 3).

Figura 3. Recta de regressió per mínims quadrats



Nota

La recta de regressió passa pel punt $(\overline{x}, \overline{y})$.

Hem fet un canvi en la notació per a distingir de manera clara entre una recta qualsevol: $y = a + bx$ i la recta de regressió per mínims quadrats obtinguda en determinar a i b.

A partir d'ara, escriurem la **recta de regressió** de la manera següent:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

El model de regressió lineal permet trobar el valor esperat de la variable aleatòria Y quan X pren un valor específic.

La **recta de regressió *Y*/*X*** permet predir un valor de y per a un determinat valor de x.

Per a cada observació (x_i,y_i) definim:

- el valor estimat o predit per a la recta de regressió:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- els paràmetres o coeficients de la recta y són donats per:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{i} \quad \hat{\beta}_1 = \frac{Cov(XY)}{S_X^2} = \frac{S_{XY}}{S_X^2}$$

On:

S_{XY} és la covariància mostral, S_X^2 la variància mostral de X, $\bar{x}$ e $\bar{y}$ són les mitjanes aritmètiques de les variables X i Y respectivament.

$\hat{\beta}_0$ és l'ordenada en l'origen de l'equació estimada de regressió.

$\hat{\beta}_1$ és el pendent de l'equació estimada de regressió.

- el residu o error com la diferència entre el valor observat y_i i el valor estimat $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Exemple 1. Estudi dels serveis oferits per un centre de documentació.

Hem comprovat en l'exemple anterior que hi ha correlació lineal entre ambdues variables. Ara calcularem la **recta de regressió per mínims quadrats *Y*/*X***.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

en la qual,

x_i = valor de funcionament per a la i-èsima entrevista

$\hat{\beta}_0$ = ordenada en l'origen de la línia estimada de regressió

$\hat{\beta}_1$ = pendent de la línia estimada de regressió

$\hat{y}_i$ = valor estimat de l'atenció a l'usuari per a la *i*-èsima entrevista

Perquè la línia estimada de regressió s'ajusti bé amb les dades, les diferències entre els valors observats i els valors estimats d'atenció a l'usuari han de ser petits.

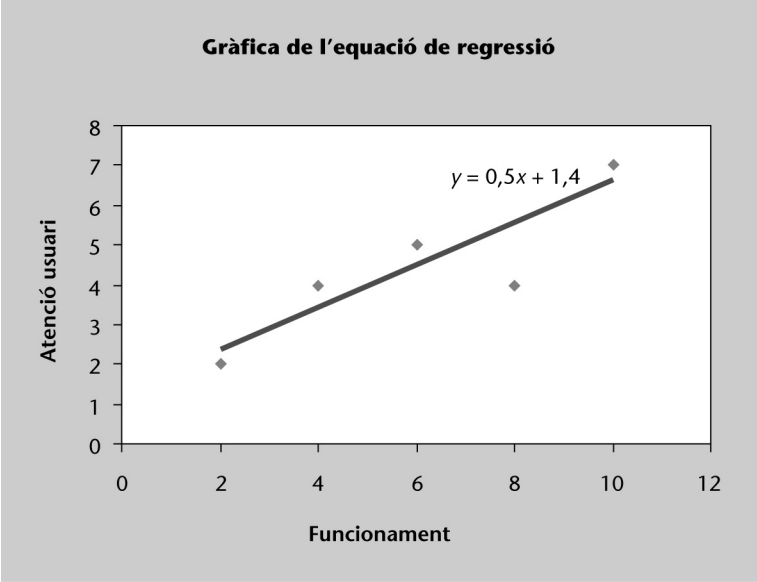
Utilitzant els valors obtinguts en la taula 2 podem determinar el pendent i l'ordenada en l'origen de l'equació estimada de regressió en aquest exemple. Els càlculs són els següents:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2} = 0,5 \; ; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1,4$$

Pel que s'ha dit abans, l'equació estimada de regressió deduïda amb el mètode de mínims quadrats, serà:

$$\hat{y} = 1,4 + 0,5x$$

Figura 4. Gràfica de l'equació de regressió exemple 1



Interpretació dels paràmetres de la recta de regressió

És important interpretar els coeficients de l'equació en el context del fenomen que estem estudiant.

- Interpretació de l'ordenada en l'origen, $\hat{\beta}_0$:

  Aquest coeficient representa l'estimació del valor de Y quan X és igual a zero. No sempre té una interpretació pràctica. Perquè sigui possible, és necessari que:

  – realment sigui possible que X prengui el valor $x = 0$,
  – es tinguin prou observacions properes al valor $x = 0$.

- Interpretació del pendent de la recta, $\hat{\beta}_1$:

  Aquest coeficient representa l'estimació de l'increment que experimenta la variable Y quan X augmenta en una unitat. Aquest coeficient ens informa de com estan relacionades les dues variables en quina quantitat varien els valors de Y quan varien els valors de la X en una unitat.

La qualitat o bondat de l'ajustament

Una vegada calculada la recta de regressió per mínims quadrats hem d'analitzar si aquest ajust al model és prou bo. Si mirem si en el diagrama de dispersió els punts experimentals queden molt a prop de la recta de regressió obtinguda, podem tenir una idea de si la recta s'ajusta o no a les dades, però ens fa falta un valor numèric que ens ajudi a necessitar-ho. La mesura de bondat d'ajust per a una equació de regressió és el **coeficient de determinació R^2.** Ens indica el grau d'ajust de la recta de regressió als valors de la mostra, i es defineix com la proporció de variància a Y explicada per la recta de regressió. L'expressió de R^2 és la següent:

$$R^2 = \frac{\text{Variància a } Y \text{ explicada per la recta de regressió}}{\text{Variància total de les dades } Y}$$

La variància explicada per la recta de regressió és la variància dels valors estimats. La variància total de les dades és la variància dels valors observats. Per tant, podem establir que:

Variància total de Y = variància explicada per la regressió +
+ variància no explicada (residual o dels errors)

És a dir, podem descompondre la variabilitat total (*SSTotal*) de les observacions de la manera següent:

$$SSTotal = SSRegressió + SSError$$

en la qual,

SSTotal, és la suma de quadrats totals $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$

SSRegressió, mesura quant es desvien els valors de $\hat{y}_i$ mesurats en la línia de regressió, dels valors de $\overline{y}_i$, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$

SSError, representa l'error que es comet en usar $\hat{y}_i$ per a estimar y_i, és la suma de quadrats d'aquests errors, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

Ara veiem com es poden utilitzar les tres sumes de quadrats, *SST, SSR* i *SSE* per a obtenir la mesura de bondat d'ajust per a l'equació de regressió, que és el coeficient de determinació R^2. Serà donat per l'expressió:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Els valors del coeficient de determinació estan compresos entre zero i u: $0 \leq R^2 \leq 1$
- R^2 =1 quan l'ajust és perfecte, és a dir, tots els punts són sobre la recta de regressió.
- R^2 = 0 mostra la inexistència de relació entre les variables *X* i *Y*.
- Com R^2 explica la proporció de variabilitat de les dades explicada pel model de regressió, com més pròxim a la unitat, millor serà l'ajust.

Relació entre R^2 i *r*

És molt important tenir clara la diferència entre el coeficient de correlació i el coeficient de determinació:

- R^2 mesura la proporció de variació de la variable dependent explicada per la variable independent.
- r^2 és el coeficient de correlació, mesura el grau d'associació lineal entre les dues variables.
- No obstant això, en la regressió lineal simple tenim que $R^2 = r^2$.

Observacions

Un coeficient de determinació diferent de zero no significa que hi hagi relació lineal entre les variables. Per exemple, R^2 = 0,5 només diu que el 50% de la variància de les observacions queda explicat pel model lineal.

La relació entre R^2 i *r* ajuda a comprendre el que s'ha exposat en l'anàlisi de la correlació, que un valor de r^2= 0,5 indica una correlació dèbil. Aquest valor re-

presentarà un $R^2 = 0,25$, és a dir, el model de regressió només explica un 25% de la variabilitat total de les observacions.

El signe de *r* dóna informació de si la relació és positiva o negativa. Així, doncs, amb el valor de *r* sempre es pot calcular el valor de R^2, però al revés quedarà indeterminat el valor del signe llevat que coneguem el pendent de la recta. Per exemple, donat un $R^2 = 0,81$, si se sap que el pendent de la recta de regressió és negatiu, aleshores es pot afirmar que el coeficient de correlació *r* serà igual a 0,9.

Predicció

La predicció constitueix una de les aplicacions més interessants de la tècnica de regressió. La predicció consisteix a determinar a partir del model estimat el valor que pren la variable endògena per a un valor determinat de l'exògena. La fiabilitat d'aquesta predicció serà tant més gran, en principi, com millor sigui l'ajust (és a dir, com més gran sigui R^2), en el supòsit que hi hagi relació causal entre la variable endògena i la variable exògena.

Nota

Variable endògena és la variable dependent que es prediu o s'explica i és representada per *Y*.

Variable exògena és la variable independent que serveix per a predir o explicar i és representada per *X*.

Exemple 1. Estudi dels serveis oferits per un centre de documentació

Una vegada obtinguda l'equació estimada de regressió $\hat{y} = 1,4 + 0,5x$ de l'exemple anterior, interpretem els resultats:

En aquest cas l'ordenada en l'origen ($\hat{\beta}_0 = 1,4$) sí que pot tenir interpretació amb sentit, ja que correspondria a l'estimació de la puntuació obtinguda per a l'atenció a l'usuari quan la puntuació del funcionament global és zero. El pendent ($\hat{\beta}_1 = 0,5$) és positiu, la qual cosa indica que l'augment en una unitat de la valoració del funcionament global del centre està associat amb un augment de 0,5 unitats en la puntuació d'atenció a l'usuari.

Si volguéssim predir la valoració de l'atenció per a una persona que ha valorat en 7 el funcionament global, el resultat seria:

$$\hat{y} = 1,4 + 0,5 \cdot 7 = 4,9$$

En l'exemple hem obtingut l'equació de regressió i hem d'analitzar la bondat d'aquest ajust que donaria resposta a la pregunta següent: les dades s'ajusten bé a aquesta equació de regressió?

Calcularem el coeficient de determinació, que és una mesura de la bondat d'ajust. Per a això hem de descompondre la variabilitat total de les observacions de la manera següent:

$$SST = SSR + SSE$$

Utilitzant els valors de la taula 2 ("Càlcul de les sumes de quadrats per a l'equació estimada de regressió amb mínims quadrats"), calculem *SST* = suma de quadrats total, és la suma de l'última columna de la taula 2.

$$SST = \sum_{i=1}^{5}(y_i - \bar{y})^2 = 13,2$$

En la taula 3 veiem els càlculs necessaris per a determinar l'*SSE* = suma de quadrats deguda a l'error

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = 3,2$$

Taula 3. Càlcul de les sumes de quadrats degudes a l'error SCE

Funcionament (*X*)	**Atenció (*Y*)**	$\hat{y} = 1,4 + 0,5x_i$	$e = y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
2	2	2,4	–0,4	0,16
4	4	3,4	0,6	0,36
6	5	4,4	0,6	0,36
8	4	5,4	–1,4	1,96
10	7	6,4	0,6	0,36

$$SSE = \sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = 3,2$$

L'*SSR* = suma de quadrats deguda a la regressió es pot calcular amb facilitat utilitzant aquesta expressió:

$$SSR = \sum_{i=1}^{5}(\hat{y}_i - \bar{y})^2$$

o bé si es coneixen *SST* i *SSE* es pot obtenir fàcilment, de la manera següent:

$$SSR = SST - SSE = 13,2 - 3,2 = 10$$

El valor del coeficient de determinació serà:

$$R^2 = \frac{SSR}{SST} = \frac{10}{13,2} = 0,7576$$

Si l'expressem en percentatge, R^2= 75,76%. Podem concloure que el 75,76% de la variació de la puntuació en l'atenció a l'usuari es pot explicar amb la relació lineal entre les valoracions del funcionament global del centre i l'atenció a l'usuari. L'ajust al model lineal és bo. Es considera un bon ajust quan R^2 és més gran o igual que 0,5.

El coeficient de correlació lineal *r* serà $\sqrt{0,75760} = |0,87|$, resultat d'acord amb l'estimació obtinguda utilitzant la covariància.

Solució de problemes de regressió lineal simple amb programes informàtics

Per a resoldre l'exercici emprem el programari R Commander.

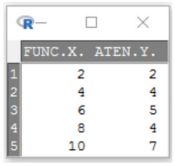
Escrivim les dades de l'exemple 1. "Estudi dels serveis oferts per un centre de documentació". A la variable independent (*Y*) l'anomenem **ATEN** (d'atenció a l'usuari) i a la variable dependent (*X*) l'anomenem **FUNC** (de funcionament global) per a facilitar la interpretació dels resultats. Escrivim les dades **FUNC** i les dades d'**ATEN**, amb encapçalaments, per a obtenir el diagrama de dispersió.

Passos a seguir

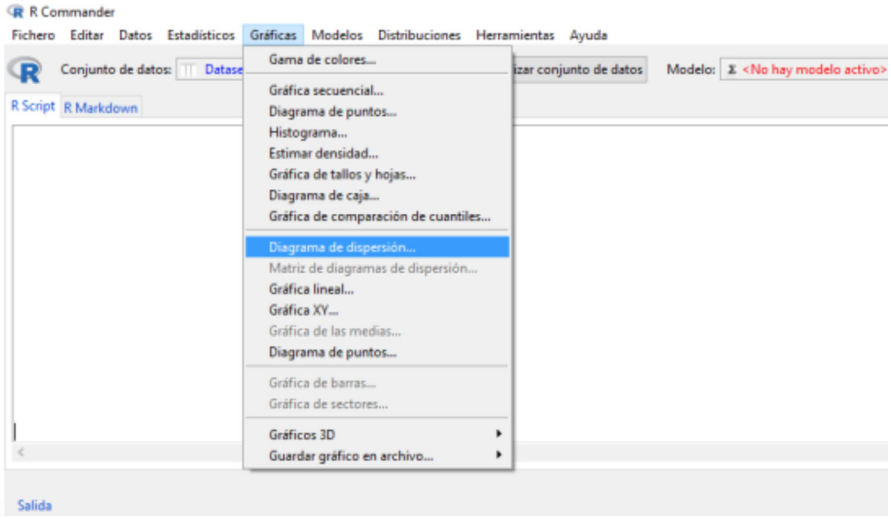
Per a fer el gràfic una vegada introduïdes les dades en el programa **(1)**, se segueix la ruta *Gráficos* > *Diagrama de dispersión* **(2)** i s'emplenen els camps a la finestra corresponent seleccionant les variables **(3)**. Seleccioneu *Aceptar* per obtenir el diagrama de dispersió.

Figura 5. Passos a seguir per a obtenir el diagrama de dispersió

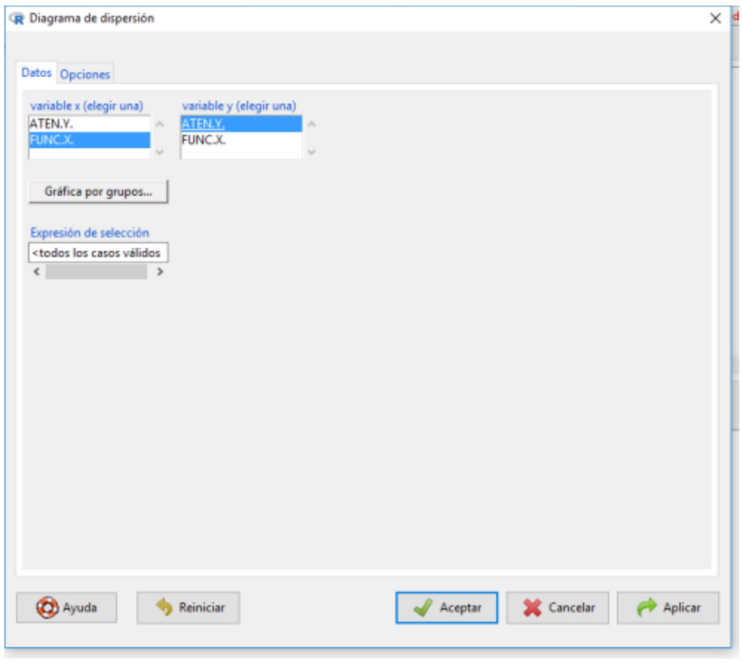
1



2

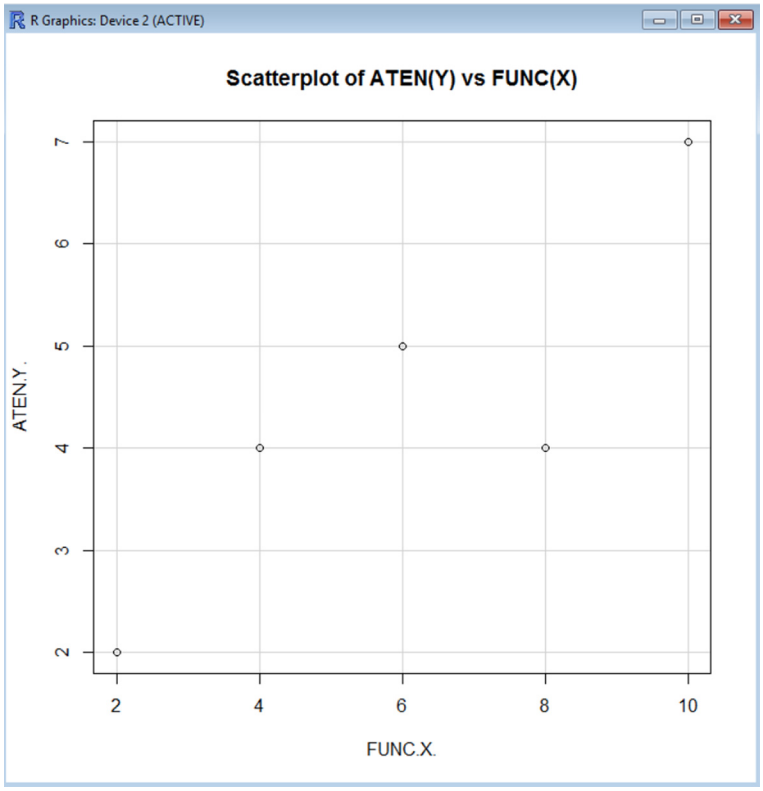


3



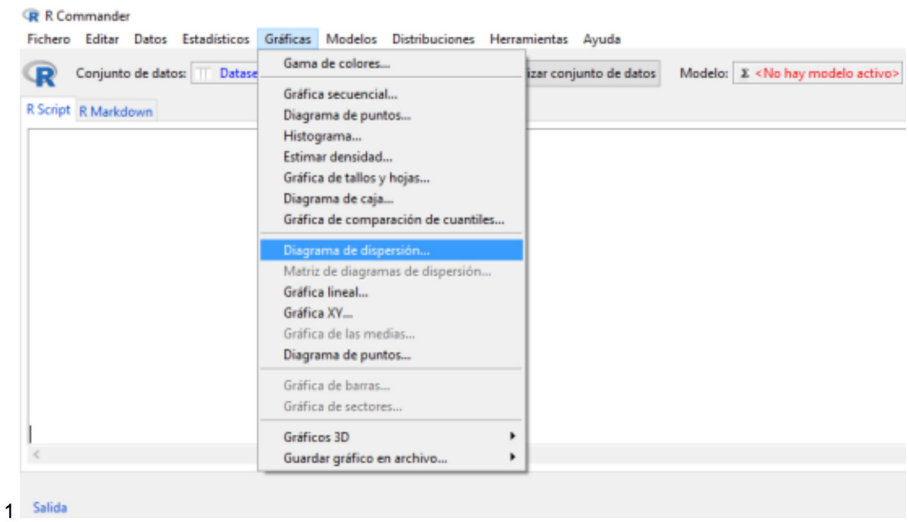
Vam obtenir el diagrama de la figura 6.

Figura 6. Diagrama de dispersió. R Commander



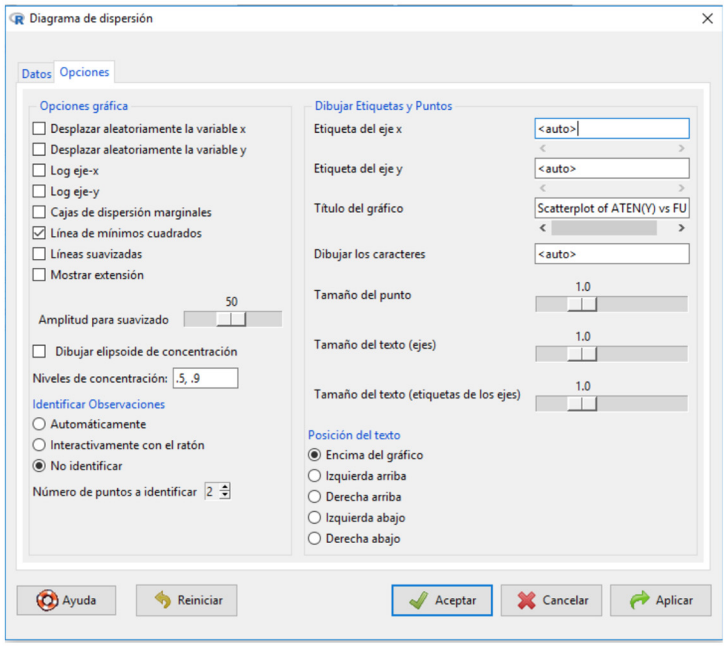
La figura 7 mostra els passos a seguir per a representar la recta de regressió de mínims quadrats:

Figura 7. Passos a seguir per a representar la recta de regressió de mínims quadrats



Passos a seguir

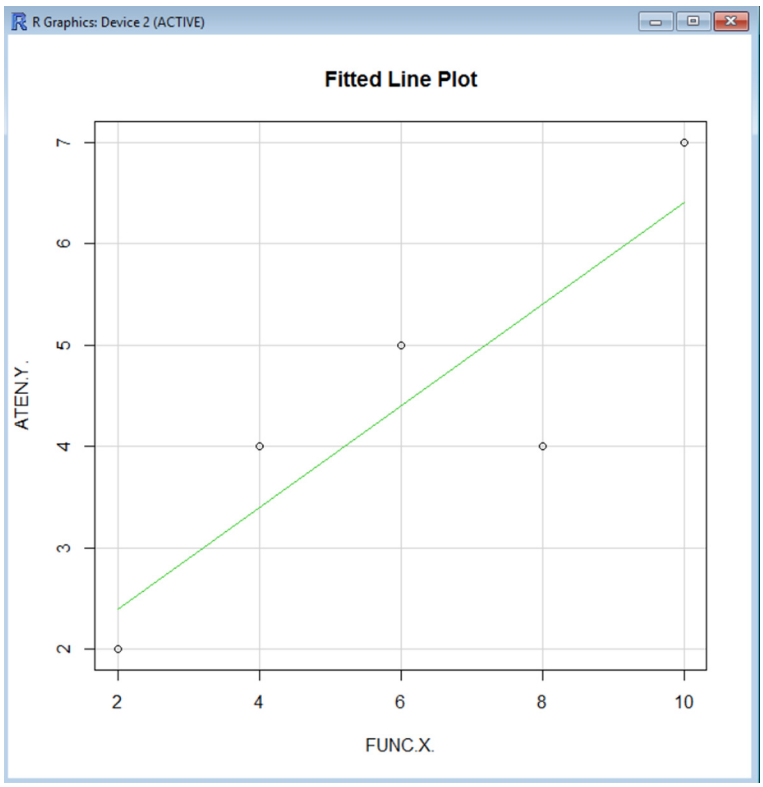
Utilitzem l'opció *Stat*, se segueix la ruta *Gráficas* > *Diagrama de dispersión* **(1)** i a *Opciones* marqueu la casella *Línea de mínimos cuadrados* **(2)**. Seleccioneu *Aplicar* i després *Aceptar* per obtenir el gràfic.



2

Vam obtenir els resultats que apareixen a la figura 8.

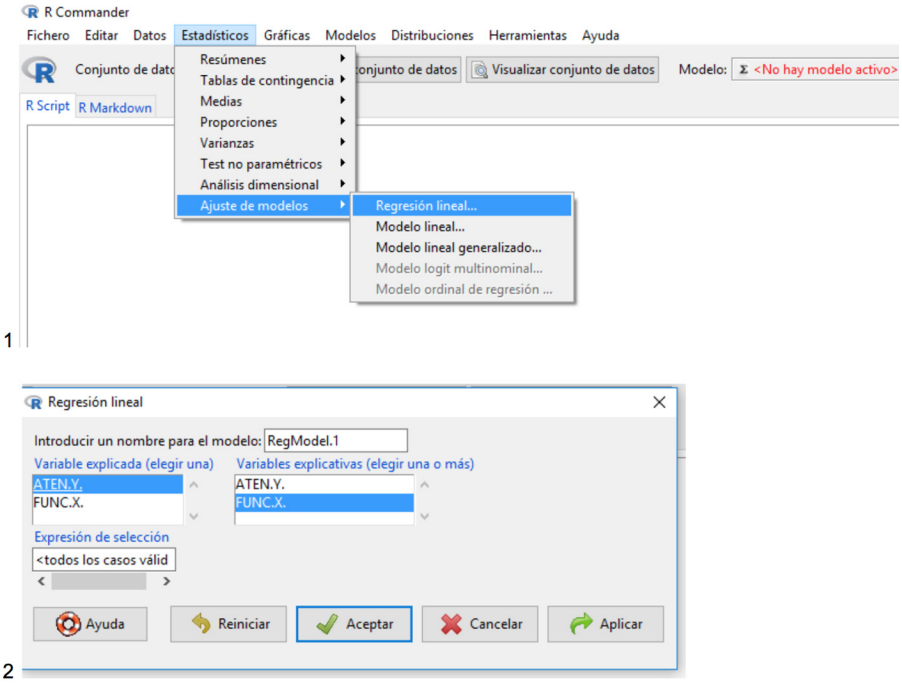
Figura 8. Gràfica de l'equació de regressió de mínims quadrats. R Commander



A continuació interpretarem els resultats:

La figura 8 mostra la gràfica de l'equació de regressió sobre el diagrama de dispersió. El pendent de l'equació de regressió $\hat{\beta}_1$ és positiu, la qual cosa implica que en augmentar les valoracions del funcionament global, les puntuacions d'atenció a l'usuari també augmenten.

Figura 9. Passos a seguir per a fer l'anàlisi de regressió



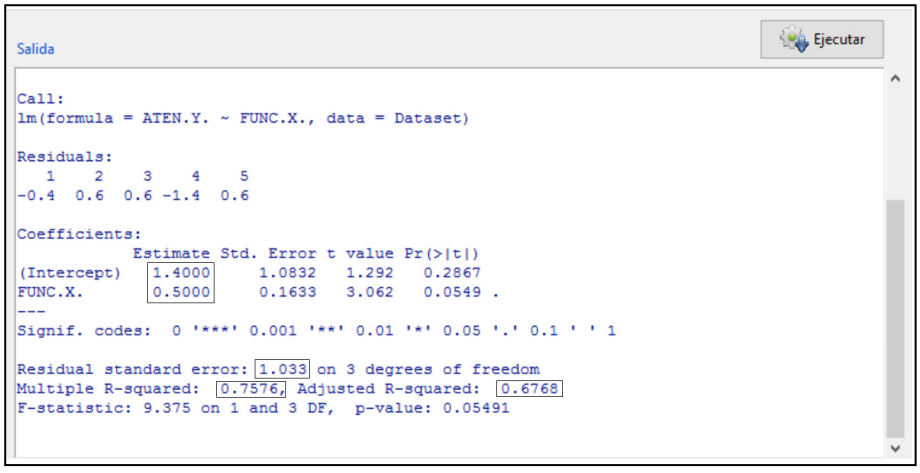
Passos a seguir

Se segueix la ruta *Ajuste de modelos* > *Regresión lineal* **(1)** i s'emplenen els camps a la finestra corresponent **(2)**. Seleccioneu *Aceptar* per obtenir l'anàlisi de regressió.

En el quadre de diàleg de R Commander podem obtenir més informació sobre resultats seleccionant les opcions desitjades. Per exemple, amb aquest quadre de diàleg es poden obtenir els residus, els residuals estandarditzats, els punts d'alta influència i la matriu de correlació (aquests resultats els comentarem més endavant).

Vam obtenir els resultats que apareixen en la figura 10.

Figura 10. Resultats de l'anàlisi de regressió. R Commander



- Interpretació de les estadístiques de regressió:

  R Commander imprimeix l'equació de regressió de la manera següent:

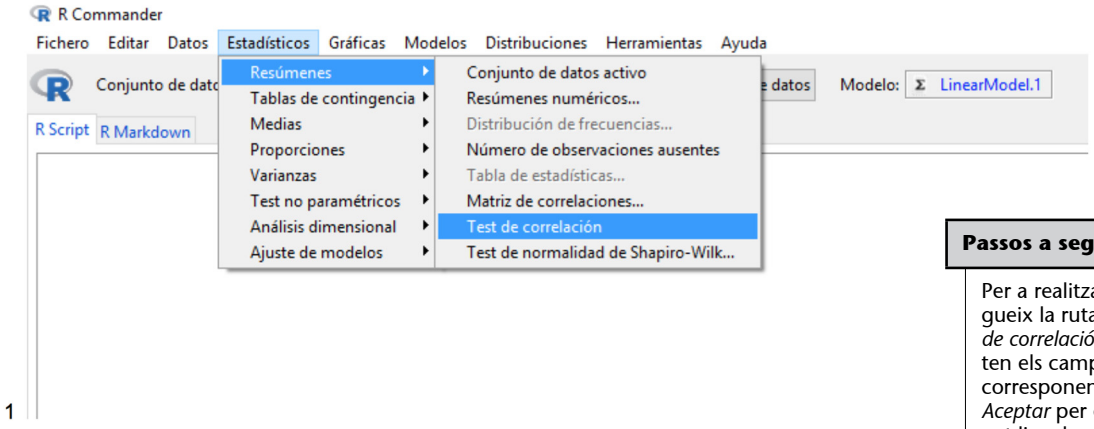
$$\text{ATEN}(Y) = 1{,}40 + 0{,}500\ \text{FUNC}(X)$$

Imprimim una taula que mostra els valors dels coeficients a i b. El coeficient *constant* (ordenada en l'origen) és 1,4, i el pendent amb base en la variable FUNC és de 0,50. *SE Coef* són les desviacions estàndard de cada coeficient. Els valors de les columnes T i P els analitzarem més endavant en estudiar la inferència en la regressió.

El programa imprimeix l'error estàndard del valor estimat, $S = 1{,}033$ mesura la mida d'una desviació típica d'un valor observat (x,y) a partir de la recta de regressió. També proporciona la informació sobre la bondat d'ajust. Observem que $R - Sq = 75{,}8\%$ ($R^2 = 0{,}758$) és el coeficient de determinació expressat en percentatge. Com hem comentat en la solució manual de l'exercici, un valor del 75,8% significa que el 75,8% de la variació en la puntuació d'atenció a l'usuari pot explicar-se per mitjà de la valoració obtinguda en el funcionament global del centre. Se suposa que el 24,2% restant de la variació es deu a la variabilitat aleatòria. El resultat $R - Sq(\text{adj}) = 67{,}7\%$ (R^2 ajustat) és un valor corregit d'acord amb la quantitat de variables independents, el qual és tingut en compte en realitzar una regressió amb diverses variables independents. Més endavant ho estudiarem en tractar la regressió múltiple.

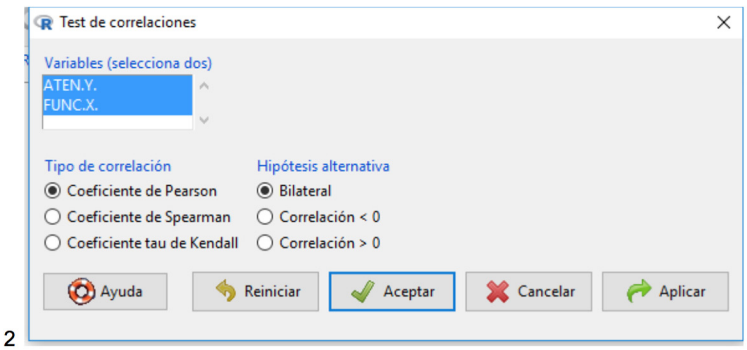
A continuació es calcularà el coeficient de correlació lineal com s'indica en la figura 11.

Figura 11. Passos a seguir per a calcular el coeficient de correlació



Passos a seguir

Per a realitzar el gràfic se segueix la ruta *Resúmenes* > *Test de correlación* **(1)** i es completen els camps a la finestra corresponent **(2)**. Seleccioneu *Aceptar* per obtenir el coeficient lineal.



Vam obtenir els resultats que apareixen en la figura 12.

Figura 12. Resultats de l'anàlisi de correlació. R Commander

```
        Pearson's product-moment correlation

data:  ATEN.Y. and FUNC.X.
t = 3.0619, df = 3, p-value = 0.05491
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05118001  0.99136856
sample estimates:
      cor
0.8703883
```

- Interpretació de l'anàlisi de correlació:

  Com que $r = 0,8703883$, podem dir que hi ha correlació lineal positiva entre les valoracions obtingudes d'atenció a l'usuari i les valoracions del funcionament global del centre. El funcionament està associat positivament a l'atenció a l'usuari.

Observem que $R^2 = 0,758$, per tant $\sqrt{R^2} = \sqrt{0,758} = 0,87 = r$.

Atenció

Per a poder fer regressió amb el **MS Excel** és necessari instal·lar prèviament un complement anomenat *Anàlisi de dades*. Per a instal·lar les eines d'anàlisi de dades feu clic a *Herramientas > complementos*, en el quadre de diàleg, activeu *Herramientas para análisis*.

Per a resoldre l'exemple 1. "Estudi dels serveis oferits per un centre de documentació" utilitzem el **Microsoft Excel.**

La figura 13 mostra la corresponent sortida que ofereix el **Microsoft Excel**.

Observem que les estadístiques de regressió coincideixen amb les obtingudes amb R Commander.

Figura 13. Resultats de l'anàlisi de regressió de l'exemple 1. "Estudi dels serveis oferts per un centre de documentació". Excel

	A	B	C	D	E	F	G	H	I
1	Resum								
2									
3	*Estadístiques de la regressió*								
4	Coeficient de correlació múltiple	0,87038828							
5	Coeficient de determinació R^2	0,757575758							
6	R^2 ajustat	0,676767677							
7	Error típic	1,032795559							
8	Observacions	5							
9									
10	Anàlisi de variància								
11		*Graus de llibertat*	*Suma de quadrats*	*Mitjana dels quadrats*	*F*	*Valor crític de F*			
12	Regressió	1	10	10	9,375	0,054912524			
13	Residus	3	3,2	1,066666667					
14	Total	4	13,2						
15									
16		*Coeficients*	*Error típic*	*Estadístic t*	*Probabilitat*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
17	Intercepció	1,4	1,083205121	1,292460655	0,28674468	-2,047242134	4,847242134	-2,047242134	4,847242134
18	Funcionament (X)	0,5	0,163299316	3,061862178	0,05491252	-0,019691305	1,019691305	-0,019691305	1,019691305
19									
20									
21									
22	Anàlisi dels residuals				Resultats de dades de probabilitat				
23									
24	*Observació*	*Pronòstic atenció (Y)*	*Residus*		*Percentil*	*Atenció (Y)*			
25	1	2,4	-0,4		10	2			
26	2	3,4	0,6		30	4			
27	3	4,4	0,6		50	4			
28	4	5,4	-1,4		70	5			
29	5	6,4	0,6		90	7			
30									

Diagnòstic de la regressió

Igual com en qualsevol procediment estadístic, quan s'efectua una regressió en un conjunt de dades es fan algunes suposicions importants. N'hi ha quatre:

1) El model de línia recta és correcte.

2) Els errors o residus segueixen una distribució aproximadament normal de mitjana zero.

3) Els errors o residus tenen una variància constant σ^2.

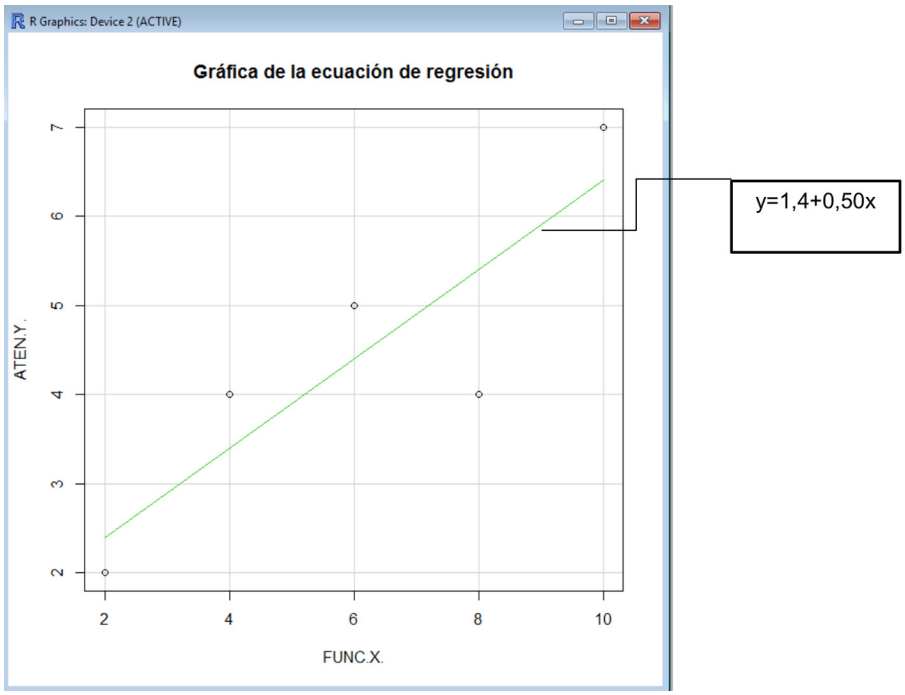
4) Els errors o residus són independents.

Sempre que feu servir regressions per a ajustar una recta a les dades, hem de considerar aquestes suposicions. Comprovar que les dades compleixen aquestes suposicions suposa passar per una sèrie de proves anomenades **diagnosis**, que es descriuen a continuació.

Prova de suposició de línia recta

Per a comprovar si és correcte el model de línia recta utilitzem el gràfic de dispersió amb l'ajust a la recta de mínims quadrats (exemple 1, figura 14).

Figura 14. Gràfica de l'equació de regressió. Exemple 1



Anàlisi de residus

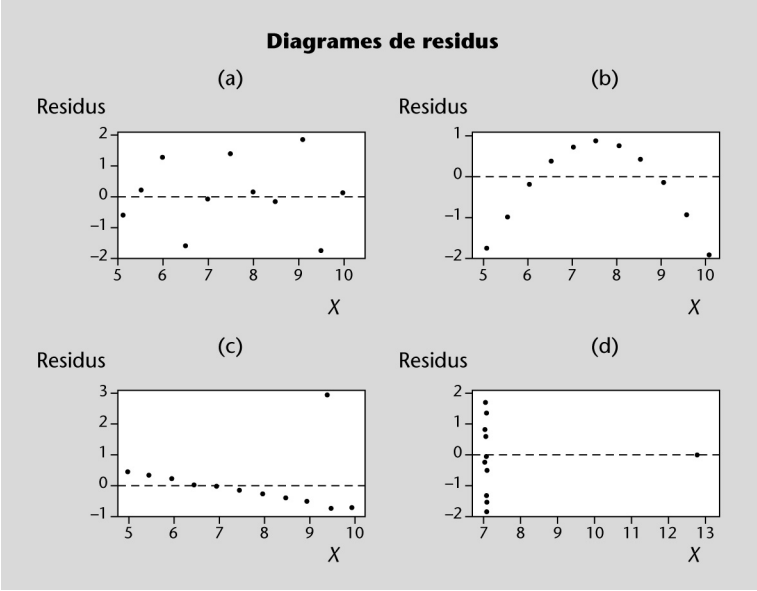
Una vegada fet l'ajust d'un model de regressió lineal a les dades mostrals, cal analitzar els residus o errors. Aquesta anàlisi, que a continuació comentarem de manera breu i intuïtiva, ens servirà per a fer un diagnòstic del model de regressió.

Una altra manera de veure si les dades s'ajusten a una recta és fer un gràfic dels residus ($e_i = y_i - \hat{y}_i$) en funció de la variable predictora (X). Es representa el valor de la variable independent (X) en l'eix horitzontal, i els valors dels residus (e_i) en l'eix vertical.

Podem calcular els residus manualment segons havíem indicat en la taula 3.

En la figura 15 presentem quatre exemples de gràfics de residus o errors.

Figura 15. Diagrama de residus



Podem observar que dels quatre només el primer no presenta cap tipus d'estructura, els residus es distribueixen aleatòriament, de manera que només tindria sentit la regressió feta sobre la mostra (a). Si els punts s'orientessin en forma de U (o U invertida), haurien problemes amb aquest supòsit. És el cas de la mostra (b). Els residus del diagrama c i d no es distribueixen aleatòriament; per tant, no es compleix el supòsit de linealitat.

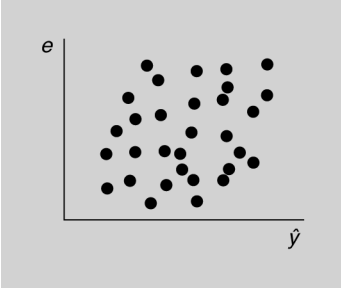
En el mateix gràfic també podem observar si els residus tenen variància constant (supòsit 3). Si la variància dels errors és constant per a tots els valors de X, la gràfica de residuals ha de mostrar un patró similar a una banda horitzontal dels punts, com en (a). Si formen una fletxa (en un extrem s'agrupen molt més que en l'altre), cas (d), aleshores aquest supòsit falla. És convenient també estar atents davant de la possible existència de valors atípics o valors extrems (*outliers*), ja que aquests podrien afectar.

Valor atípic

Per *valor atípic* entenem un valor molt diferent dels altres i que molt probablement és erroni.

També podem fer servir un gràfic de residus en funció de valor estimat o predit $\hat{y}$. Això ho representarem gràficament mitjançant un diagrama de dispersió dels punts $(\hat{y}_i, e_i)$. És a dir, sobre l'eix de les abscisses representem el valor estimat $\hat{y}$, i sobre l'eix d'ordenades, el valor corresponent del residu, de la manera següent: $e_i = y_i - \hat{y}_i$.

Figura 16. Gràfic de residus en funció de valor estimat o predit $\hat{y}$



Si el model lineal obtingut s'ajusta bé a les dades mostrals, llavors el núvol de punts $(\hat{y}_i, e_i)$ no ha de mostrar cap tipus d'estructura. Per a la regressió lineal simple, la gràfica de residus en funció de X i els de residus en funció de $\hat{y}$ donen la mateixa informació. Per a la regressió múltiple, la gràfica de residus en funció de $\hat{y}$ s'usa amb més freqüència, perquè es maneja més d'una variable independent.

Per a comprovar el segon presumpte que els errors o residus segueixen una distribució aproximadament normal utilitzarem la gràfica de probabilitat normal.

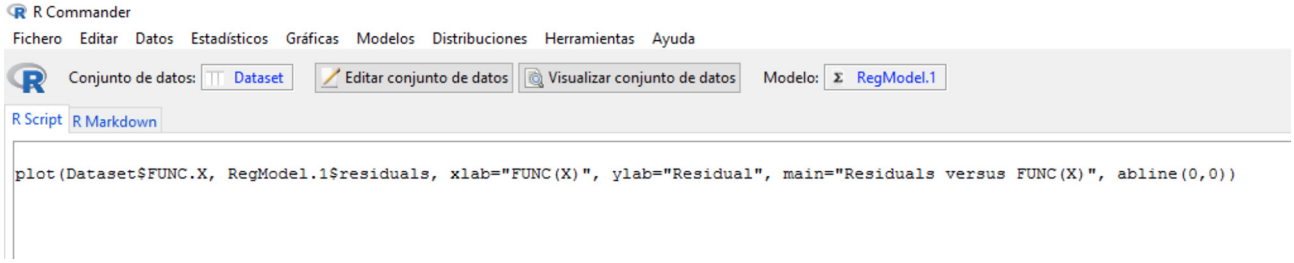
Considerem de nou l'exemple 1. "Estudi dels serveis oferts per un centre de documentació", i fem la diagnosi amb R Commander, a fi de comprovar si es compleixen les condicions del model.

En la figura 17 s'indiquen els passos a seguir per a crear un gràfic dels residus en funció de la variable de predicció amb R Commander. En aquest cas, el gràfic s'ha de programar manualment amb el següent codi:

Passos a seguir

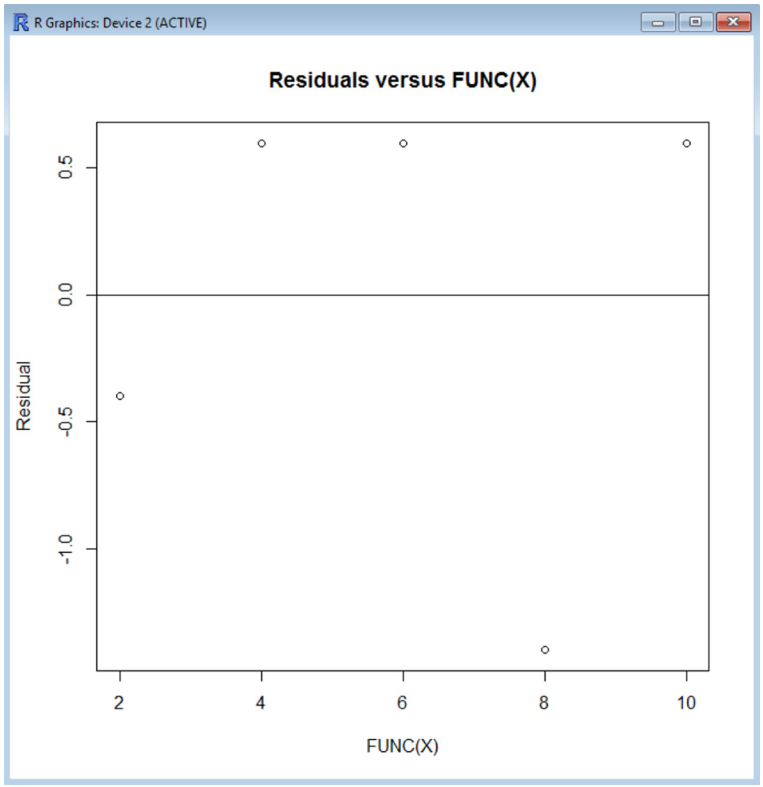
Primer cal seleccionar els valors de la variable *X* (FUNC) i després els residus del model. Els arguments *xlab* i *ylab* són les etiquetes dels eixos corresponents. Per últim, l'argument *main*, és el títol del gràfic.

Figura 17. Codi per a crear un gràfic dels residus en funció de la predicció



Obtenim la gràfica que apareix a la figura 18.

Figura 18. Gràfica dels residus en funció de la variable independent

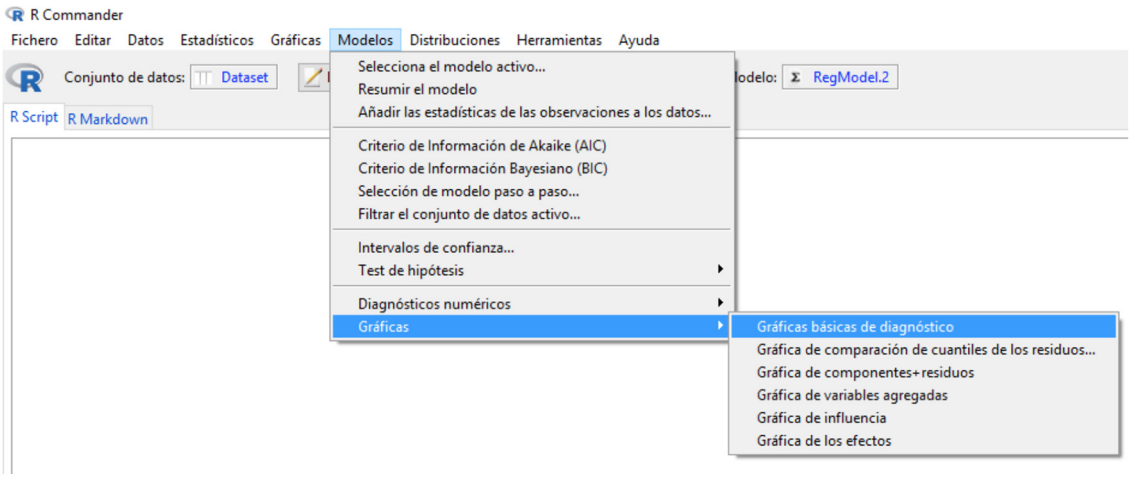


Els valors residuals es distribueixen aleatòriament, sense presentar cap tipus d'estructura; per tant, concloem que la gràfica dels residus no mostra evidència d'incomplir el supòsit de linealitat, i per ara és vàlid el model lineal simple per a l'exemple 1. "Estudi dels serveis oferts per un centre de documentació".

En el mateix gràfic podem observar que els residus tenen variància constant, ja que semblen estar a la banda horitzontal.

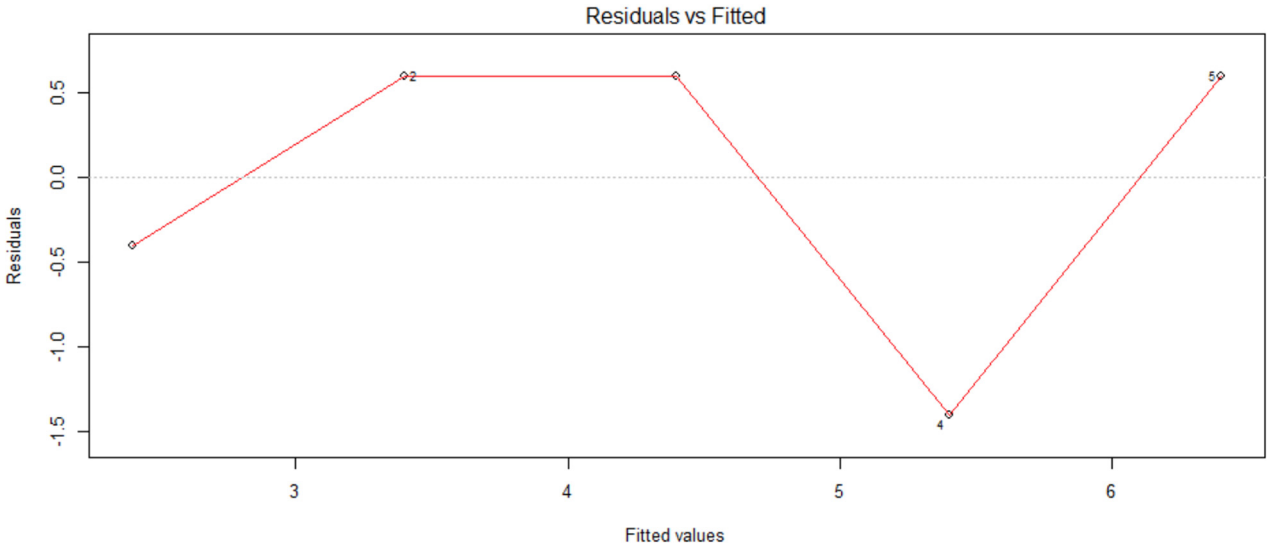
A fi de comprovar si es compleixen la resta de les condicions del model, seleccionem l'opció ***Modelos > Gráficas > Gráficas básicas de diagnóstico*** i completem els camps segons s'indica en la figura 19:

Figura 19. Passos a seguir per a crear un gràfic dels residus en funció dels valors estimats (fits)



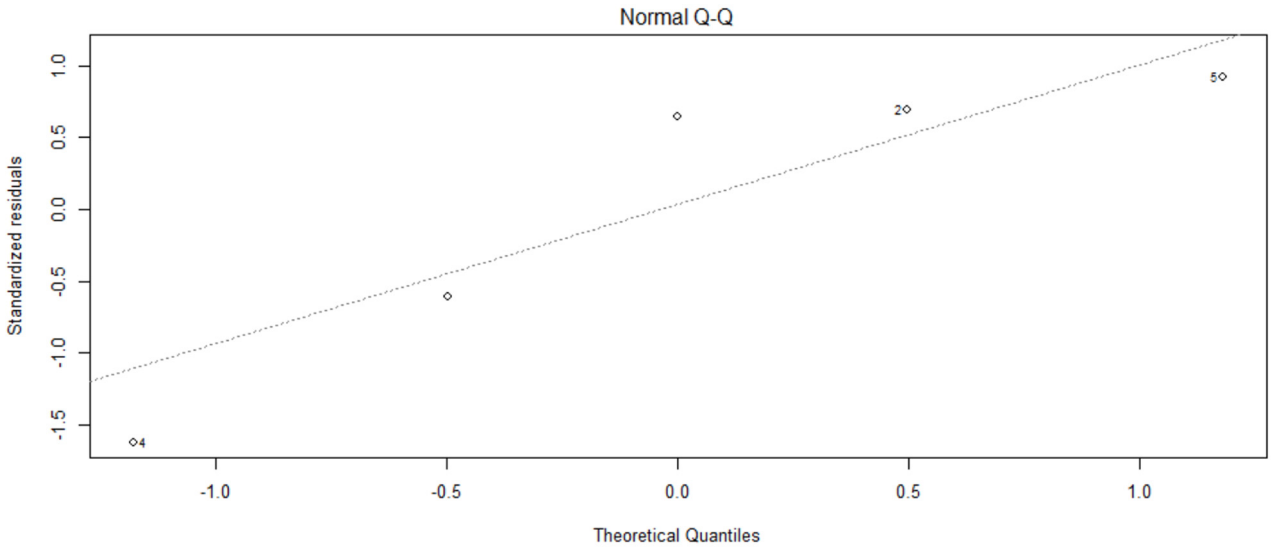
La figura 20 presenta el gràfic dels valors residuals davant els valors estimats i el significat és anàleg al de la figura 18. Els residus es distribueixen aleatòriament, sense presentar cap tipus d'estructura, i podem concloure que és vàlid el model lineal simple.

Figura 20. Gràfica dels residus en funció dels valors estimats



En la gràfica de la figura 21 podem comprovar que els residus segueixen una distribució aproximadament normal, ja que els punts s'apropen bastant en una recta (només si aquests punts s'allunyessin de la forma lineal tindríem dificultats amb aquesta hipòtesi):

Figura 21. Gràfica de probabilitat normal. Exemple 1



Inferència en la regressió: contrastos d'hipòtesi i intervals de confiança

En fer una anàlisi de regressió es comença proposant una hipòtesi quant al model adequat de la relació entre les variables dependent i independent. Per al cas de regressió lineal simple, el model de regressió suposat és

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A continuació apliquem el mètode de mínims quadrats per determinar els valors dels estimadors $\hat{\beta}_0$ i $\hat{\beta}_1$ dels paràmetres del model. L'equació estimada de regressió que en resulta és:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Ja s'ha vist que el valor del coeficient de determinació (R^2) és una mesura de bondat d'ajust d'aquesta equació. Tanmateix, fins i tot amb un valor gran de R^2, no s'hauria de fer servir l'equació de regressió sense abans fer una anàlisi de l'adequació del model proposat. Per a això cal determinar el significat (o importància estadística) de la relació. Les proves de significació en l'anàlisi de regressió es basen en els supòsits següents quant al terme de l'error ε:

1) El terme de l'error ε és una variable aleatòria amb distribució normal amb mitjana, o valor esperat, igual a zero.

2) La variància de l'error, representada per σ^2 és igual per a tots els valors de x.

3) Els valors dels errors són independents.

En l'anàlisi de regressió aplicat, primer es vol conèixer si hi ha una relació entre les variables X i Y. En el model es veu que si β_1 és 0, aleshores no hi ha relació lineal: Y no augmentaria o disminuiria quan augmenta X. Per a esbrinar si hi ha una relació lineal, es pot contrastar la hipòtesi

$$H_0 : \beta_1 = 0$$

enfront de

$$H_1 : \beta_1 \neq 0$$

Es pot contrastar aquesta hipòtesi utilitzant l'estadístic t de Student

Base per a la inferència sobre el pendent de la regressió poblacional

Sigui β_1 el pendent del model de regressió i $\hat{\beta}_1$ la seva estimació per mínims quadrats (basada en observacions mostrals). Si es compleixen els supòsits quant al terme de l'error exposats anteriorment, el pendent del model de regressió β_1, es distribueix com una t de Student amb $(n-2)$ graus de llibertat.

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

Per obtenir l'estadístic de contrast calcularem:

$S_{\hat{\beta}_1}$ és la desviació estàndard estimada de β_1,

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_i^n (x_i - \bar{x})^2}}$$

s és l'error estàndard dels estimats. Para calcular-lo es divideix la suma de les desviacions al quadrat per $n-2$ que són els graus de llibertat.

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}},$$

que es distribueix com una t de Student amb $n-2$ graus de llibertat. La majoria dels programes que es fan servir per a estimar regressions la desviació estàndard del coeficient i l'estadístic t de Student per a $\beta_1 = 0$. Les figures 10 i 13 mostren respectivament les sortides del R Commander i l'Excel corresponents a l'exemple de l'estudi dels serveis oferts per un centre de documentació.

En el cas del model d'exemple, el coeficient del pendent és $\hat{\beta}_1 = 0,50$ amb una desviació estàndard $S_{\hat{\beta}_1} = 0,1633$. Per a saber si hi ha relació entre l'atenció a l'usuari, Y, i el funcionament global, X, es pot contrastar la hipòtesi $H_0 : \beta_1 = 0$ enfront de $H_1 : \beta_1 \neq 0$. Aquest resultat s'obté en el cas d'un contrast de dues cues amb un nivell de significació $\alpha = 0,05$ i 3 graus de llibertat.

L'estadístic t calculat és:

$$t = \frac{0,50 - 0}{0,1633} = 3,06$$

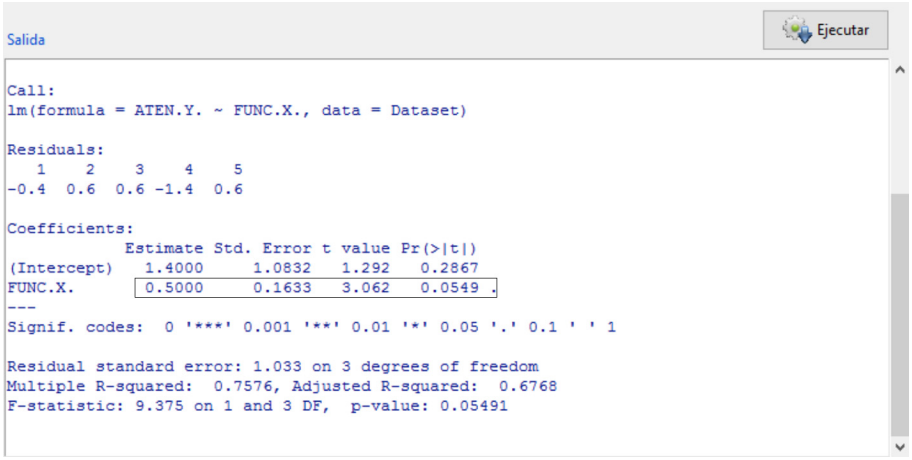
Recordeu

El *p*-valor és la probabilitat que una variable aleatòria superi el valor observat per a l'estadístic de contrast.

- Si *p*-valor $< \alpha$, es rebutja H_0.
- Si *p*-valor $\geq \alpha$, no es rebutja H_0.

L'estadístic *t* resultant, $t = 3,06$, mostrat a la sortida de regressió de la figura 22, és la prova definitiva per a rebutjar o acceptar la hipòtesi nul□la. En aquest cas el *p*-valor és 0,055; com que *p*-valor > 0,05 (no podem rebutjar la H_0: $\beta_1 = 0$ al nivell de significació de $\alpha = 0,05$), s'accepta que $\hat{\beta}_1 = 0$. Per tant, no es pot afirmar que hi hagi una relació lineal entre les valoracions del funcionament global i l'atenció a l'usuari a un nivell de confiança del 95% (nivell de significació del 0,05).

Figura 22. Resum de la figura 10. Resultats de l'anàlisi de regressió. R Commander



Si el nivell de significació s'hagués fixat del 10% ($\alpha = 0,10$), es podria rebutjar H_0, ja que el *p*-valor < 0,10, els resultats indicarien que $\beta_1 \neq 0$ i en aquest cas es podria dir que a un nivell de confiança del 90% hi ha relació lineal entre totes dues variables.

Interval de confiança per al pendent

Es poden obtenir intervals de confiança per al pendent β_1 del model de regressió utilitzant els estimadors dels coeficients i de les variàncies que s'han desenvolupat i el raonament utilitzat en el mòdul 2.

En la sortida de l'anàlisi de regressió de l'atenció a l'usuari quant al funcionament global del centre de documentació de la figura 22 s'observa que

$$n = 5 \; \hat{\beta}_1 = 0,50 \; S_{\hat{\beta}_1} = 0,1633$$

Per a obtenir l'interval de confiança al 95% de β_1, $(1 - \alpha) = 0,95$ i $n - 2 = 3$ graus de llibertat, és necessari calcular el valor crític de la *t* de Student, en aquest cas amb $n - 2 = 5 - 2 = 3$ graus de llibertat i $\alpha/2 = 0,05/2 = 0,025$. Es pot obtenir utilitzant les taules de la distribució *t* de Student o amb l'ordinador.

Si els errors de la regressió ε_i segueixen una distribució normal i es compleixen els supòsits de la regressió, s'obté un interval de confiança al $(1 - \alpha)\%$ del pendent del model de regressió simple β_1 de la manera següent:

$$\hat{\beta}_1 - t_{n-2,\alpha/2} s_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2} s_{\hat{\beta}_1}$$

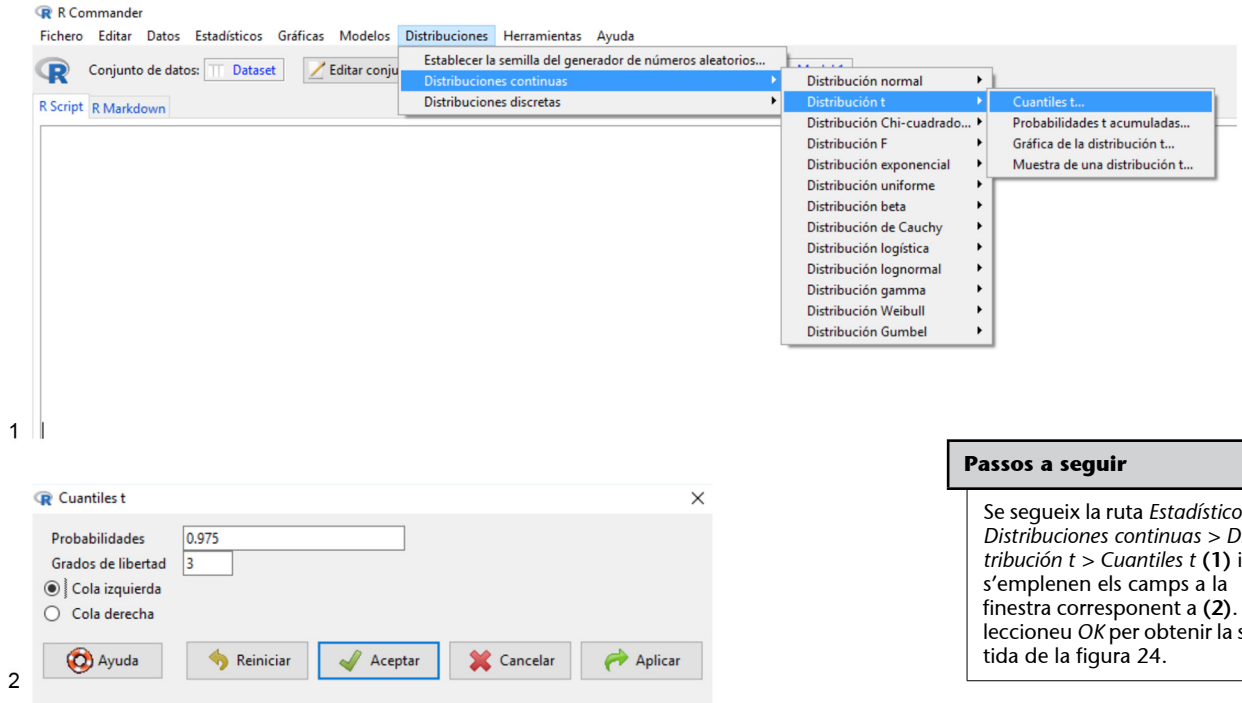
on $t_{n-2,\alpha/2}$ és el nombre per al qual

$$P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$$

l'estadístic t_{n-2} segueix una distribució t de Student amb $(n - 2)$ graus de llibertat.

Si es fa servir el R Commander, els passos a seguir es mostren en la figura 23.

Figura 23. Passos a seguir per a calcular el valor crític *t*



Passos a seguir

Se segueix la ruta *Estadísticos* > *Distribuciones continuas* > *Distribución t* > *Cuantiles t* **(1)** i s'emplenen els camps a la finestra corresponent a **(2)**. Seleccioneu *OK* per obtenir la sortida de la figura 24.

Figura 24. Resultats de càlcul del valor crític *t*. R Commander

```
> qt(c(0.975), df=3, lower.tail=TRUE)
[1] 3.182446
```

el valor de $t_{n-2,\alpha/2} = t_{3;0,025} = 3,18$

Per tant, l'interval de confiança al 95% serà

$$0,50 - (0,1633)\ (3,18) < \beta_1 < 0,50 + (0,1633)\ (3,18)$$

O sigui

$$-0{,}019 < \beta_1 < 1{,}0193$$

Per tant, l'interval de confiança buscat és: $0{,}50 \pm 3{,}18245 \square 0{,}163$; per exemple, es pot afirmar amb una probabilitat del 95% que β_1 és a l'interval d'extrems –0,0197 i 1,0197.

En la taula 4 hi ha la representació de l'interval de confiança calculat amb l'Excel. El resum mostra en les últimes columnes els valors estimats d'interval de confiança del 95% per als paràmetres de regressió β_0 i β_1 , també les desviacions estàndards estimades (columna "Error típic"), el valor estadístic *t* (columna "Estadístic *t*") i els *p*-valors (columna "Probabilitat").

Taula 4. Resum de la figura 13 (resultats de l'anàlisi de regressió. Excel)

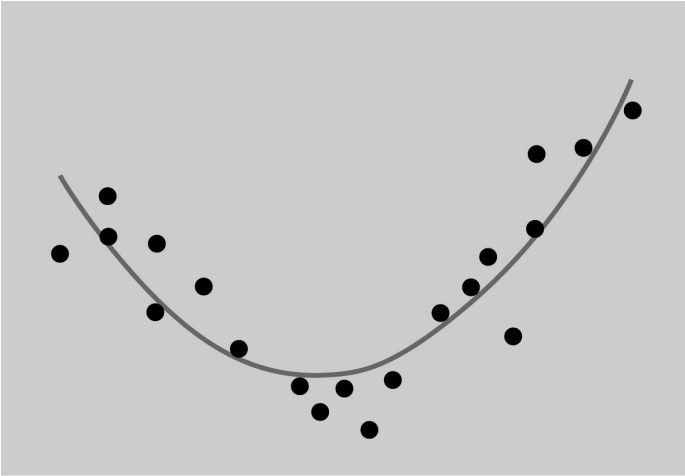
	Coeficients	Error típic	Estadístic *t*	Probabilitat	Inferior 95%	Superior 95%
Intercepció	1,4	1,08320512	1,29246066	0,286745	–2,047242	4,847242134
Funcionament (*X*)	0,5	0,16329932	3,06186218	0,054913	–0,019691	1,019691305

3.2. Models de regressió simple no lineals: model quadràtic i cúbic

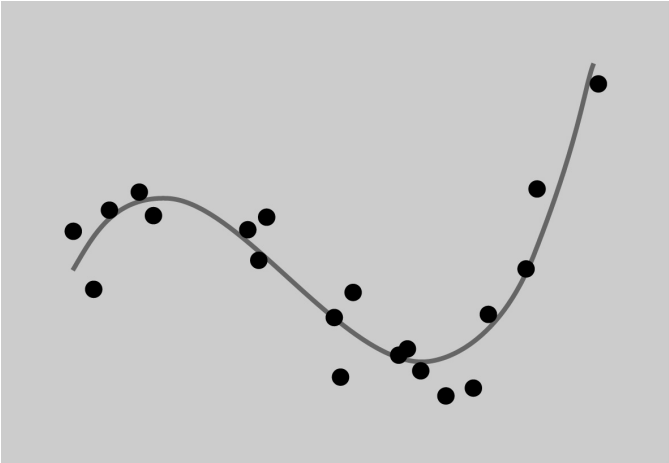
Hi ha algunes relacions que no són estrictament lineals i es poden desenvolupar mètodes per a poder utilitzar els mètodes de regressió i estimar els coeficients del model.

A part dels models de regressió lineals, en podem establir d'altres que no són lineals, entre els quals destaquem: el model quadràtic i el cúbic, que són models curvilinis. Cada model correspon al grau de l'equació, on *Y* és la resposta i *X* és la variable predictora, β_0 és l'ordenada a l'origen, i β_1, β_2, i β_3 són els coeficients. És important escollir el model apropiat quan es modelitzen dades usant regressió i anàlisi de tendència.

Model quadràtic: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$



Model cúbic: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$



Per a determinar quin model utilitzar, es representen prèviament les dades (diagrama de dispersió) i es calcula el coeficient de correlació lineal de Pearson. Convé recordar que l'esmentat coeficient r mesura el grau d'associació lineal que hi ha entre les variables X i Y quan s'ajusta al seu núvol de punts una línia recta, però no mesura el grau d'ajust d'una corba al núvol de punts. Es podria donar el cas que la relació entre les variables fos gran, només que distribuïda al llarg d'una corba, i aleshores, en ajustar a una recta s'obtindria un coeficient de correlació lineal r i un coeficient de determinació R^2 baix. Calcularíem l'ajust simultani als models no lineals (quadràtic i cúbic) i es calcularien els coeficients de determinació per a ambdós models, per a determinar la bondat de l'ajust. El millor model serà el que presenti el valor més elevat de R^2.

Els mètodes d'inferència per als models no lineals transformats són els mateixos que s'han desenvolupat per als models lineals. Així, doncs, si tenim un model quadràtic, l'efecte d'una variable X està indicat per als coeficients tant dels termes lineals com dels termes quadràtics.

Exemple 2. Nombre de visitants a un museu (estimació d'un model quadràtic utilitzant el R Commander)

Es vol estudiar la variació entre el nombre de visitants a un museu en funció del nombre d'obres visitades. La taula 5 mostra el nombre de visitants i el nombre d'obres visitades. S'han seleccionat aleatòriament les dades corresponents a 6 dies.

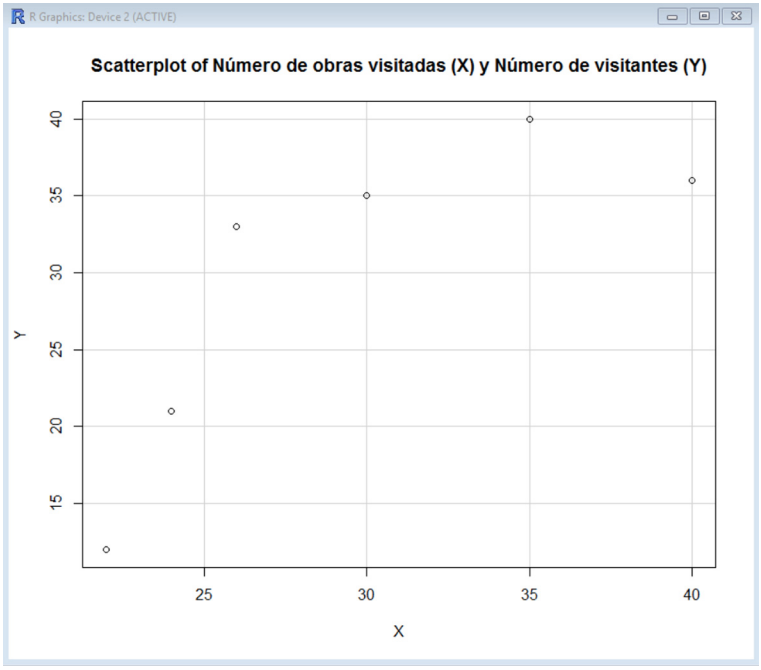
Taula 5. Nombre de visitants a un museu

Nombre de visitants (X)	22	24	26	30	35	40
Nombre d'obres visitades (Y)	12	21	33	35	40	36

Amb aquestes dades podem deduir si hi ha relació entre ambdues variables i, si les variables estan relacionades, podem establir el millor model.

La figura 25 representa el diagrama de dispersió per a aquestes dades. El diagrama de dispersió indica que possiblement hi ha una relació curvilínia entre el nombre d'obres visitades i el nombre de visitants.

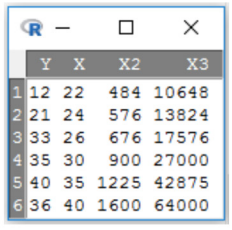
Figura 25. Diagrama de dispersió per a exemple 2. R Commander



Abans de deduir l'equació curvilínia entre el nombre d'obres visitades i el nombre de visitants, es fa l'ajust a un model de regressió lineal simple (de primer ordre) tot seguint els passos que mostra la figura 26.

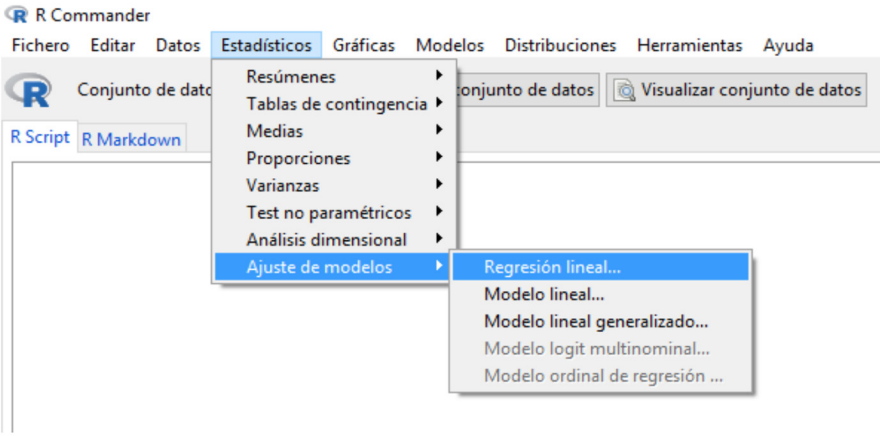
Figura 26. Passos a seguir per a comprovar el model lineal

1



	Y	X	X2	X3
1	12	22	484	10648
2	21	24	576	13824
3	33	26	676	17576
4	35	30	900	27000
5	40	35	1225	42875
6	36	40	1600	64000

2



Passos a seguir

Primer es carreguen les dades. Per calcular el model quadràtic i cúbic s'han de crear les variables *x2* i *x3* com es mostra a **(1)**. A continuació, se segueix la ruta *Estadísticos* > *Ajuste de modelos* > *Regresión lineal* **(2)** i s'emplenen els camps a la finestra corresponent. Seleccioneu *Aceptar* per obtenir la sortida de la figura 28.

Figura 27. Gràfica de l'equació de regressió de mínims quadrats

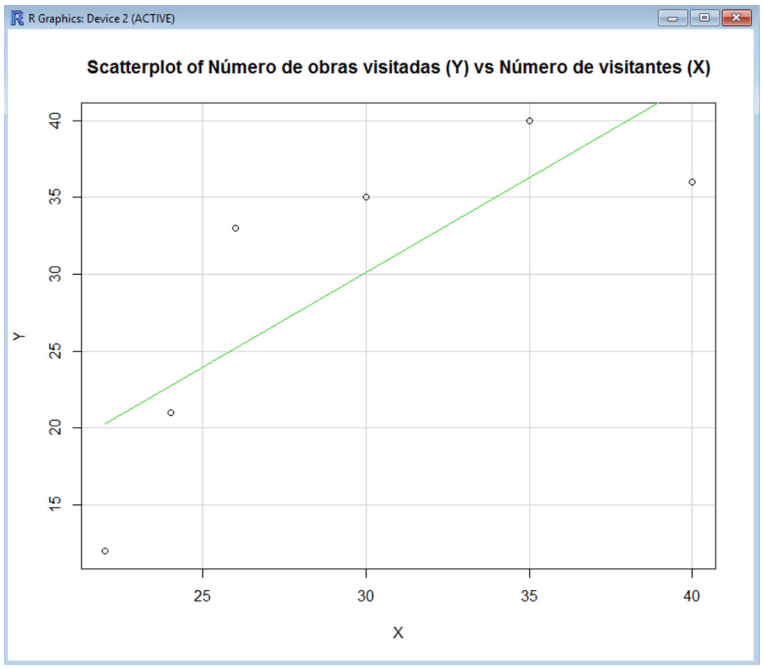


Figura 28. Resultats de l'anàlisi de regressió. Model lineal simple

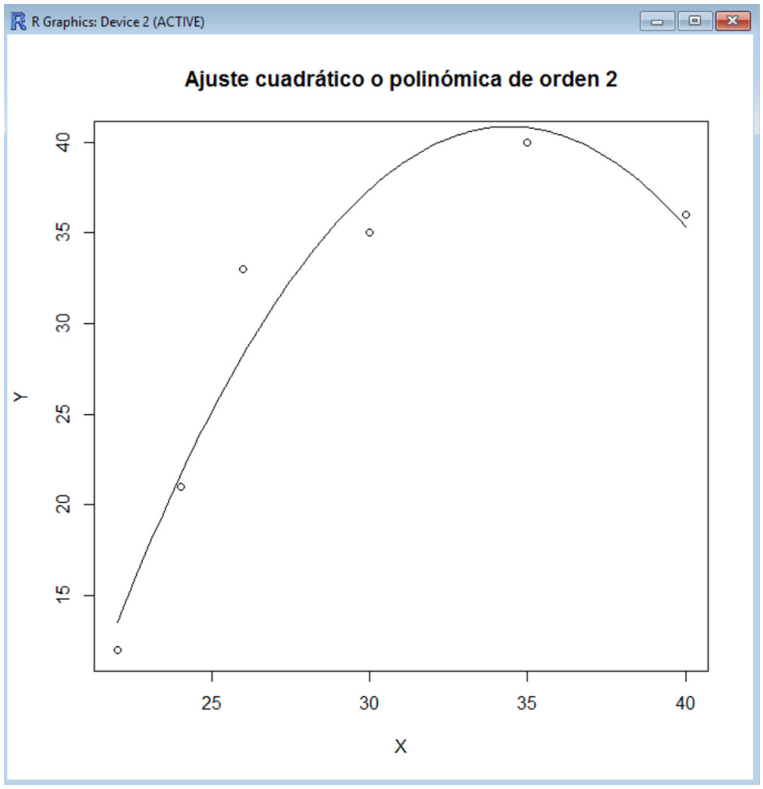
```
> summary(RegModel.3)

Call:
lm(formula = Y ~ X, data = Dataset)

Residuals:
     1      2      3      4      5      6
-8.278 -1.737  7.804  4.885  3.737 -6.411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.7745    14.1709  -0.478   0.6576
X             1.2296     0.4697   2.618   0.0589 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.269 on 4 degrees of freedom
Multiple R-squared:  0.6314, Adjusted R-squared:  0.5393
F-statistic: 6.853 on 1 and 4 DF,  p-value: 0.05893
```

Observem que amb el model lineal s'explica un 63,1% de la variabilitat del nombre de visitants (R^2 = 63,1%). L'equació d'ajust és:

Nombre de visitants (Y) = –6,77 + 1,2296 Nombre d'obres visitades (X)

A continuació es presenta l'ajust del model quadràtic i com podeu veure en la gràfica de la figura 29 els punts s'ajusten millor a una funció no lineal.

Figura 29. Gràfica de l'ajust quadràtic



Observem que l'ajust quadràtic és molt bo amb un valor de R^2 = 94,48%, el qual millora l'ajust lineal. L'equació d'ajust és:

$$\text{Nombre de visitants } (Y) = -168{,}9 + 12{,}19 \text{ nombre d'obres visitades} - 0{,}1770 \text{ nombre d'obres visitades}^2$$

Figura 30. Resultats de l'anàlisi de regressió. Model quadràtic

```
Call:
lm(formula = Y ~ X + X2, data = Dataset)

Residuals:
      1       2       3       4       5       6
-1.5441 -0.6309  4.6987 -2.3934 -0.7918  0.6615

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -168.8848    39.7862  -4.245   0.0239 *
X             12.1870     2.6632   4.576   0.0196 *
X2            -0.1770     0.0429  -4.127   0.0258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.248 on 3 degrees of freedom
Multiple R-squared:  0.9448, Adjusted R-squared:  0.908
F-statistic: 25.68 on 2 and 3 DF,  p-value: 0.01297
```

A continuació es presenta l'ajust del model cúbic:

Figura 31. Gràfica de l'ajust cúbic

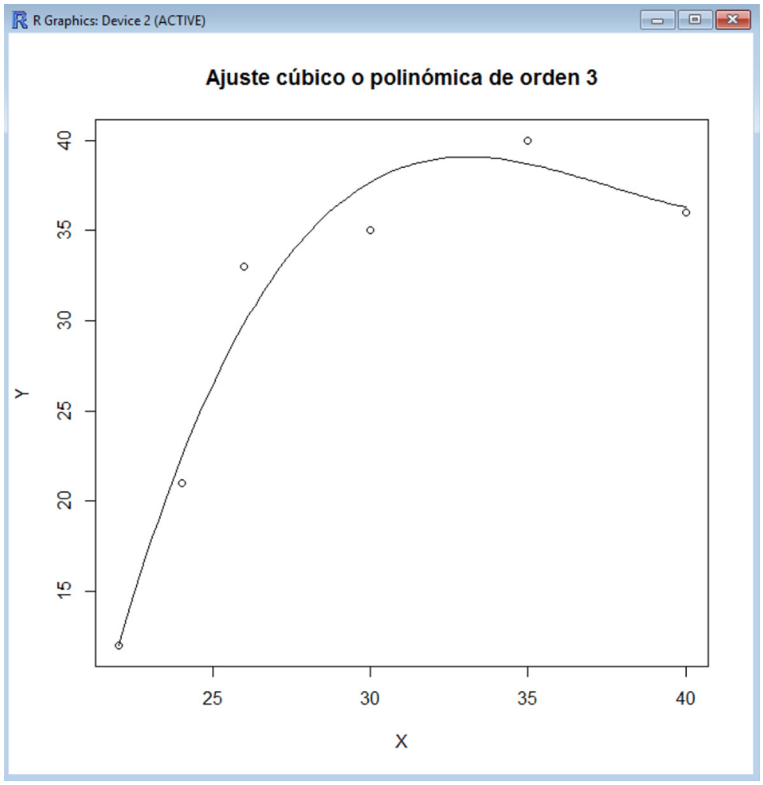


Figura 32. Resultats de l'anàlisi de regressió. Model cúbic

```
Call:
lm(formula = Y ~ X + X2 + X3, data = Dataset)

Residuals:
       1        2        3        4        5        6
-0.03848 -1.42495  3.12856 -2.68802  1.31448 -0.29158

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.209e+02  2.509e+02  -1.678    0.235
X            3.775e+01  2.527e+01   1.494    0.274
X2          -1.021e+00  8.310e-01  -1.229    0.344
X3           9.081e-03  8.926e-03   1.017    0.416

Residual standard error: 3.229 on 2 degrees of freedom
Multiple R-squared:  0.9636, Adjusted R-squared:  0.9091
F-statistic: 17.66 on 3 and 2 DF,  p-value: 0.05406
```

L'ajust al model cúbic també és bo amb un valor alt de R^2 = 96,36% que millora l'ajust lineal i iguala el quadràtic.

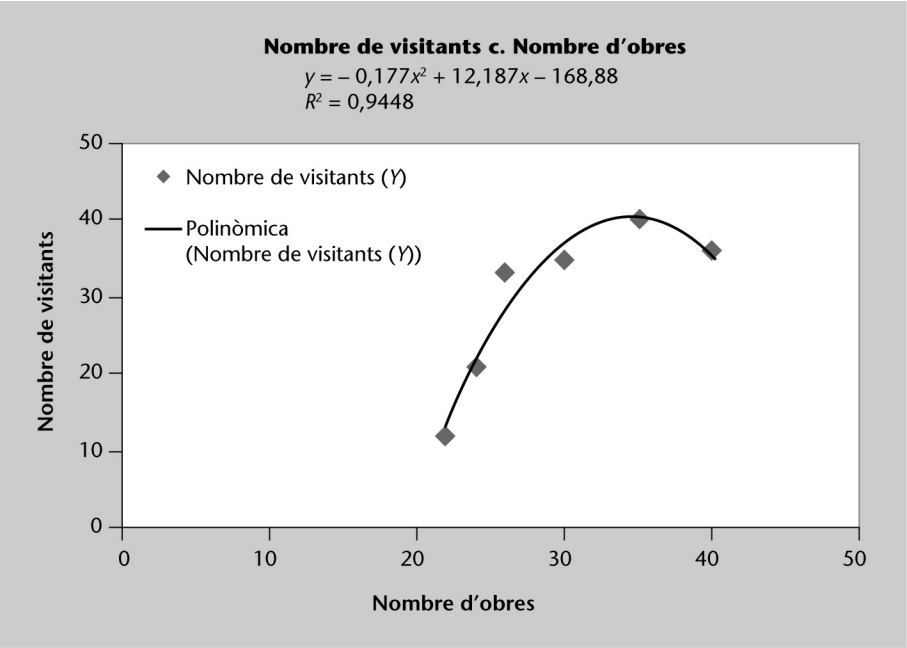
L'equació d'ajust és:

Nombre de visitants (Y) = –420,9 + 37,75 nombre d'obres visitades – 1,021 nombre d'obres visitades^2 + 0,009081 nombre d'obres visitades^3

Analitzant la significativitat dels models per mitjà del *p*-valor, el model quadràtic pel fet de tenir el *p*-valor més petit (0,01297) és el més significatiu. Per això s'escolliria com a millor ajust el quadràtic.

La figura 33 mostra la corresponent sortida que ofereix el **Microsoft Excel** de l'exemple 2. "Nombre de visitants a un museu". Seleccionant l'opció *Tipus de tendència poligonal de segon ordre*, que coincideix amb l'ajust quadràtic elegit amb R Commander (figures 29 i 30). L'equació d'ajust i el valor de R^2 coincideixen amb les que hem obtingut amb R Commander.

Figura 33. Gràfica de l'ajust quadràtic. Excel



3.3. Transformacions de models de regressió no lineals: models exponencials

Algunes relacions entre variables poden analitzar-se mitjançant models exponencials. Per exemple, les relacions entre la variable temps (X) i altres variables (Y) com la població, els preus d'alguns productes, el nombre d'ordinadors infectats, són exponencials. Els models exponencials de demanda es fan servir força en l'anàlisi de conducta del mercat.

Corba en un model exponencial

En el model lineal s'ajusta el núvol de punts a una recta d'equació:

$$y = a + bx$$

En el model exponencial s'ajusta a una corba d'equació:

$$y = ka^x \quad con\, a > 0,\ k > 0$$

El model exponencial és del tipus:

$$y = ka^x \quad con\, a > 0, k > 0$$

on k i a són valors constants.

Per a tractar aquest model es farà una transformació de les variables de manera que el model es converteixi en lineal.

Si a l'equació $y = ka^x$ es prenen logaritmes $\ln y = \ln(ka^x)$, s'obté, per l'aplicació de les propietats dels logaritmes:

Propietats dels logaritmes

$\ln ab = \ln a + \ln b$

$\ln a^x = x \ln a$

$$\ln y = \ln k + x \ln a$$

Aquesta equació mostra un model lineal entre les variables X i $\ln Y$.

Si es representa el diagrama de dispersió dels punts $(x_i, \ln y_i)$ i el núvol de punts presenta una estructura lineal, es pot pensar que entre les variables X i Y hi ha una relació exponencial.

4. Models de regressió múltiple

En l'apartat 3.1 s'ha presentat el mètode de regressió simple per a obtenir una equació lineal que prediu una variable dependent o endògena en funció d'una única variable independent o exògena: nombre total de llibres venuts en funció del preu. Tot i això, en moltes situacions, diverses variables independents influeixen conjuntament en una variable dependent. La regressió múltiple permet esbrinar l'efecte simultani de diverses variables independents en una variable dependent fent servir el principi dels mínims quadrats.

Hi ha moltes aplicacions de la regressió múltiple per a respondre preguntes com les següents:

- En quina mesura el preu d'un ordinador depèn de la velocitat del processador, de la capacitat del disc dur i de la quantitat de memòria RAM?
- Com relacionar l'índex d'impacte d'una revista científica amb el nombre total de documents publicats i el nombre de citacions per document?
- El sou d'un titulat depèn de l'edat, dels anys que fa que va acabar els estudis, dels anys d'experiència en 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 ciutat, etc.?
- El preu d'un cotxe depèn de la potència del motor, del nombre de portes i de multitud d'accessoris que pot portar: coixins de seguretat, ordinador de viatge, equip d'alta fidelitat, volant esportiu, llandes especials, etc.?

Els mètodes per a ajustar models de regressió múltiple es basen en el mateix principi de mínims quadrats explicat en l'apartat 3.1.

El nostre objectiu és aprendre a fer servir la regressió múltiple per a crear i analitzar models. Per tant, s'aprendrà com funciona la regressió múltiple i algunes directrius per a interpretar-la. En comprendre perfectament la regressió múltiple, és possible resoldre una àmplia varietat de problemes aplicats. Aquest estudi dels mètodes de regressió múltiple és paral□lel al de regressió simple. El primer pas per a desenvolupar un model consisteix en la selecció de les variables i de la forma del model. A continuació, s'estudia el mètode de mínims quadrats i s'analitza la variabilitat per a identificar els efectes de cada una de les variables de predicció.

Després s'estudia l'estimació, els intervals de confiança i el contrast d'hipòtesi. Fem servir aplicacions informàtiques per a indicar com s'aplica la teoria a problemes reals.

Desenvolupament del model

Quan s'aplica la regressió múltiple, es construeix un model per a explicar la variabilitat de la variable dependent. Per a això cal incloure les influències simultànies i individuals de diverses variables independents. Se suposa, per exemple, que es vol desenvolupar un model que predigui el preu de les impressores làser que vol liquidar una empresa. Un estudi inicial indicava que el preu estava relacionat amb el nombre de pàgines per minut que la impressora és capaç d'imprimir i els anys d'antiguitat de la impressora en qüestió. Això duria a especificar el model següent de regressió múltiple amb dues variables independents.

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

on:

Y = preu en euros
X_1 = nombre de pàgines impreses per minut
X_2 = anys d'antiguitat de la impressora

La taula 6 conté 12 observacions d'aquestes variables. Es faran servir aquestes dades per a desenvolupar el model lineal que predigui el preu de les impressores en funció del nombre de pàgines impreses per minut i dels anys d'antiguitat de la impressora.

Taula 6. Dades de l'exemple 3. "Estudi sobre el preu d'impressores làser en funció de la seva velocitat d'impressió i l'antiguitat del model"

X_1	6	6	6	6	8	8	8	8	12	12	12	12
X_2	6	4	2	0	6	4	2	0	6	4	2	0
Y	466	418	434	487	516	462	475	501	594	553	551	589

Nota

En el cas general emprarem k per a representar el nombre de variables independents.

Però abans de poder estimar el model cal desenvolupar i comprendre el mètode de regressió múltiple.

El model de regressió múltiple és

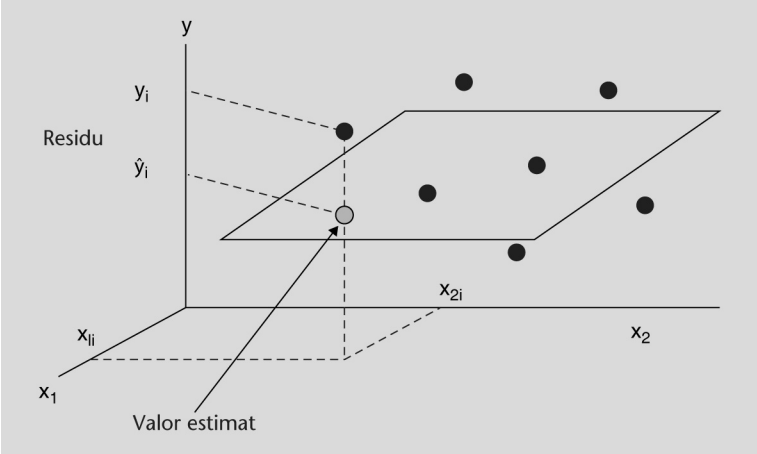
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

On $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ són els coeficients de les variables independents o exògenes i ε (lletra grega èpsilon) és l'error o residu i és una variable aleatòria. Més endavant descriurem tots els supòsits del model per al model de regressió múltiple i per a ε.

Els coeficients, en general, no es coneixen i cal determinar-los a partir de les dades d'una mostra i fent servir el **mètode de mínims quadrats** per a arribar a l'equació estimada de regressió que més s'aproxima a la relació lineal entre les variables independents i dependent. El procediment és similar al que es fa

servir en la regressió simple. En la regressió múltiple el millor ajust és un hiperplà en espai *n*-dimensional (espai tridimensional en el cas de dues variables independents, figura 34).

Figura 34. Gràfica de l'equació de regressió, per a l'anàlisi de regressió múltiple amb dues variables independents



Els valors estimats de la variable dependent es calculen amb l'equació estimada de regressió múltiple:

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

On $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ són els valors dels estimadors dels paràmetres o coeficients de l'equació de regressió múltiple. La deducció d'aquests coeficients requereix l'ús de l'àlgebra de matrius i se surt del propòsit d'aquest text. Així, en descriure la regressió múltiple ho enfocarem cap a com podem emprar els programes informàtics de càlcul per a obtenir l'equació estimada de regressió i altres resultats i com interpretar-la, i no cap a com fer els càlculs de la regressió múltiple.

Criteri de mínims quadrats

$$\min \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

on:

y_i = valor observat de la variable dependent en la i-èsima observació.

$\hat{y}_i$ = valor estimat de la variable dependent en la *i*-èsima observació.

Considerant de nou el model de regressió amb dues variables independents de l'exemple 3. "Estudi sobre el preu d'impressores làser en funció de la seva velocitat d'impressió i l'antiguitat del model". Utilitzant les dades de la taula 6 s'ha estimat un model de regressió múltiple, que s'observa a la sortida R Commander de la figura 35.

Figura 35. Resultats de l'exemple 3 de l'anàlisi de regressió múltiple per a dues variables independents

```
Call:
lm(formula = Y ~ X1 + X2, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-31.971 -18.611  -3.343  18.488  35.629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  330.407     29.795  11.089  1.5e-06 ***
X1            20.161      3.097   6.509  0.00011 ***
X2            -0.350      3.455  -0.101  0.92154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.76 on 9 degrees of freedom
Multiple R-squared:  0.8248, Adjusted R-squared:  0.7859
F-statistic: 21.19 on 2 and 9 DF,  p-value: 0.0003941
```

Passos a seguir

Per a estimar el model de regressió múltiple introduïm les dades en R Commander per a calcular el model.

Se segueix la ruta *Estadísticos* > *Ajuste de modelos* > *Regresión lineal* i es completen els camps a la finestra corresponent. Seleccioneu *Aceptar* per obtenir l'anàlisi de regressió.

Els coeficients estimats s'identifiquen a la sortida dels programes informàtics

L'equació de regressió múltiple és: $Y = 330{,}4 + 20{,}2\ X1 - 0{,}35\ X2$

La interpretació dels coeficients és la següent:

- Coeficient de $X1$ (20,2 euros): seria l'augment del preu de la impressora quan augmenta en una unitat el nombre de pàgines per minut que imprimeix, quan les altres variables independents es mantenen constants (en aquest cas $X2$, l'antiguitat no varia).

- Coeficient $X2$ (–0,35 euros): seria la disminució del preu per cada any més d'antiguitat de la impressora, quan $X1$ roman constant (el nombre de pàgines per minut no varia).

- Terme independent (330,4): no té molt sentit interpretar-lo en aquest cas, ja que representaria el preu d'una impressora que no pot imprimir cap pàgina.

El coeficient de determinació múltiple

Coeficient de determinació R^2

Coeficient de determinació R^2 en el R Commander es designa com a ***Multiple R-squared***.

En la regressió lineal simple, vam veure que la suma total de quadrats es pot descompondre en dos components: la suma de quadrats deguda a la regressió i la suma de quadrats deguda a l'error. Aquest mateix procediment s'aplica a la suma de quadrats de la regressió múltiple. El coeficient de determinació múltiple mesura la bondat d'ajust per a l'equació de regressió múltiple. Aquest coeficient es calcula com segueix:

$$R^2 = \frac{SSR}{SST}$$

Podem interpretar-ho com la proporció de variabilitat de la variable dependent que podem explicar amb l'equació de regressió múltiple. Quan es multi-

plica per cent, s'interpreta com la variació percentual de y que s'explica amb l'equació de regressió.

En general R^2 augmenta quan s'afegeixen variables independents (variables explicatives o predictores) al model. Si s'afegeix una variable al model, R^2 es fa més gran (o roman igual), tot i que aquesta variable no sigui estadísticament significativa. El **coeficient de determinació corregit** o ***Adjusted R-squared*** elimina l'efecte que es produeix sobre *Multiple R-squared* quan s'augmenta el nombre de variables independents.

El **coeficient de correlació múltiple** es defineix com l'arrel quadrada positiva de *Multiple R-squared*. Aquest coeficient ens proporciona la correlació existent entre la variable dependent (resposta) i una nova variable formada per la combinació lineal dels predictors.

Continuant amb l'exemple 3. "Estudi sobre el preu d'impressores làser en funció de la seva velocitat d'impressió i l'antiguitat del model", interpretarem el resultat del coeficient de determinació *Multiple R-squared* = 82,48% (figura 35), significa que el 82,48% de la variabilitat en el preu d'impressores làser s'explica amb l'equació de regressió múltiple, amb el nombre de pàgines que imprimeix per minut i els anys d'antiguitat. La figura 35 mostra que el valor *Adjusted R-squared* = 78,59%, significa que si s'agregués una variable independent (predictora) el valor de R^2 no augmentaria.

Supòsits del model

Els supòsits sobre el terme de l'error ε, en el model de regressió múltiple, són similars als del model de regressió lineal simple.

Per simplicitat, considerarem un model de regressió amb només dues variables explicatives ($X1$ i $X2$). L'equació de regressió múltiple, amb dues variables independents, serà:

$$y = \beta_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

on els β_i representen coeficients reals i ε representa l'error aleatori.

1) L'error és una variable aleatòria el valor mitjà o esperat de la qual és zero; això és $E(\varepsilon) = 0$.

2) Per a tots els valors de $X1$ i $X2$, els valors de Y (o, alternativament, els valors de εi) mostren variància constant σ^2.

3) Per a cada valor de $X1$ i $X2$, la distribució de Y (o, alternativament, la de ε) és aproximadament normal.

4) Els valors de Y obtinguts (o, alternativament, els de ε) són independents.

Hi ha tota una sèrie de gràfics que ens poden ajudar a analitzar els resultats d'una regressió lineal múltiple i a comprovar si es compleixen o no els supòsits anteriors:

1) Un gràfic de la variable dependent enfront dels valors estimats pel model ens ajudarà a comprovar visualment la bondat de l'ajust.

2) Representant els residus enfront dels valors estimats podrem comprovar la variabilitat vertical en les dades. Això ens permetrà saber si es compleix el supòsit de variància constant.

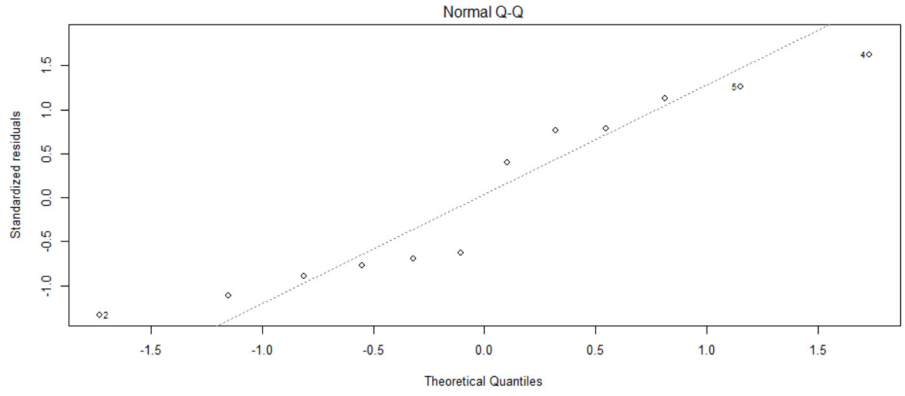
3) Un gràfic de residus enfront de cada una de les variables explicatives pot revelar problemes addicionals que no s'hagin detectat en el gràfic anterior.

4) Per a comprovar la hipòtesi de normalitat sol ser convenient realitzar un test i un gràfic de normalitat per als residus.

En l'exercici es comprova si es compleixen els supòsits del model utilitzat R Commander.

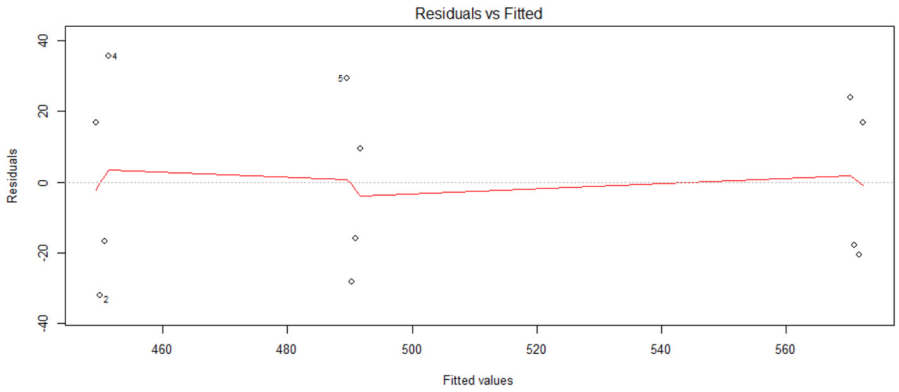
En la gràfica de la figura 36 podem comprovar que els residus segueixen una distribució aproximadament normal, ja que els punts s'apropen bastant a una recta.

Figura 36. Gràfica de probabilitat normal



La figura 37 presenta el gràfic dels valors residuals davant els valors estimats. Els residus es distribueixen aleatòriament, sense presentar cap tipus d'estructura, i podem concloure que és vàlid el model lineal múltiple. També observem en aquest gràfic que les variàncies dels residus són constants. El procediment i la interpretació dels supòsits es van explicar en l'apartat 3.1 ("Models de regressió lineal simple") i són iguals als corresponents de regressió múltiple.

Figura 37. Gràfica dels residus en funció dels valors estimats



Proves de significació

Les proves de significació que emprem en la regressió lineal van ser una prova *t* i una prova *F*. En aquest cas, ambdues proves donen com a resultat la mateixa conclusió: si es rebutja la hipòtesi nul·la la conclusió és que $\beta_1 \neq 0$. En la regressió múltiple la prova *t* i *F* té diverses finalitats.

La prova *F* s'usa per a determinar si hi ha una relació significativa entre la variable dependent i el conjunt de totes les variables independents. En aquestes condicions se l'anomena **prova de significació global.**

La prova *t* s'aplica per a determinar si cada una de les variables independents té significat. Es fa una prova *t* separadament per a cada variable independent en el model. Cada una d'aquestes proves rep el nom de **prova de significació individual.**

Prova *F* o anàlisi de la variància en regressió lineal

Les hipòtesis per a la prova *F* impliquen els paràmetres del model de regressió múltiple:

Hipòtesi nul□la: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Hipòtesi alternativa: H_1: un o més dels paràmetres no és igual a zero (almenys un paràmetre són $\neq 0$). Hem de fixar el nivell de significació α.

Si es rebutja H_0 tindrem prou evidència estadística per a concloure que un paràmetre o més no són igual a zero, i que la relació general entre y i el conjunt de variables independents x_1, x_2, …, x_k és significativa. Tanmateix, si no podem rebutjar H_0, no tenim l'evidència suficient per a arribar a la conclusió que la relació és significativa.

Per a fer el contrast hem de calcular l'estadístic de contrast *F*. **L'estadístic *F* és una variable aleatòria que es comporta segons una distribució *F* de Snedecor amb *k* graus de llibertat en el numerador (*DF-Regressió*) i *n* – *k* – 1 graus de llibertat en el denominador (*DF-Error* o *Residuals*).** On *k* són els graus de llibertat de la regressió i són iguals a la quantitat de variables independents, i *n* és el nombre d'observacions. Així, doncs, l'estadístic de contrast és:

$$F^* = \frac{SSR/k}{SSE/n-k-1}$$

Quadrat mitjà

És la suma de quadrats dividida pels graus de llibertat (DF) corresponents. Aquesta quantitat es fa servir a la prova *F* per a determinar si hi ha diferències significatives entre mitjanes.

També podem definir l'estadístic de contrast com el quocient de quadrats mitjà (*mean squares*).

El *quadrat mitjà a causa de la regressió* o simplement *regressió del quadrat mitjà* es representa per ***MSR*** (*mean square regression*):

$$MSR = \frac{SSR}{\text{graus de llibertat de la regressió}} = \frac{SSR}{k}$$

El quadrat mitjà a causa dels errors o residus s'anomena *quadrat mitjà residual* o *quadrat mitjà de l'error* i es representa per ***MSE*** (*mean square residual error*):

$$MSE = \frac{SSE}{\text{graus de llibertat de l'error}} = \frac{SSR}{n-k-1}$$

El valor de l'estadístic de contrast *F* podem definir-lo com a: $F^* = \frac{MSR}{MSE}$

Regla de decisió del contrast d'hipòtesi

Podem actuar de dues maneres:

a) A partir del *p*-valor. Aquest valor és: *p*-valor = $P(F_{\alpha;\, k,n-k-1} > F^*)$. On F_α és un valor de la distribució *F* amb *k* graus de llibertat en el numerador i *n* – *k* – 1 graus de llibertat en el denominador.

- Si *p*-valor < α es rebutja la hipòtesi nul□la H_0; es rebutja la hipòtesi nul□la H_0; per tant, el model en conjunt explica de manera significativa la variable *Y*. És a dir, el model sí que contribueix amb informació a explicar la variable *Y*.

- Si *p*-valor ≥ α no es rebutja la hipòtesi nul□la H_0; per tant, no hi ha una relació significativa. El model en conjunt no explica de manera significativa la variable *Y*.

b) A partir dels valors crítics

- Si $F^* > F_{\alpha\,;\,k,n-k-1}$, es rebutja la hipòtesi nul·la H_0
- Si $F^* < F_{\alpha\,;\,k,n-k-1}$, no es rebutja la hipòtesi nul·la H_0

Taula d'anàlisi de variància

A la primera columna es posa la **font de variació**, els elements del model responsables de la variació. A la segona columna posem la **suma de quadrats** corresponents.

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

A la quarta columna amb el nombre de **quadrats mitjans** es posen les sumes de quadrats dividides pels graus de llibertat corresponents. Només per a *SSR* i *SSE*.

A la cinquena columna posem l'estadístic de contrast ***F***.

Podem resumir els càlculs necessaris a la taula 7, coneguda com a taula d'anàlisi de la variància:

Taula 7. Anàlisi de variància per a un model de regressió múltiple amb *k* variables independents

Font de variació	Suma de quadrats	Graus de llibertat	Quadrats mitjans	*F*
Regressió	SSR	k	$MSR = SSR/k$	$F = \frac{MSR}{MSE}$
Error	SSE	$n - k - 1$	$MSE = SSE/n - k - 1$	
Total	SST	$n - 1$		

Aplicarem la prova *F* a l'exemple 3. Amb dues variables independents *nombre de pàgines per minut* ($X1$) i *antiguitat de la impressora* ($X2$).

Les hipòtesis es formulen de la manera següent:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: \beta_1 \text{ o } \beta_2 \text{ no és igual a zero}$$

Fixem un nivell de significació del 5% ($\alpha = 0{,}05$).

La figura 38 mostra els resultats del model de regressió múltiple, a la part de resultats corresponent a l'anàlisi de variància.

Figura 38. Resultats obtinguts amb R Commander

```
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 30348.6 30348.6 42.3668 0.0001103 ***
X2         1     7.3     7.3  0.0103 0.9215372
Residuals  9  6447.0   716.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor de l'estadístic de contrast és F^* per a la variable $X1$ és 42,3668 i per a la variable $X2$ és 0,0103. El *p*-valor per a la variable $X1$ és 0,0001103 i per a la variable $X2$ és 0,9215372. Per tant, la variable $X1$ explica de manera significativa la variable Y. Mentre que la variable $X2$ no explica de manera significativa la variable Y. En altres paraules, el nombre de pàgines per minut influeix sobre el preu d'impressores. En canvi, l'antiguitat de la impressora no influeix sobre el preu.

Prova *t*

S'utilitza per a determinar el significat de cada un dels paràmetres individuals. Les hipòtesis per a la prova *t* impliquen els paràmetres del model de regressió múltiple, es realitza un contrast per a cada paràmetre β:

Hipòtesi nul□la: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Hipòtesi alternativa: H_1: un paràmetre o més no són igual a zero (almenys un paràmetre és $\neq 0$). Hem de fixar el nivell de significació α.

L'estadístic de contrast és:

$$t^* = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Segueix una distribució *t* de Student amb $n - k - 1$ graus de llibertat.

Regla de decisió del contrast d'hipòtesi

Podem actuar de dues maneres:

a) A partir del *p*-valor. Aquest valor és: $p = 2P(t_{n-k-1} > | t^*|)$.

- Si $p < \alpha$ es rebutja la hipòtesi nul□la H_0; es rebutja la hipòtesi nul□la H_0; per tant, hi ha una relació lineal entre la variable X_i i *Y*. Per tant, l'esmentada variable ha de romandre en el model.

- Si $p \geq \alpha$ no es rebutja la hipòtesi nul□la H_0; per tant, no hi ha una relació lineal entre la corresponent variable X_i i *Y*. Diem que la variable implicada X_i és no explicativa i podem eliminar-la del model.

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 és significativa.

- Si $|t^*| > t\alpha/2, n - k - 1$, no es rebutja la hipòtesi nul□la H_0; per tant, la variable no és significativa. Diem que la variable implicada X_i no és explicativa.

Si la prova *F* de l'exemple (figura 38) ha mostrat que la relació múltiple té significat, es pot fer una prova *t* per a determinar el significat de cada un

dels paràmetres individuals. El nivell de significació és $\alpha = 0{,}05$. Observeu que els valors dels estadístics *t* apareixen a la figura 39, i els *p*-valors dels contrastos individuals són per al contrast de β_1 el *p*-valor = 0,000 i per a β_2 *p*-valor = 0,92154.

Figura 39. Resultats obtinguts amb R Commander

```
Call:
lm(formula = Y ~ X1 + X2, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-31.971 -18.611  -3.343  18.488  35.629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  330.407     29.795  11.089  1.5e-06 ***
X1            20.161      3.097   6.509  0.00011 ***
X2            -0.350      3.455  -0.101  0.92154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.76 on 9 degrees of freedom
Multiple R-squared:  0.8248, Adjusted R-squared:  0.7859
F-statistic: 21.19 on 2 and 9 DF,  p-value: 0.0003941
```

Interpretem el contrast per al paràmetre β_1 la H_0: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$. Com que 0,000 < 0,05 es rebutja H_0; per tant, la variable *X*1 (nombre de pàgines impreses per minut) és significativa.

El contrast per al paràmetre β_2, la H_0: $\beta_2 = 0$, H_1: $\beta_2 \neq 0$. Com que 0,92154 > 0,05 no podem rebutjar H_0; per tant, la variable *X*2 (antiguitat) no és significativa i podríem eliminar-la del model perquè no influeix significativament en el preu.

Com que *p*-valor <0,05 el model **en conjunt** explica de manera significativa la variable *Y* (el R2 es elevat 82,48%). No obstant, seria recomanable realitzar el model sense la variable *X2*.

El problema de la multicolinealitat

En els problemes de regressió lineal múltiple esperem trobar dependència entre la variable *Y* i les variables explicatives *X*1, *X*2, ..., *Xk*. Però, en alguns problemes de regressió podem tenir també algun tipus de dependència entre algunes de les variables *Xj*. En aquest cas tenim informació redundant en el model. Aquest fenomen es diu **multicolinealitat**, i sol ser bastant freqüent en els models de regressió lineal múltiple.

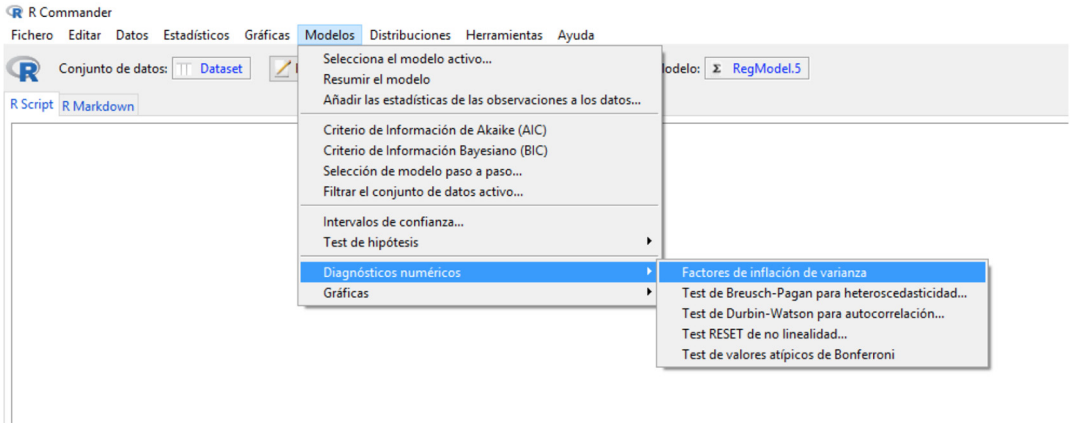
El terme **multicolinealitat**, en anàlisi de regressió múltiple, indica la correlació entre variables independents. La multicolinealitat pot tenir efectes molt importants en les estimacions dels coeficients de la regressió i, per

tant, sobre les posteriors aplicacions del model estimat. Quan les variables independents estan molt correlacionades no és possible determinar l'efecte separadament d'una d'elles sobre la variable dependent. Quan hi ha multicolinealitat, els resultats dels contrastos d'hipòtesi sobre el model conjunt i els resultats dels contrastos individuals són aparentment contradictoris, però que realment no ho són. Aquest efecte el veurem en l'exemple proposat (figura 40). El R Commander disposa d'una opció, per calcular la ***Variance Inflation Factors*** (VIF), la qual ens permet identificar la multicolinealitat entre els predictors del model. La figura 40 indica els passos a seguir.

Passos a seguir

Se segueix la ruta *Modelos > Diagnósticos numéricos > Factores de inflación de varianza.*

Figura 40. Passos a seguir per a identificar la multicolinealitat



Ara, la figura 41 dels resultats de l'anàlisi de regressió múltiple conté els valors VIF. Cada coeficient VIF és d'1,000. Aquests valors són baixos, la qual cosa indica que les variables independents no estan correlacionades. Ja que aquests valors indiquen que el grau de colinearitat és baix. No hi ha multicolienalitat en el model proposat.

Figura 41. Resultats de l'anàlisi de regressió múltiple de l'exemple 3, inclou *Variance Inflation Factors* (VIF)

```
> vif(RegModel.5)
X1 X2
 1  1
```

Utilitzant el **Microsoft Excel** per a obtenir l'anàlisi de regressió de l'exemple 3. "Estudi sobre el preu d'impressores làser en funció de la seva velocitat d'impressió i l'antiguitat del model".

Passos a seguir

Per a efectuar la regressió múltiple amb l'**MS Excel**, i una vegada introduïdes les dades en el full de càlcul, se segueix la ruta següent: clic a *Herramientas > Análisis de datos > Regresión > OK.*

A continuació, seleccioneu els rangs de dades de les variables.

La taula 8 mostra la corresponent sortida que ofereix el **Microsoft Excel**.

Taula 8. Resultats de l'anàlisi de regressió de l'exemple 3. "Estudi sobre el preu d'impressores làser en funció de la seva velocitat d'impressió i l'antiguitat del model". Excel

	B	C	D	E	F	G	H
6	Resum						
7							
8	*Estadístiques de la regressió*						
9	Coeficient de correlació múltiple	0,910524728					
10	Coeficient de determinació	0,829055281					
11	R^2 ajustat	0,791067565					
12	Error típic	26,40996235					
13	Observacions	12					
14							
15	Anàlisi de variància						
16		*Graus de llibertat*	*Suma de quadrats*	*Mitjana dels quadrats*	*F*	*Valor crític de F*	
17	Regressió	2	30444,29167	15222,14583	21,8242996	0,000353062	
18	Residus	9	6277,375	697,4861111			
19	Total	11	36721,66667				
20							
21		*Coeficients*	*Error típic*	*Estadístic t*	*Probabilitat*	*Inferior 95%*	*Superior 95%*
22	Intercepció	330,375	29,40041791	11,23708517	1,3446E-06	263,8666342	396,883366
23	X1	20,1875	3,056359247	6,605080872	9,8697E-05	13,27353505	27,1014649
24	X2	-0,5	3,409511478	-0,146648575	0,88664178	-8,212850796	7,2128508
25							

Resum

En aquest mòdul hem introduït conceptes de relacions funcionals i estadístiques, i també el de variables dependents i el de variables independents. Hem comentat la construcció d'un diagrama de dispersió com a pas inicial a l'hora de buscar algun tipus de relació entre dues variables. Si el diagrama mostra una estructura lineal, aleshores es buscarà la recta que millor s'ajusta a les observacions. Hem posat de manifest la importància d'interpretar correctament els coeficients de la recta. També hem vist com s'ha de fer servir la recta de regressió per a fer prediccions. Hem introduït una mesura numèrica de la bondat d'ajust. Aquesta mesura s'obté amb el coeficient de determinació, discutint els valors que pot prendre. Finalment, hem comentat la importància d'analitzar els residus per a fer un diagnòstic del model lineal obtingut.

En aquest mòdul de regressió lineal simple s'ha considerat que les observacions sobre dues variables X i Y són una mostra aleatòria d'una població i que es fan servir per a extraure algunes conclusions del comportament de les variables sobre la població, i per això hem vist com inferir sobre el pendent de la recta obtinguda a partir de la mostra i com fer un contrast d'hipòtesi per a decidir si la variable X explica realment el comportament de la variable Y. També hem comentat algunes relacions no lineals i la manera com es pot transformar en una de lineal.

Hem tractat la regressió lineal múltiple com una generalització del model de regressió lineal simple en els casos en què hi ha més d'una variable explicativa. Finalment, hem vist com inferir sobre els coeficients de regressió obtinguts a partir de la mostra, com fer un contrast d'hipòtesi per a cada un dels coeficients obtinguts per a decidir si les variables independents expliquen realment el comportament de la variable dependent o es pot prescindir d'alguna. També hem dut a terme un contrast conjunt del model. Finalment, hem presentat el possible problema de multicolinealitat que hi pot haver i que és degut a la relació entre algunes de les variables explicatives que suposadament són independents.

Exercicis d'autoavaluació

1) Els preus d'una pantalla TFT d'una coneguda marca són els següents:

Mida (polzades)	15	17	19	24
Preu (euros)	251	301	357	556

Calculeu la recta de regressió per a explicar el preu a partir de la mida.

2) Amb les dades de la qüestió anterior volem decidir si es tracta d'un bon model. Quin mètode proposeu per a determinar si s'ajusta bé? Què podem dir del cas concret de l'exemple anterior?

3) Considerem un model lineal per a explicar el rendiment d'un sistema informàtic (variable Y) en relació amb el nombre de memòries intermèdies i el nombre de processadors (variables $X1$ i $X2$ respectivament). S'obté el model $Y = -3{,}20 + 2X1 + 0{,}0845X2$ amb un coeficient de determinació de 0,99. Es tracta d'un bon model? Quin serà el rendiment esperat si tenim 1 memòria intermèdia i 1 processador? Comenteu si aquest valor us sembla lògic i si es pot relacionar amb la bondat del model.

4) L'empresa Ibérica editors ha de decidir si firma o no un contracte de manteniment per al seu nou sistema de processament de paraules. Els directius creuen que la despesa de manteniment ha d'estar relacionada amb l'ús i han reunit la informació que veiem a la taula següent sobre l'ús setmanal, en hores, i la despesa anual de manteniment (centenar d'euros).

Ús setmanal (hores)	**Despeses anuals de manteniment**
13	17,0
10	22,0
20	30,0
28	37,0
32	47,0
17	30,5
24	32,5
31	39,0
40	51,5
38	40,0

a) Determineu l'equació de regressió que relaciona el cost anual de manteniment amb l'ús setmanal.
b) Proveu el significat de la relació obtinguda en l'apartat *a* al nivell de significació 0,05.
c) Ibèrica editors espera utilitzar 30 hores setmanals el processador de paraules. Determineu un interval de predicció del 95% per a la despesa de l'empresa en manteniment anual.
d) Si el contracte de manteniment costa 3.000 euros anuals, recomanaríeu firmar-lo? Per què?

5) Una biblioteca pública d'una ciutat espanyola ofereix un servei via Internet de préstecs de llibres als seus usuaris. Es vol estudiar la correlació entre el nombre d'usuaris d'aquesta biblioteca virtual i quants d'ells acaben realitzant els préstecs.

Les dades dels últims 12 mesos són:

Usuaris	296	459	602	798	915	521	362	658	741	892	936	747
Préstecs	155	275	322	582	761	324	221	415	562	628	753	569

a) Determineu el coeficient de correlació entre les dues variables. Calculeu i representeu la recta de regressió.
b) Quin nombre de préstecs s'esperaria si el nombre d'usuaris augmentés 1.000?

6) Un expert documentalista necessita saber si l'eficiència d'un nou programa de recerca bibliogràfica depèn del volum de les dades entrants. L'eficiència es mesura amb el nombre de peticions per hora processades. Aplicant el programa en diferents volums de dades, obtenim els resultats següents:

Volum (gigabytes), *X*	6	7	7	8	10	10	15
Peticions processades, *Y*	40	55	50	41	17	26	16

a) Calculeu la recta de regressió per a explicar les peticions processades per hora a partir del volum de dades i interpreteu els paràmetres obtinguts.
b) Realitzeu el gràfic d'ajust a la recta de mínims quadrats.
c) Determineu el coeficient de correlació lineal entre les dues variables i interpreteu-ne el significat.
d) Determineu el coeficient de determinació R^2 i interpreteu-ne el significat.
e) Calculeu, a partir de la recta anterior, quantes peticions podem esperar per a un volum de dades de 12 gigabytes.
f) Realitzeu el contrast d'hipòtesi sobre el pendent. Podem afirmar a un nivell de significació de 0,05 que el pendent de la recta és zero?

Solucionari

1) Preu = –279,11 + 34,42 · mida.

2) Per a estudiar la qualitat de l'ajust, es calcula el coeficient de correlació mostral $r = 0{,}994$.

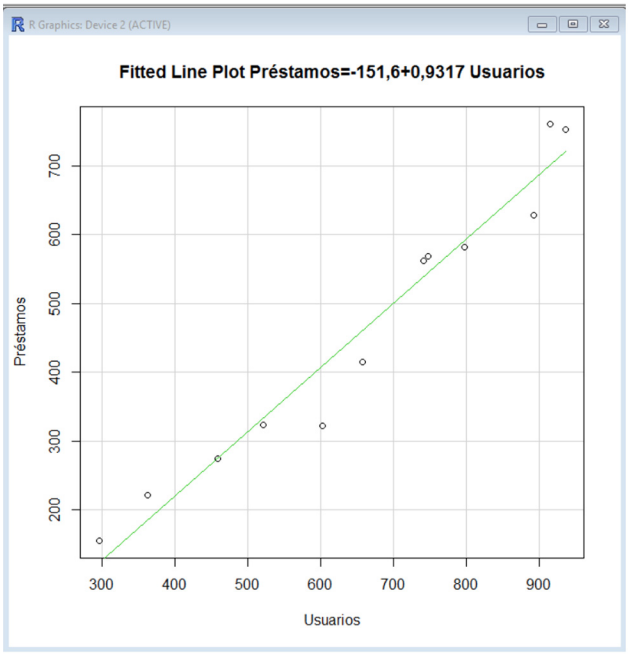
3) És un bon model, ja que el coeficient de determinació és molt proper a 1. El rendiment si tenim una memòria intermèdia i un processador seria: $Y = -3{,}20 + 2 \cdot 1 + 0{,}0845 \cdot 1 = -1{,}1155$. Aquest valor no té sentit, ja que el rendiment no pot ser negatiu. De totes maneres, aquest fet no és contradictori amb el fet de tenir un bon model, ja que som fora de l'interval on la regressió funciona.

4)
a) $\hat{y} = 10{,}5 + 0{,}953x$.
b) Relació significativa; p-valor = 0,000.
c) $[2.874;54.952]$ euros.
d) Sí, la probabilitat de trobar la despesa de manteniment dins de l'interval de confiança és del 95%.

5)
a) $r = 0{,}9775$.

```
        Pearson's product-moment correlation

data:  Préstamos and Usuarios
t = 14.685, df = 10, p-value = 4.287e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9196495 0.9938831
sample estimates:
      cor
0.9775904
```

```
Call:
lm(formula = Préstamos ~ Usuarios, data = Dataset)

Residuals:
   Min     1Q Median     3Q    Max
-87.33 -19.09  11.03  31.20  60.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -151.55168   43.91506  -3.451  0.00622 **
Usuarios       0.93170    0.06345  14.685 4.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.42 on 10 degrees of freedom
Multiple R-squared:  0.9557, Adjusted R-squared:  0.9513
F-statistic: 215.6 on 1 and 10 DF,  p-value: 4.287e-08
```



b) –151,6 + 0,9317 × 1.000 ≃ 780 préstecs

6)
a)

```
Call:
lm(formula = Peticiones.Y ~ Volumen.X, data = Dataset)

Residuals:
      1       2       3       4       5       6       7
 -7.429  11.714   6.714   1.857 -13.857  -4.857   5.857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.286     12.491   5.787  0.00217 **
Volumen.X     -4.143      1.324  -3.129  0.02599 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.908 on 5 degrees of freedom
Multiple R-squared:  0.6619, Adjusted R-squared:  0.5943
F-statistic:  9.79 on 1 and 5 DF,  p-value: 0.02599
```

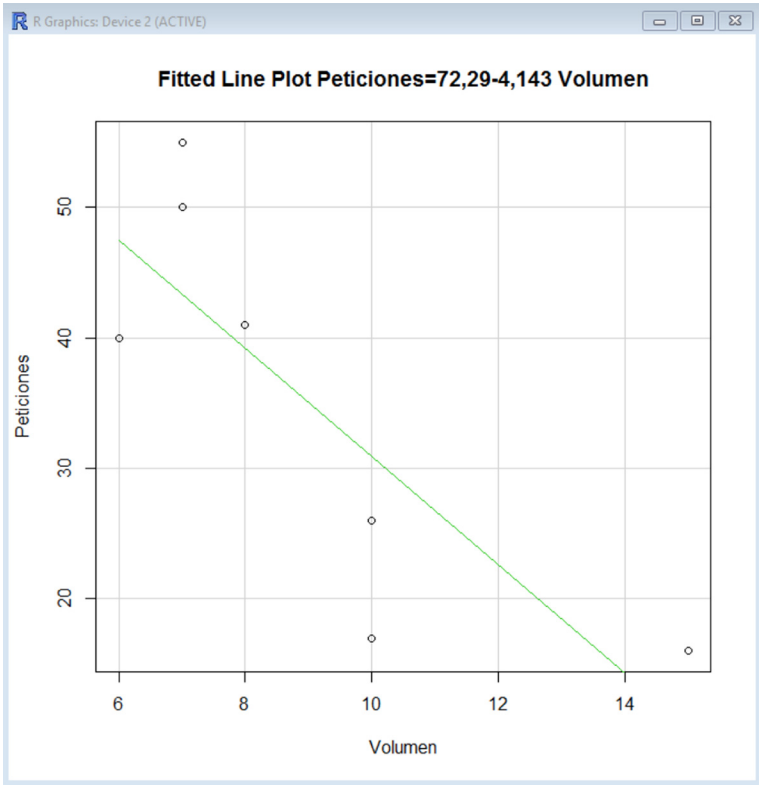
La recta de regressió serà:

Peticions processades = 72,29 – 4,143 volum (gigabytes).

L'ordenada en l'origen: 72,29, en aquest cas el seu significat no té cap sentit.

El pendent de la recta: –4,143; és negatiu: indica que per cada unitat de volum de dades (gigabytes) que augmenten les dades entrants el nombre de peticions processades disminueix en 4,143 unitats.

b) El gràfic d'ajust a la recta de mínims quadrats és:



c)

```
        Pearson's product-moment correlation

data:  Peticiones.Y and Volumen.X
t = -3.129, df = 5, p-value = 0.02599
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9714575 -0.1563115
sample estimates:
       cor
-0.8135995
```

El coeficient de correlació $r = -0{,}814$ ens indica que hi ha una correlació alta negativa entre el volum de dades entrants i el nombre de peticions processades.

d) El coeficient de determinació *R-Sq* és el 66,19%. Això vol dir que el nostre model lineal explica el 66,19% del comportament de la variable *Y* (en aquest cas, nombre de peticions processades).

e) Amb 12 gigabytes, hi haurà $72{,}3 - 4{,}14 \cdot 12 = 22{,}57$ peticions.

f) A la sortida anterior podem veure que el *p*-valor associat al contrast d'hipòtesis anterior és 0,026. Com que aquest valor és menor que $\alpha = 0{,}05$, hem de rebutjar la hipòtesi nul□la, és a dir, podem concloure que el pendent de la recta és diferent de zero o, dit d'una altra manera, que el coeficient de correlació poblacional és no nul (és a dir, que les dues variables estan correlacionades i que, per tant, el model estudiat té sentit).