Regressió lineal simple

Josep Gibergans Bàguena

P08/05057/02311

Universitat Oberta de Catalunya
www.uoc.edu

Índex

El model de regressió simple

1. Introducció

Després d'estudiar com s'ha d'organitzar, representar gràficament i analitzar un conjunt de dades a partir d'alguns paràmetres, ens proposem estudiar les relacions entre variables.

Origen dels models de regressió

Aquests models van ser utilitzats per Laplace i Gauss en els seus treballs d'astronomia i física desenvolupats durant el segle XVIII, però el nom de *models de regressió* té el seu origen en els treballs de Galton en biologia de final del segle XIX. L'expressió de Galton:

"regression towards mediocrity"

donà nom a la regressió.

Per exemple, podem estudiar les distribucions dels pesos i de les alçades d'un conjunt de persones per separat. Ara l'objectiu és determinar si hi ha alguna relació entre aquestes variables.

Volem construir models que descriguin la relació entre les variables amb el propòsit, principalment, de predir els valors d'una variable a partir dels valors de l'altra. Això ho farem amb el model de regressió lineal simple.

2. Relacions entre dues variables

El model de regressió lineal simple ens permet de construir un model per a explicar la relació entre dues variables.

L'objectiu és explicar el comportament d'una variable, Y, que denominarem **variable explicada** (o **dependent** o **endògena**), a partir d'una altra variable X, que anomenarem **variable explicativa** (o **independent** o **exògena**).

Exemple de relació entre dues variables

Si les dues variables són els ingressos mensuals i les despeses en activitats d'oci, aleshores podríem triar la segona com a variable explicada Y i la primera com a variable explicativa X, ja que, en principi, les despeses en oci dependran molt dels ingressos: com més diners guanyem, més gran és la part que gastem en oci.

És important observar que també podríem escollir les variables a l'inrevés, és a dir, les despeses en oci com a variable explicativa X i els ingressos com a variable explicada Y. Com més diners gastem en oci, voldrà dir que tenim més ingressos.

No és fàcil la decisió de triar quina és la variable explicativa i quina és la variable explicada. Com veurem més endavant, dependrà molt de les característiques de les dades que tindrem.

Les relacions entre dues variables poden ser de dos tipus:

Exemple de relació funcional

Podem conèixer l'àrea d'un quadrat a partir de la longitud del seu costat.

1) Funcionals (o **deterministes**): quan hi ha una fórmula matemàtica que permet de calcular els valors d'una de les variables a partir dels valors que pren l'altra.

2) Estadístiques (o **estocàstiques**): quan no existeix una expressió matemàtica que les relacioni de forma exacta.

Exemple de relació estadística

Sabem que hi ha una relació entre l'alçada i el pes de les persones: en general, és a dir, com més alçada, més pes. Però no hi ha cap fórmula matemàtica que ens en doni una en funció de l'altra, ja que, si no, totes les persones que tenen la mateixa alçada tindrien el mateix pes i això sabem que no és cert.

En la relació entre el pes i l'alçada és evident que hi ha molts factors, com poden ser factors genètics, l'activitat física, l'alimentació, etc. que fan que una persona d'una determinada alçada tingui un pes o un altre. Tots aquests factors i d'altres que no coneixem fan que la relació entre aquestes dues variables sigui estadística i no funcional.

3. Diagrames de dispersió i corbes de regressió

A partir d'un conjunt d'observacions de dues variables X i Y sobre una mostra d'individus, el primer pas en una anàlisi de regressió és representar aquestes dades sobre uns eixos coordenats *x-y*. Aquesta representació és l'anomenat *diagrama de dispersió*. Ens pot ajudar molt en la cerca d'un model que descrigui la relació entre totes dues variables.

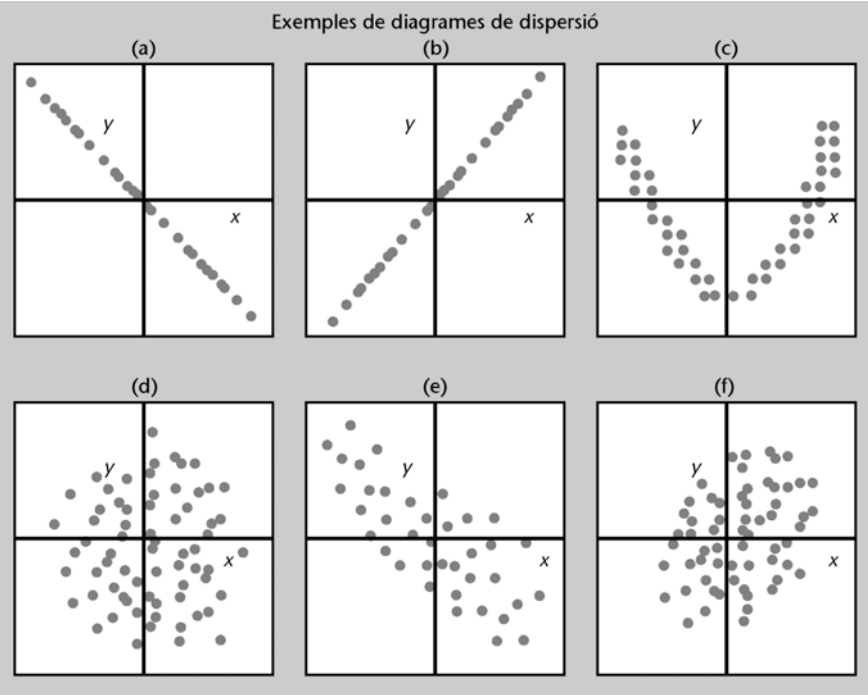
El **diagrama de dispersió** l'obtenim representant cada observació (x_i, y_i) com un punt en el pla cartesià XY.

Terminologia

El diagrama de dispersió també es coneix com a *núvol de punts.*

Exemple de diagrames de dispersió

El diagrama de dispersió pot presentar formes diverses:



En els casos (a) i (b) tenim que les observacions es troben sobre una recta. En el primer cas, amb pendent negatiu, que ens indica que a mesura que X es fa gran la Y és cada ve-

gada més petita i el contrari en el segon cas, en què el pendent és positiu. 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 es troben situats en una franja prou 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).

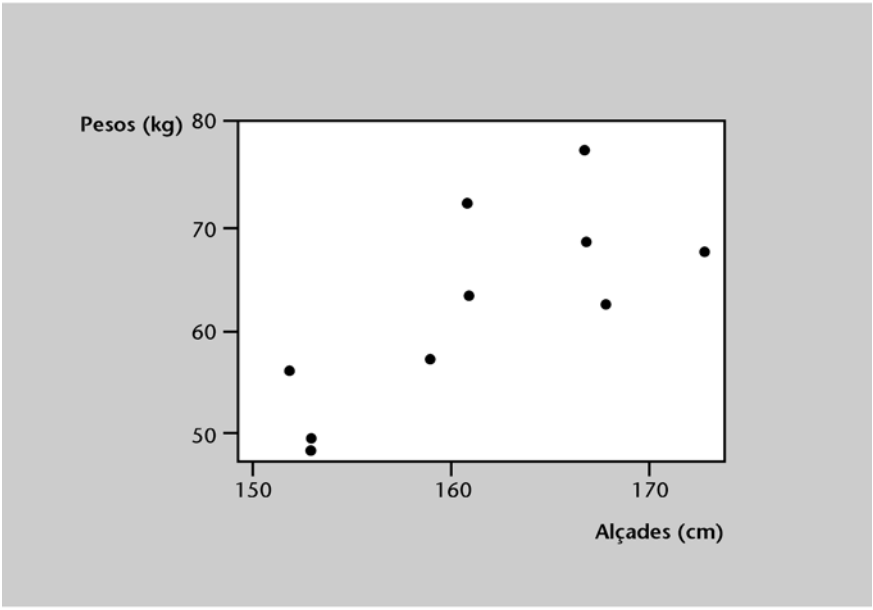
En el cas (d) no tenim cap tipus de relació entre les variables. El núvol de punts no presenta una forma "tubular" ben determinada; els punts es troben absolutament dispersos.

En els casos (e) i (f) podem observar que sí que existeix algun tipus de relació entre totes dues variables. En el cas (e) podem veure una mena 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'hi acosten bastant, de manera que podem pensar en una forta relació lineal. En el cas (f), observem una relació lineal amb pendent positiu, però no tan forta com l'anterior.

Exemple de les alçades i els pesos

Considerem les observacions dels pesos i alçades d'un conjunt de 10 persones: l'individu 1 té 161 cm d'alçada i 63 kg de pes, l'individu 2 té 152 cm d'alçada i 56 kg de pes, etc., tal com es veu en la taula següent:

Individu	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
***X* Alçada (cm)**	161	152	167	153	161	168	167	153	159	173
***Y* Pes (kg)**	63	56	77	49	72	62	68	48	57	67



El diagrama de dispersió també ens pot ajudar a trobar algun valor atípic entre les dades de la mostra que pugui tenir l'origen en una mala observació o en el fet de ser una observació corresponent a un individu excepcional dintre la mostra. Quan tenim un valor atípic, hem de controlar les influències que pugui tenir en l'anàlisi.

Definició i exemple de valor atípic

Per *valor atípic* entenem un valor molt diferent dels altres i que molt possiblement és erroni. Per exemple, una persona de 150 cm d'alçada i 150 kg de pes. En el diagrama de dispersió sortirà com un punt solitari allunyat dels altres.

4. Recta de regressió

Una vegada hem fet el diagrama de dispersió i després d'observar una possible relació lineal entre les dues variables, ens proposem trobar l'equació de la recta que millor s'ajusti al núvol de punts. Aquesta recta s'anomena **recta de regressió**.

4.1. Estimació dels paràmetres: Mètode dels mínims quadrats

Una recta queda ben determinada si el valor del seu **pendent** (b) i de la seva **ordenada a l'origen** (a) són coneguts. D'aquesta manera, l'equació de la recta ve donada per:

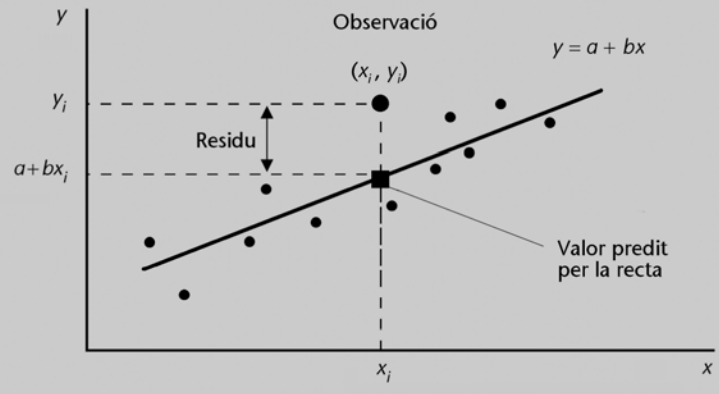
$$y = a + bx$$

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

$$y_i - (a + bx_i)$$

Per cada recta que considerem, tindrem una col·lecció diferent de residus. Cercarem la recta que doni lloc als residus més petits quant a la suma dels quadrats.

Interpretació geomètrica del residu



Per a determinar una recta de regressió, utilitzarem el mètode dels mínims quadrats.

El **mètode dels mínims quadrats** consisteix a cercar els valors dels paràmetres *a* i *b* de manera que la suma dels quadrats dels residus sigui mínima. Aquesta recta és la **recta de regressió per mínims quadrats**.

Essent la suma dels quadrats l'expressió:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

per a trobar els valors de *a* i *b*, només cal determinar les derivades parcials respecte dels paràmetres *a* i *b*:

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n} (y_i - a - bx_i)^2\right) = -2\sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\frac{\partial}{\partial b}\left(\sum_{i=1}^{n} (y_i - a - bx_i)^2\right) = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i$$

i les igualem a zero. Així obtenim el sistema d'equacions següent, conegut com a *sistema d'equacions normals*:

$$\begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 \end{cases}$$

Les solucions d'aquest sistema d'equacions són:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{i} \quad a = \bar{y} - b\bar{x}$$

en què:

- $s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ és la **covariància mostral** de les observacions (x_i, y_i)
- $s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ és la **variància mostral** de les observacions x_i

És molt important obsevar que, de totes les rectes, la recta de regressió lineal per mínims quadrats és aquella que fa mínima la suma dels quadrats dels residus.

Terminologia

La suma dels quadrats dels residus també s'anomena *suma dels errors quadràtics*.

La resolució d'aquests sistema d'equacions es troba a l'annex 1. d'aquesta sessió.

En rigor...

... caldria provar que, efectivament, aquests valors dels paràmetres fan mínima la suma dels quadrats dels residus.

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

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

en què els **paràmetres de la recta** $\hat{\beta}_0$ i $\hat{\beta}_1$ vénen donats per:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{i} \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

D'ara endavant, als **residus calculats** amb la recta de regressió els anomenarem e_i, és a dir:

$$e_i = y_i - \hat{y}_i$$

en què $\hat{y}_i$ és el **valor estimat** per la recta de regressió.

Notació

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:

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

obtinguda en determinar a i b.

5. Interpretació dels paràmetres de la recta de regressió

Un cop determinada la recta de regressió, és molt important interpretar els paràmetres de l'equació en el context del fenomen que s'estudia.

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

Aquest paràmetre representa l'estimació del valor de Y quan X és igual a zero:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 0 = \hat{\beta}_0$$

No sempre té una interpretació pràctica. Perquè sigui possible, cal que:

1. Sigui realment possible que X prengui el valor $x = 0$

2. Es tinguin suficients observacions properes al valor $x = 0$

$\hat{\beta}_0$ en l'exemple dels pesos i les alçades

En l'exemple dels pesos i les alçades, el valor de l'ordenada a l'origen no tindrà sentit, ja que correspondria al pes que tindrien les persones d'alçada nul·la.

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

Aquest paràmetre representa l'estimació de l'increment que experimenta la variable Y quan X augmenta en una unitat. Aquest paràmetre ens informa de com estan relacionades les dues variables en el sentit que ens diu en quina quantitat (i si és positiva o negativa) varien els valors de Y quan varien els valors de la X en una unitat.

Pendent en l'exemple dels pesos i les alçades

En l'exemple dels pesos i les alçades havíem observat en el diagrama de dispersió que, en general, augmenta el pes de les persones a mesura que augmenta l'alçada.

6. Construcció de la taula per a determinar els paràmetres

Vegem ara com hem de determinar, a la pràctica, la recta de regressió. Ho il·lustrarem a partir de les dades de l'exemple dels pesos i les alçades.

Exemple de les alçades i els pesos

Continuem amb l'exemple de les alçades i pesos d'un grup de deu persones anterior. Per a determinar la recta de regressió, calculem la covariància mostral s_{xy}, la variància mostral s_x^2 i les mitjanes $\bar{x}$ i $\bar{y}$.

Podem calcular totes aquestes quantitats a partir de la taula de càlculs de la recta de regressió.

i	x_i	y_i	$\bar{x}_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	161	63	–0,4	1,1	0,16	–0,44
2	152	56	–9,4	–5,9	88,36	55,46
3	167	77	5,6	15,1	31,36	84,56
4	153	49	–8,4	–12,9	70,56	108,36
5	161	72	–0,4	10,1	0,16	–4,04
6	168	62	6,6	0,1	43,56	0,66
7	167	68	5,6	6,1	31,36	34,16
8	153	48	–8,4	–13,9	70,56	116,76
9	159	57	–2,4	–4,9	5,76	11,76
10	173	67	11,6	5,1	134,56	59,16
Σ	**1.614**	**619**			**476,40**	**466,40**

Mitjanes mostrals: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 161{,}4$ i $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 61{,}9$

Variància mostral: $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{476{,}40}{10-1} = 52{,}933$

Covariància mostral: $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{466{,}40}{10-1} = 51{,}822$

Els paràmetres són:

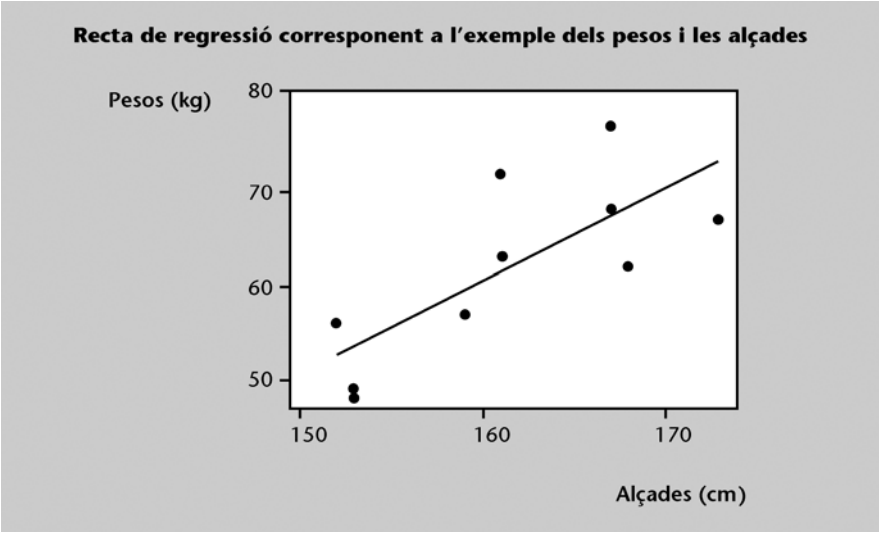
$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{51{,}822}{52{,}933} = 0{,}979009$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 61{,}9 - 0{,}979009 \cdot 161{,}4 = -96{,}1121$$

Tenim la recta de regressió següent:

$$\hat{y} = -96{,}1121 + 0{,}979009x$$

Podem representar la recta de regressió en el diagrama de dispersió:



Interpretem els paràmetres obtinguts:

- Ordenada a l'origen: evidentment, no té sentit pensar que el pes d'una persona d'alçada zero és –96,1121 kg. Ja hem comentat abans que moltes vegades no té sentit la interpretació d'aquest paràmetre.
- Pendent: tenim un pendent de 0,979009. Un valor positiu que ens informa que el pes augmenta amb l'alçada a raó de 0,979 kg per cada centímetre.

7. Interpolació i extrapolació

Un dels objectius més importants de la regressió és l'aplicació del model per al pronòstic del valor de la variable dependent (Y) per a un valor de la variable independent (X) no observat en la mostra.

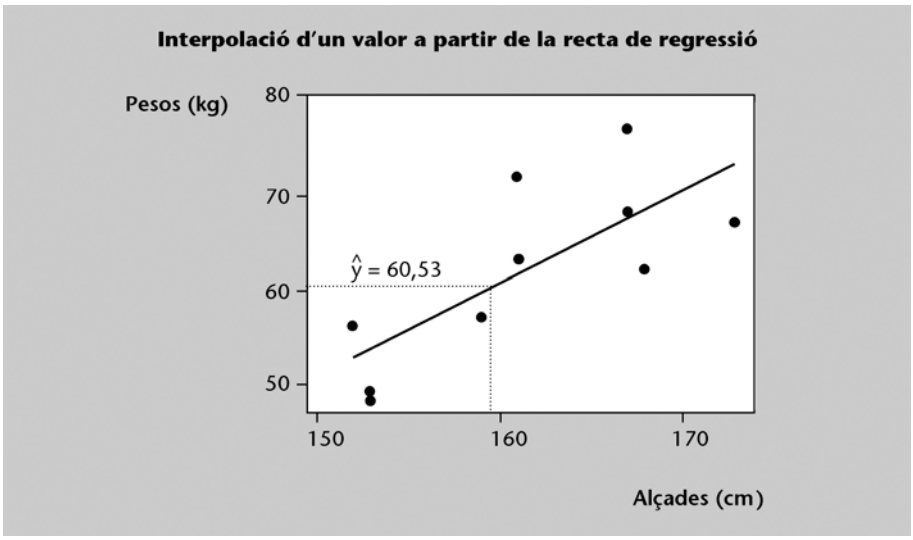
Exemple de les alçades i els pesos

En el nostre problema dels pesos i les alçades, podríem estar interessats a conèixer el pes d'una persona d'alçada 1,60 m. A partir de la nostra recta de regressió:

$$\hat{y} = -96{,}1121 + 0{,}979009x$$

Per a un valor de X de 160 cm, tenim un valor estimat per a la Y de 60,53 kg:

$$\hat{y} = -96{,}1121 + 0{,}979009 \cdot 160 = 60{,}53$$



Un aspecte important a l'hora d'aplicar el model de regressió obtingut és el **risc de l'extrapolació**. És a dir, quan volem conèixer el valor que presentarà la variable Y per a un determinat valor de X que es trobi fora de l'interval de valors que pren la mostra. Aleshores hem d'anar amb molt de compte:

1) Hem determinat el model amb la informació continguda en la mostra, de manera que no hem tingut cap informació del comportament de la variable Y per a valors de X de fora del rang de la mostra.

2) És possible que no tingui sentit l'extrapolació que volem fer. Abans de fer servir el model de regressió, ens hem de preguntar per allò que estem fent.

Extrapolació fora de rang

Si volem saber el pes d'un nadó que només mesura quaranta centímetres, no podrem fer servir la recta de regressió obtinguda. Les característiques biològiques del nadó, molt diferents a les de les persones adultes, faran que la relació entre el pes i l'alçada sigui diferent. Hauríem de fer una anàlisi de regressió a partir d'una mostra de nadons.

Sentit de l'extrapolació

No té cap sentit fer servir el model de regressió per a calcular el pes de persones de deu centímetres o tres metres d'alçada. El model ens donarà un resultat numèric, que, en tot cas, cal interpretar.

8. Models de regressió no lineals

A banda dels models lineals, se'n poden establir d'altres, entre els quals destaca l'exponencial.

Corba en un model exponencial

En el model lineal hem ajustat el núvol de punts a una recta d'equació:

$$y = a + bx$$

En el model exponencial volem ajustar als punts una corba d'equació:

$y = ka^x$ amb $a > 0$ i $k > 0$

El **model exponencial** és del tipus:

$$y = ka^x \text{ amb } a > 0,\ k > 0$$

on k i a són valors constants.

Així com en el cas lineal és molt fàcil veure si hi pot haver una relació lineal entre les variables a partir del diagrama de dispersió, en el cas exponencial és una mica més difícil.

Exemples de relacions exponencial

Les relacions entre la variable temps (X) i altres variables (Y) com la població, el nombre d'ordinadors infectats per un virus en els primer dies de contaminació, els preus d'alguns productes, etc., són exponencials.

Per a tractar-lo, linealitzem el problema, és a dir, transformem les variables de manera que el problema esdevingui lineal. Si a l'equació $y = ka^x$ prenem logaritmes $\ln y = \ln(ka^x)$ obtenim, per aplicació de les propietats dels logaritmes:

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

Propietats dels logaritmes

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

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

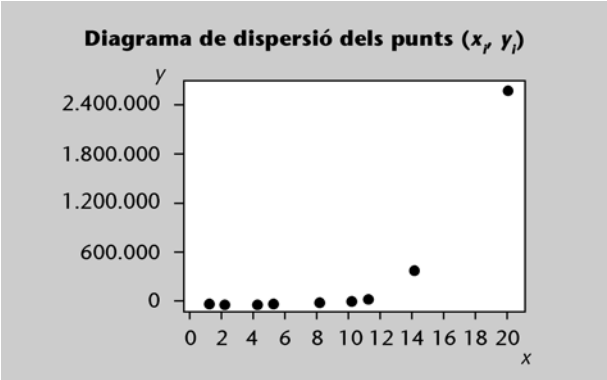
Aquesta última equació ens mostra un model lineal entre les variables X i $\ln Y$. Així, si representem el diagrama de dispersió dels punts $(x_i, \ln y_i)$ i el núvol de punts presenta una estructura lineal, podem pensar que entre les variables X i Y hi ha una relació exponencial.

Exemple de la propagació d'un virus informàtic

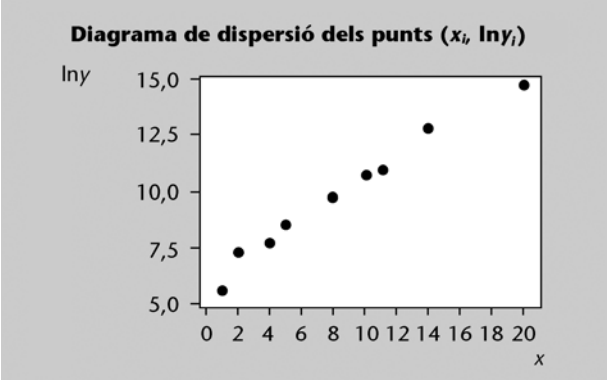
La taula registra el nombre de dies que han passat des que s'ha detectat un nou virus informàtic i el nombre d'ordinadors infectats en un país.

Nombre de dies x_i	Nombre d'ordinadors infectats y_i	Transformació de Y $\ln y_i$
1	255	5,5413
2	1.500	7,3132
4	2.105	7,6521
5	5.050	8,5271
8	16.300	9,6989
10	45.320	10,7215
11	58.570	10,9780
14	375.800	12,8368
16	1.525.640	14,2379
20	2.577.000	14,7621

El diagrama de dispersió dels punts següents ens fa pensar en l'existència d'alguna mena de relació entre les variables que no és lineal. Estudiarem si es tracta d'una relació exponencial.



Calculem el logaritme de les dades de la variable Y i representem el diagrama de dispersió corresponent.



Podem observar que entre les variables X i ln Y hi ha una relació lineal; per tant, entre les variables originals X i Y hi haurà una relació exponencial.

Si calculem la recta de regressió de ln y sobre x: $\ln\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Obtenim: $\ln\hat{y} = 5{,}84 + 0{,}482x$, és a dir, $\hat{y} = e^{5{,}84 + 0{,}482x}$.

De manera que, si volem estimar el nombre d'ordinadors infectats al cap de dotze dies, farem el següent:

Per a $x = 12$: $\ln\hat{y} = 5{,}84 + 0{,}482 \cdot 12 = 11{,}624$.

I prenent exponencials, podem aïllar $\hat{y}$:

$$\hat{y} = \exp(11{,}624) = 111.747{,}8195$$

Per tant, al cap de dotze dies el nombre estimat d'ordinadors infectats ha estat de 111.748 unitats.

9. Resum

En aquesta primera sessió hem introduït els conceptes de relacions funcionals i estadístiques i també els de variables dependents (o explicades) i el de variables independents (o explicatives). A continuació, s'ha comentat la construcció d'un diagrama de dispersió com a pas inicial a l'hora de cercar alguna mena de

relació entre dues variables. Si el diagrama ens mostra una estructura lineal, llavors busquem la línia recta que millor s'ajusta a les nostres observacions. Ho fem mitjançant el mètode dels mínims quadrats. Hem posat de manifest la importància d'interpretar correctament els paràmetres de la recta. També hem vist com hem d'utilitzar la recta de regressió per a fer interpolacions. Finalment, hem comentat una relació no lineal molt important com és la relació exponencial i la manera en què la podem transformar en una de lineal.

Exercicis

1.

El departament de personal d'una empresa informàtica dedicada a la introducció de dades ha fet un programa de formació inicial del personal. La taula següent indica el progrés en pulsacions per minut (p.p.m.) obtingut en mecanografia de vuit estudiants que van seguir el programa i el nombre de setmanes que fa que el segueixen:

Nombre de setmanes	Guany en velocitat (p.p.m.)
3	87
5	119
2	47
8	195
6	162
9	234
3	72
4	110

a) Representeu el diagrama de dispersió. Creieu que és raonable suposar que hi ha una relació lineal entre el nombre de setmanes i el guany de velocitat?

b) Busqueu la recta de regressió. Interpreteu els paràmetres obtinguts.

c) Quin guany de velocitat podem esperar d'una persona que fa set setmanes que va a classe?

2.

Ha sortit al mercat un nou model de gravador de DVD, una mica més car que els anteriors, però amb unes prestacions molt superiors, de manera que la feina dels tècnics dels grans centres comercials és molt important a l'hora de presentar aquest producte al client. Amb l'objectiu de saber si el "nombre de tècnics comercials presents en una botiga" (X) pot tenir alguna incidència en el "nombre d'aparells venuts durant una setmana" (Y), es van observar quinze centres comercials amb els resultats que es mostren a continuació:

$$\sum_{i=1}^{15} x_i = 215\,;\ \sum_{i=1}^{15} x_i^2 = 3.567\,;\ \sum_{i=1}^{15} y_i = 1.700\,;\ \sum_{i=1}^{15} x_i y_i = 28.300$$

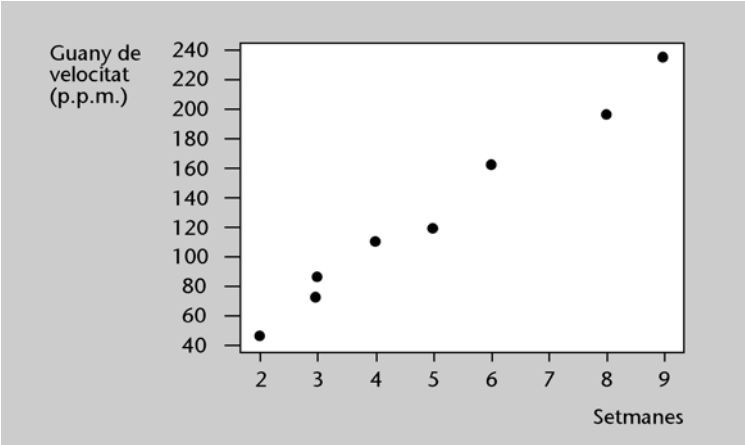
a) Busqueu la recta de regressió.

b) Quin és el nombre d'aparells que es pot estimar que es vendran en un centre amb disset comercials?

Solucionari

1.

Diagrama de dispersió:



El diagrama de dispersió ens mostra que la relació entre totes dues variables és lineal amb pendent positiu, de manera que com més setmanes passen, més gran és el guany de velocitat. Per tant, té sentit buscar la recta de regressió. A partir de la taula de càlculs següent:

i	x_i	y_i	$\bar{x}_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3	87	–2	–41,25	4	82,5
2	5	119	0	–9,25	0	0
3	2	47	–3	–81,25	9	243,75
4	8	195	3	66,75	9	200,25
5	6	162	1	33,75	1	33,75
6	9	234	4	105,75	16	423
7	3	72	–2	–56,25	4	112,5
8	4	110	–1	–18,25	1	18,25
Σ	**40**	**1.026**			**44**	**1.114**

Mitjanes mostrals: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{40}{8} = 5{,}0$ i

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1.206}{8} = 128{,}250$$

Variància mostral: $s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{44{,}00}{7} = 6{,}286$

Covariància mostral: $s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1.114{,}00}{7} = 159{,}143$

Ja podem calcular els coeficients de la recta de regressió:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{159{,}143}{6{,}286} = 25{,}318 \text{ i}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 128{,}250 - 25{,}318 \cdot 5 = 1{,}659$$

La recta de regressió obtinguda és:

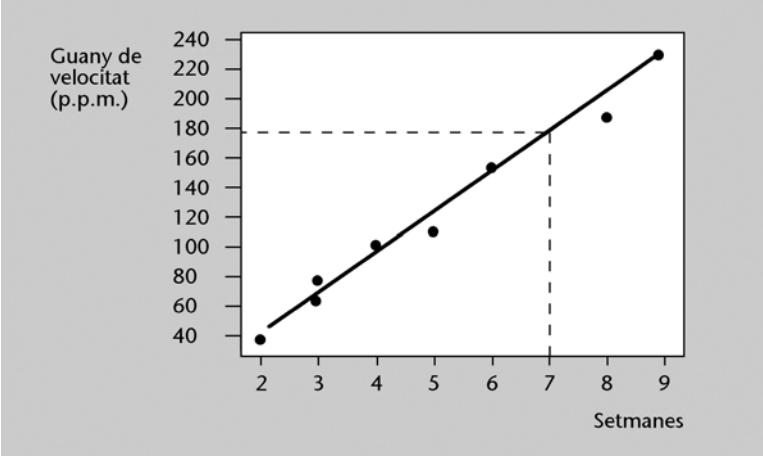
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 1{,}659 + 25{,}318x$$

En aquest cas, l'ordenada a l'origen no té cap interpretació amb sentit, ja que correspondria al guany de velocitat per zero setmanes de classes. Evidentment, no té sentit pensar que sense fer classes es té un guany de velocitat de 1,659 p.p.m. El pendent de la recta sí ens dóna una informació útil: per cada setmana de classe, és te un guany de velocitat d'aproximadament vint-i-cinc p.p.m.

Per a una persona que fa set setmanes que va a classe, podem calcular el guany de velocitat a partir de la recta de regressió, considerant $x = 7$:

$$\hat{y} = 1{,}659 + 25{,}318 \cdot 7 = 178{,}885$$

És a dir, aproximadament un guany de 179 pulsacions per minut.



2.

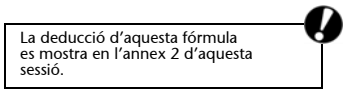
a) Per a trobar la recta de regressió, hem de trobar abans les mitjanes i covariàncies mostrals de les variables X i Y i també la variància mostral de X. A partir de les dades que ens dona l'enunciat:

- Mitjanes mostrals: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{215}{15} = 14{,}333$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1.700}{15} = 113{,}333$$

- Variància mostral:

Per a calcular la variància mostral a partir de les dades de l'enunciat, farem servir l'expressió equivalent:


La deducció d'aquesta fórmula es mostra en l'annex 2 d'aquesta sessió.

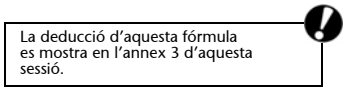
$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}$$

De manera que:

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1} = \frac{3.567 - 15 \cdot 14{,}333^2}{14} = 34{,}667$$

- Covariància mostral:

També ara farem servir una nova expressió per a calcular la covariància mostral:


La deducció d'aquesta fórmula es mostra en l'annex 3 d'aquesta sessió.

$$s_{xy} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1}$$

De manera que:

$$s_{xy} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1} = \frac{28.300 - 15 \cdot 14{,}333 \cdot 113{,}333}{14} = 280{,}952$$

Els paràmetres de la recta de regressió són:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{280{,}952}{34{,}667} = 8{,}104$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 113{,}333 - 8{,}104 \cdot 14{,}333 = -2{,}829$$

La recta de regressió obtinguda és:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2{,}829 + 8{,}104x$$

b) Per a un centre amb disset comercials, podem estimar les vendes d'aparells de DVD mitjançant la recta de regressió obtinguda:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2{,}829 + 8{,}104 \cdot 17 = 134{,}939$$

Per tant, en un centre amb disset comercials s'hauran venut aproximadament uns 135 aparells.

Annexos

Annex 1

Resolució del sistema d'equacions normals:

$$\begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

A partir de la primera equació del sistema:

$$\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = n\bar{y} - n\beta_0 - n\beta_1 \bar{x} = 0$$

Dividint per n: $\bar{y} = \beta_0 + \beta_1 \bar{x}$ i aïllant la β_0: $\beta_0 = \bar{y} - \beta_1 \bar{x}$

De la segona equació del sistema:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 = \sum_{i=1}^{n} x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$\sum_{i=1}^{n} x_i y_i = n\beta_0 \bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2$, però tenim en compte que: $\beta_0 = \bar{y} - \beta_1 \bar{x}$

aleshores $\sum_{i=1}^{n} x_i y_i = n(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 = n\bar{x}\bar{y} - \beta_1 n\bar{x}^2 + \beta_1 \sum_{i=1}^{n} x_i^2$

Aïllant β_1:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

podem donar una expressió equivalent a partir de la definició de variància mostral:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s_x^2(n-1) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

i de la definició de covariància mostral:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy}(n-1) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + n\bar{x}\bar{y} =$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Tenint en compte la variància i covariància, podem expressar els paràmetres de la recta de regressió de la manera següent:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \quad \text{i} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Annex 2

Variància mostral:

Podem deduir a partir de la fórmula de la seva definició:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

una expressió equivalent desenvolupant el quadrat del numerador:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\bar{x}x_i + \sum_{i=1}^{n} (\bar{x})^2 =$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n(\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n(\bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - n(\bar{x})^2$$

De manera que:

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2 \right) - n(\bar{x})^2}{n-1}$$

Annex 3

Covariància mostral:

A partir de la definició de la covariància:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

si desenvolupem el producte de dins el sumatori del numerador:

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-\bar{x}y_i+\bar{x}\bar{y}) = \sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\bar{y}-\sum_{i=1}^{n}\bar{x}y_i+\sum_{i=1}^{n}\bar{x}\bar{y}=$$

$$= \sum_{i=1}^{n}x_iy_i-\bar{y}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}y_i+\bar{x}\bar{y}n = \left(\sum_{i=1}^{n}x_iy_i\right)-n\bar{x}\bar{y}$$

De manera que:

$$s_{xy} = \frac{\left(\sum_{i=1}^{n}x_iy_i\right)-n\bar{x}\bar{y}}{n-1}$$

La qualitat de l'ajust

1. Introducció

La recta de regressió per mínims quadrats minimitza la suma dels quadrats dels residus. Ara ens preguntem si aquest ajust és prou bo. Mirant 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 precisar-ho.

2. El coeficient de determinació, R^2

Volem avaluar en quin grau el model de regressió lineal que hem trobat a partir d'un conjunt d'observacions explica les variacions que es produeixen en la variable dependent d'aquestes.

La mesura més important de la bondat de l'ajust és el **coeficient de determinació R^2.** Aquest coeficient ens indica el grau d'ajust de la recta de regressió als valors de la mostra, i es defineix com la proporció de variància explicada per la recta de regressió, és a dir:

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

Notació

La variància explicada per la recta de regressió és la variància dels valors estimats $\hat{y}_i$.

La variància total de les dades és la variància dels valors observats y_i.

Buscarem una expressió que ens permeti de calcular el coeficient de determinació. Veurem que la variància de les observacions es pot descompondre en dos termes: la variància que queda explicada pel model de regressió lineal i una variància deguda als residus.

A partir de la definició de residus (e_i) de la regressió com la diferencia entre els valor observats (y_i) i els valors estimats ($\hat{y}_i$) per la recta de regressió:

Notació

Anomenarem indistintament *valors estimats* o *valors predits* ($\hat{y}_i$) els obtinguts mitjançant la recta de regressió.

$$e_i = y_i - \hat{y}_i$$

podem escriure:

$$y_i = \hat{y}_i + e_i$$

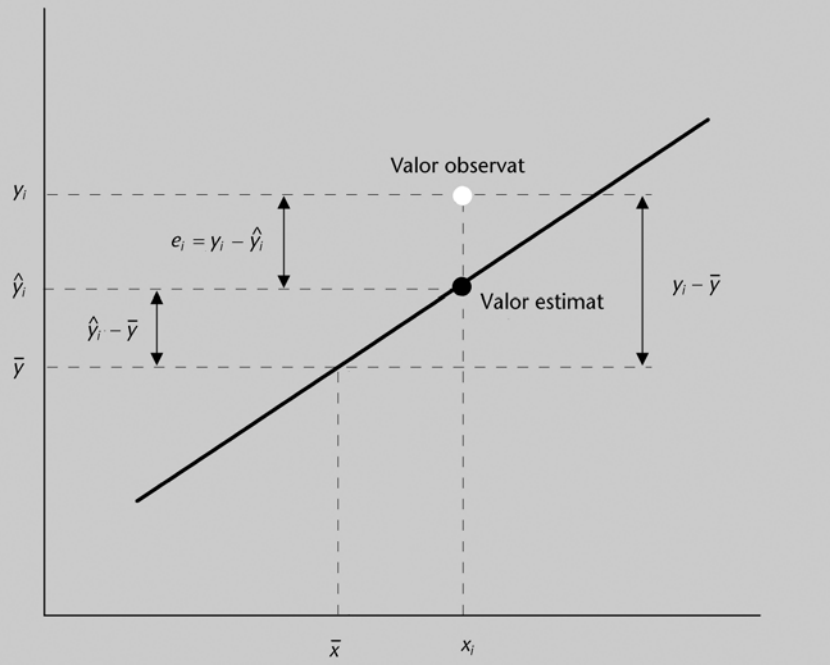
Si ara restem al dos membres d'aquesta igualtat la mitjana de les observacions y_i, obtenim una expressió que ens relaciona les desviacions respecte

de la mitjana de les observacions amb les desviacions respecte la mitjana dels valors estimats.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

Representarem gràficament les desviacions respecte de la mitjana, les observacions i els valors estimats amb la recta de regressió.

Desviacions respecte de la mitjana de les observacions i dels valors estimats amb la recta de regressió



Observació

La recta de regressió passa per $(\bar{x}, \bar{y})$.

Elevant al quadrat i sumant tots els valors, es pot demostrar que:

Aquesta deducció matemàtica es troba desenvolupada en l'annex 1 d'aquesta sessió.

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

Donant noms a aquestes quantitats, podem escriure d'una manera més compacta aquesta expressió:

$\sum_{i=1}^{n} (y_i - \bar{y})^2 = SQT$ Suma de quadrats totals.

$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SQR$ Suma de quadrats de la regressió.

$\sum_{i=1}^{n} e_i^2 = SQE$ Suma de quadrats dels errors.

Així, tenim que:

$$SQT = SQR + SQE.$$

Podem interpretar aquesta última expressió en el sentit que la variància total observada (*SQT*) en la variable *Y* es descompon en dos termes: la variància explicada pel model de regressió lineal (*SQR*) més la variància que no queda explicada pel model, és a dir, la variància dels residus (*SQE*).

Aleshores, podem escriure la definició del **coeficient de determinació** d'aquesta manera:

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

o també:

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

Mirant aquestes expressions, és fàcil veure les característiques d'aquest coeficient. Sempre serà: $0 \leq R^2 \leq 1$, de manera que:

- $R^2 = 1$ quan l'ajust és perfecte, és a dir, quan tots els punts es troben sobre la recta de regressió. En aquest cas, els residus són zero i la suma dels seus quadrats també i, per tant, *SQR* = *SQT*.
- $R^2 = 0$ denota la inexistència de relació entre les variables *X* i *Y*. En aquest cas, la suma de residus és màxima i tenim que *SQE* = *SQT*.
- Com que R^2 ens explica la proporció de variabilitat de les dades que queda explicada pel model de regressió, com més proper a la unitat sigui, millor és l'ajust.

Observació

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

Exemple de les alçades i els pesos

Considerem les observacions dels pesos (kg) i les alçades (cm) d'un conjunt de deu persones: l'individu 1 té 161 cm d'alçada i 63 kg de pes, l'individu 2 té 152 cm d'alçada i 56 kg de pes, etc.

Individus (*i*)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Alçada (x_i)	161	152	167	153	161	168	167	153	159	173
Pes (y_i)	63	56	77	49	72	62	68	48	57	67

A partir de la recta de regressió:

$$\hat{y} = -96{,}1121 + 0{,}979009x$$

podem calcular els valors estimats i els residus. És molt convenient, per comoditat, disposar de les dades i els càlculs en forma de taula; en concret, construirem una taula de càlculs del coeficient de determinació:

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	e_i	e_i^2
1	161	63	61,51	1,10	1,21	–0,39	0,15	1,49	2,23
2	152	56	52,70	–5,90	34,81	–9,20	84,69	3,30	10,91
3	167	77	67,38	15,10	228,01	5,48	30,06	9,62	92,50
4	153	49	53,68	–12,90	166,41	–8,22	67,63	–4,68	21,87
5	161	72	61,51	10,10	102,01	–0,39	0,15	10,49	110,07
6	168	62	68,36	0,10	0,01	6,46	41,75	–6,36	40,47
7	167	68	67,38	6,10	37,21	5,48	30,06	0,62	0,38
8	153	48	53,68	–13,90	193,21	–8,22	67,63	–5,68	32,22
9	159	57	59,55	–4,90	24,01	–2,35	5,52	–2,55	6,50
10	173	67	73,26	5,10	26,01	11,36	128,97	–6,26	39,14
Σ		**619**			**812,90**		**456,61**		**356,29**

Tenim que:

$$SQR = 456{,}61$$
$$SQT = 812{,}90$$

Per tant, tenim un coeficient de determinació:

$$R^2 = 456{,}61 \,/\, 812{,}90 = 0{,}5617$$

Amb aquest exemple, podem comprovar l'equivalència entre les dues expressions obtingudes abans pel coeficient de determinació. A partir de la suma dels quadrats dels residus:

$$SQE = 356{,}29$$

tenim per al coeficient de determinació:

$$R^2 = 1 - (\,356{,}29 \,/\, 812{,}90\,) = 1 - 0{,}4383 = 0{,}5617$$

Evidentment, coincideixen els resultats.

Hem obtingut un coeficient de determinació $R^2 = 0{,}5617$ que ens informa que el model de regressió lineal només ens explica el 56,17% de la variància de les observacions.

3. El coeficient de correlació mostral, *r*

A partir del diagrama de dispersió podem veure si hi ha algun tipus de relació entre dues variables X i Y.

S'acostuma a dir que X i Y tenen una **relació positiva** si els valors grans de X estan aparellats amb valors grans de Y i valors petits de X, amb valors petits de Y. De manera anàloga, es diu que X i Y tenen una **relació negativa** si els valors grans de X estan aparellats amb els valors petits de Y i els petits de X, amb grans de Y.

Ara volem mesurar aquestes relacions de forma numèrica. La covariància mostral entre dues variables X i Y:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

ens pot servir per a mesurar aquestes relacions positives i negatives entre les variables X i Y.

- Si tenim una relació positiva, aleshores la majoria dels punts de coordenades $((x_i - \bar{x}),\ (y_i - \bar{y}))$ estaran en el primer i tercer quadrant en què $(x_i - \bar{x})(y_i - \bar{y}) \geq 0$, de manera que contribuiran positivament a la suma.

- Si tenim una relació negativa, aleshores la majoria dels punts de coordenades $((x_i - \bar{x}),\ (y_i - \bar{y}))$ estaran en el segon i quart quadrant en què $(x_i - \bar{x})(y_i - \bar{y}) \leq 0$, de manera que contribuiran negativament a la suma.

- Si al contrari, no hi ha cap tipus de relació positiva o negativa, la covariància serà una quantitat petita en trobar-se tots els punts aproximadament igual repartits pels quatre quadrants, cosa que compensa de forma aproximada les quantitats positives i negatives del sumatori.

La covariància té el gran inconvenient de dependre de les unitats de les variables que estudiem.

Vegeu la figura dels exemples de diagrames de dispersió en la secció 3 de la sessió "El mòdul de regressió simple" d'aquesta assignatura.

Esquema de relacions entre *X* i *Y*

Relacions positives i negatives entre les variables X i Y

	2n. quadrant $(x_i-\bar{x})\leq 0$ $(y_i-\bar{y})\geq 0$	1r. quadrant $(x_i-\bar{x})\geq 0$ $(y_i-\bar{y})\geq 0$
$\bar{y}$	3r. quadrant $(x_i-\bar{x})\leq 0$ $(y_i-\bar{y})\leq 0$	4t. quadrant $(x_i-\bar{x})\geq 0$ $(y_i-\bar{y})\leq 0$
		$\bar{x}$

Definim el **coeficient de correlació mostral** com a:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Unitats del coeficient de correlació mostral

En dividir la covariància per les desviacions típiques de X i de Y, hem aconseguit una mesura adimensional que no depèn de les unitats de les variables.

El coeficient de correlació es caracteritza per $-1 \leq r \leq 1$, de manera que:

- $r = 1$ o $r = -1$ quan hi hagi una associació lineal exacta entre les variables (en el primer cas positiva i en el segon negativa).

- $-1 < r < 1$ quan la relació entre les variables no sigui lineal de forma exacta.

- Per als altres valors sempre es formula la mateixa 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 $0 < |r| < 0{,}5$; forta si $0{,}8 < |r| < 1$; i moderada si té un altre valor.

Per a calcular el coeficient de correlació mostral, podem fer servir la mateixa taula de càlculs que per a obtenir la recta de regressió. Ho il·lustrarem amb l'exemple de les alçades i els pesos.

Exemple de les alçades i els pesos

Considerem de nou l'exemple dels pesos i alçades. Buscarem el coeficient de correlació. Abans haurem de calcular la covariància i les variàncies mostrals.

i	x_i	y_i	$\bar{x}-x_i$	$\bar{y}-y_i$	$(\bar{x}-x_i)^2$	$(\bar{y}-y_i)^2$	$(\bar{x}-x_i)(\bar{y}-y_i)$
1	161	63	0,4	–1,1	0,16	1,21	–0,44
2	152	56	9,4	5,9	88,36	34,81	55,46
3	167	77	–5,6	–15,1	31,36	228,01	84,56
4	153	49	8,4	12,9	70,56	166,41	108,36
5	161	72	0,4	–10,1	0,16	102,01	–4,04
6	168	62	–6,6	–0,1	43,56	0,01	0,66
7	167	68	–5,6	–6,1	31,36	37,21	34,16
8	153	48	8,4	13,9	70,56	193,21	116,76
9	159	57	2,4	4,9	5,76	24,01	11,76
10	173	67	–11,6	–5,1	134,56	26,01	59,16
Σ	**1.614**	**619**			**476,40**	**812,90**	**466,40**

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{466,40}{10-1} = 51,822$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{476,40}{10-1} = 52,933 \text{ de manera que } s_x = 7,276$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{812,90}{10-1} = 90,322 \text{ de manera que } s_y = 9,504$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{51,822}{7,276 \cdot 9,504} = 0,749$$

El coeficient de correlació lineal obtingut pel nostre exemple del pes i l'alçada és $r = 0,749$, que ens informa de l'existència d'una moderada relació entre aquestes dues variables. I també que a mesura que l'alçada creix, el pes també ho fa (ja que és positiu).

4. 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: mesura el grau d'associació entre les dues variables.

No obstant això, en la regressió lineal simple tenim que $R^2 = r^2$, com molt fàcilment podem comprovar.

Observació

En la regressió lineal múltiple ja no tindrem la igualtat $R^2 = r^2$.

Comprovació que en regressió lineal simple, $R^2 = r^2$

A partir de l'equació del coeficient de correlació:

$$r = \frac{s_{xy}}{s_x s_y}$$

i de l'equació del pendent de la recta de regressió:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

tenim la relació següent:

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

D'altra banda, tenim l'altre paràmetre de la recta de regressió: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ i l'equació dels valores estimats: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. D'aquestes dues expressions podem escriure:

$$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) .$$

Aplicant totes aquestes relacions a l'equació del coeficient de determinació, i a partir de la definició de variància mostral, tenim:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \hat{\beta}_1^2\frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = r^2 \underbrace{\frac{s_y^2\sum(x_i - \bar{x})^2}{s_x^2\sum(y_i - \bar{y})^2}}_{1} = r^2$$

Aquesta relació ens ajuda a comprendre per què abans hem considerat que un valor de $r = 0{,}5$ el consideràvem dèbil. Aquest valor representarà un $R^2 = 0{,}25$, és a dir, que el model de regressió només ens explica un 25% de la variabilitat total de les observacions.

També és important tenir present que r ens dóna més informació que R^2. El signe de r ens informa de si la relació és positiva o negativa. Així, doncs, amb el valor de r sempre podrem calcular el valor de R^2, però al revés sempre ens quedarà indeterminat el valor del signe llevat que coneguem el pendent de la recta. Per exemple, donat un $R^2 = 0{,}81$, si sabem que el pendent de la recta de regressió és negatiu, aleshores podrem afirmar que el coeficient de correlació serà $r = -0{,}9$.

Exemple de les alçades i els pesos

Podem comprovar la relació entre el coeficient de determinació i el coeficient de correlació amb els resultats del nostre exemple.

Hem obtingut: $R^2 = 0{,}5617$ i $r = 0{,}749$.

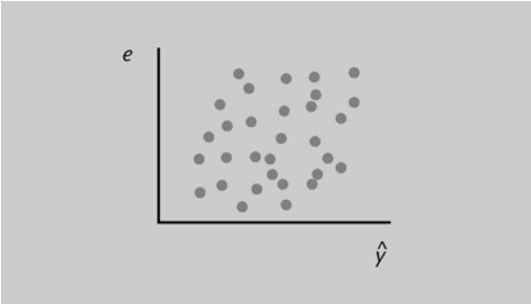
De manera que $r^2 = 0{,}749^2 = 0{,}561$.

5. Diagnòstic de la regressió: anàlisi dels residus

Una vegada fet l'ajust d'un model de regressió lineal a les nostres dades mostrals, cal fer l'anàlisi dels residus.

Aquesta anàlisi, que a continuació comentarem de forma breu i molt intuïtiva, ens servirà per a fer un diagnòstic del nostre model de regressió.

L'anàlisi dels residus consisteix a veure la distribució dels residus. Això ho farem gràficament representant un diagrama de dispersió dels punts $(\hat{y}_i, e_i)$, és a dir, sobre l'eix de les abscisses representem el valor estimat $\hat{y}_i$ i sobre l'eix d'ordenades el valor corresponent del residu, és a dir, $e_i = y_i - \hat{y}_i$. Vegem-ne un exemple:



Si el model lineal obtingut s'ajusta prou bé a les dades mostrals, aleshores el núvol de punts $(\hat{y}_i, e_i)$ no ha de mostrar cap tipus d'estructura.

Ho il·lustrarem amb un exemple ja clàssic en la bibliografia: l'**exemple d'Anscombe** (1973). A partir de les taules de dades que es mostren a continuació, discutirem quatre casos:

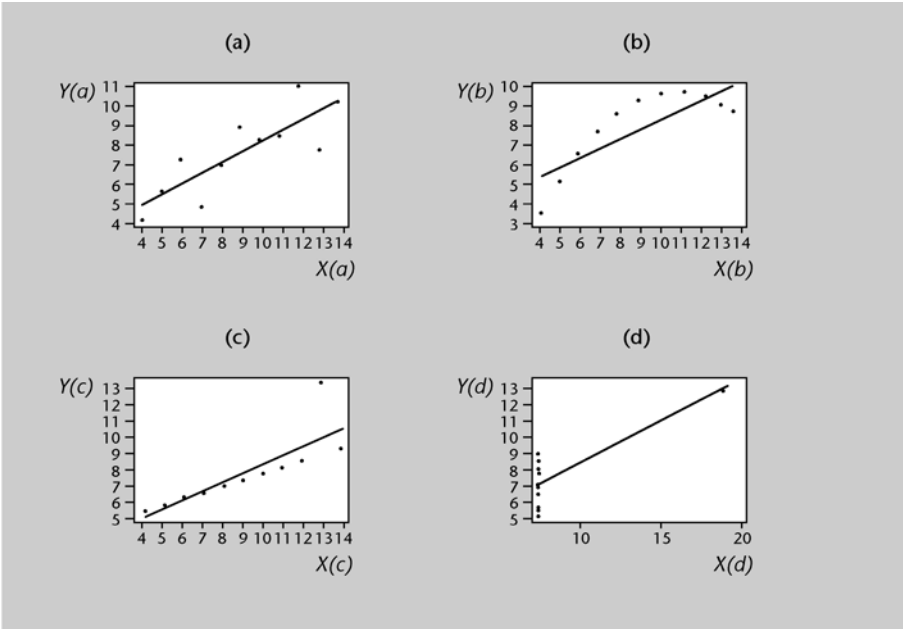
Lectura complementària

Trobareu l'exemple d'Anscombe a l'article següent:

T.W. Anscombe (1973). "Graphs in Statistical Analysis". *The American Statistician* (núm. 27, pàg. 17-21).

Cas (a)		Cas (b)		Cas (c)		Cas (d)	
X(a)	*Y(a)*	*X(b)*	*Y(b)*	*X(c)*	*Y(c)*	*X(d)*	*Y(d)*
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Dibuixarem a continuació el diagrama de dispersió i les rectes de regressió en l'exemple d'Anscombe.

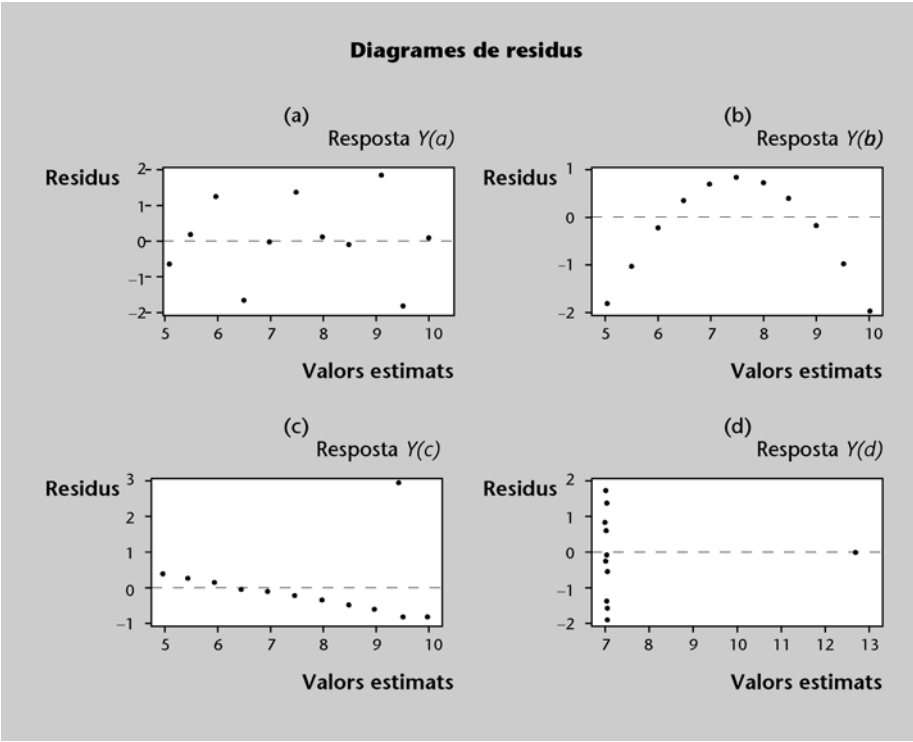


Si fem la regressió de Y sobre X, en els quatre casos obtenim la mateixa recta:

$$\hat{y} = 3 + 0{,}5x$$

El coeficient de correlació és el mateix per a les quatre amb valor $r = 0{,}82$.

Si ara fem l'estudi dels residus tal com hem indicat abans, tenim la representació dels diagrames dels residus següents:



Podem observar que de les quatre, només la primera no presenta cap tipus d'estructura sobre el núvol de punts, de manera que només tindria sentit la regressió feta sobre la mostra (a).

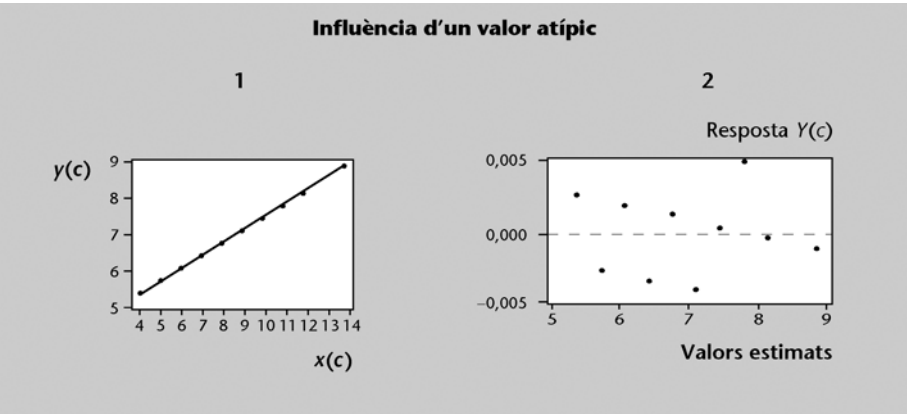
Considerem a continuació el cas (b) del diagrama de dispersió. S'hi observa un comportament curvilini que ens fa pensar que un ajust lineal no seria el més convenient. Això es manifesta de forma molt més evident en el diagrama de residus.

Si considerem la mostra (c), en el diagrama de dispersió podem observar la presencia del valor atípic (13, 12,74) que ens ha fet ajustar un model erroni a la resta de les observacions, ja que si l'eliminem, aleshores obtenim una recta de regressió diferent:

$$\hat{y} = 4{,}01 + 0{,}345x$$

i un coeficient de correlació $r = 1$. Podem observar tots els punts sobre la recta de regressió.

El diagrama dels residus també ens suggereix un bon model de regressió per la mostra resultant d'eliminar el valor atípic. A continuació, fem el diagrama de dispersió i el diagrama de residus.



Influència d'un valor atípic

En la mostra (c) hem eliminat el valor atípic i hem representat de nou el diagrama de dispersió i la recta de regressió **1** i el diagrama de residus **2**.

Finalment, en la mostra (d) el pendent està determinat per un únic valor. Tampoc no és un model gaire fiable.

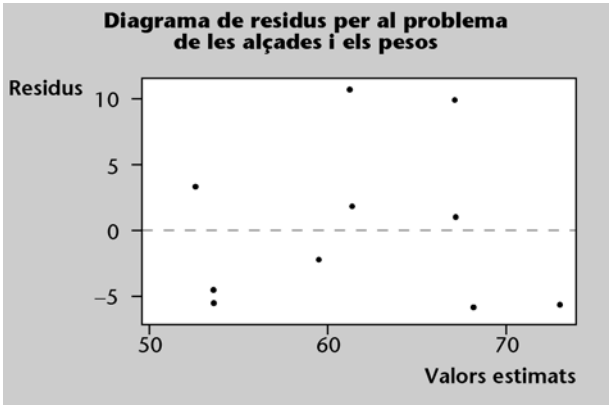
Exemple de les alçades i els pesos

Un últim exemple que encara podem examinar és el de la relació de les alçades i pesos. A partir de les dades de la taula ja vista:,

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	e_i	e_i^2
1	161	63	61,51	1,10	1,21	–0,39	0,15	1,49	2,23
2	152	56	52,70	–5,90	34,81	–9,20	84,69	3,30	10,91
3	167	77	67,38	15,10	228,01	5,48	30,06	9,62	92,50
4	153	49	53,68	–12,90	166,41	–8,22	67,63	–4,68	21,87
5	161	72	61,51	10,10	102,01	–0,39	0,15	10,49	110,07

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	e_i	e_i^2
6	168	62	68,36	0,10	0,01	6,46	41,75	–6,36	40,47
7	167	68	67,38	6,10	37,21	5,48	30,06	0,62	0,38
8	153	48	53,68	–13,90	193,21	–8,22	67,63	–5,68	32,22
9	159	57	59,55	–4,90	24,01	–2,35	5,52	–2,55	6,50
10	173	67	73,26	5,10	26,01	11,36	128,97	–6,26	39,14
Σ		**61,9**			**812,90**		**456,61**		**356,29**

és fàcil representar el diagrama de residus:



No podem observar cap tipus d'estructura en la representació; per tant, podem concloure que el model de regressió obtingut és un bon model per a explicar la relació entre totes dues variables.

6. Resum

En aquesta segona sessió, hem introduït una mesura numèrica de la bondat de l'ajust de la recta de regressió a les observacions. Aquesta mesura la fem amb el coeficient de determinació R^2. S'ha discutit la interpretació dels valors que pot prendre. A continuació hem vist el coeficient de correlació mostral, r, que ens mesura el grau d'associació entre dues variables. Hem comprovat que en la regressió lineal simple, R^2 i r coincideixen. Finalment, hem comentat la importància de fer una anàlisi dels residus per a fer un diagnòstic del model lineal obtingut.

Exercicis

1.

Una botiga d'ordinadors va fer un estudi per a determinar la relació entre les despeses de publicitat setmanal i les vendes. Es van obtenir les dades següents:

Despeses en publicitat (× 1000 €)	Vendes (× 100.000 €)
40	380
25	410
20	390
22	370
31	475
52	450
40	500
20	390
55	575
42	520

Amb aquestes dades s'han obtingut les quantitats següents:

$$\sum_{i=1}^{10} x_i = 347 \qquad \sum_{i=1}^{10} y_i = 4.460 \qquad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 6.018$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1.522,1 \qquad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 43.590,0$$

$$\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 23.793,66$$

I la recta de regressió: $\hat{y} = 308,88 + 3,95x$

A partir de tota aquesta informació, calculeu el coeficient de determinació i el coeficient de correlació.

2.

El departament de personal d'una empresa informàtica dedicada a la introducció de dades ha fet un programa de formació inicial del personal. La taula següent indica el progrés obtingut en mecanografia de vuit estudiants que van seguir el programa i el nombre de setmanes que fa que el segueixen:

Nombre de setmanes	Guany en velocitat (p.p.m.)
3	87
5	119
2	47
8	195

Nombre de setmanes	Guany en velocitat (p.p.m.)
6	162
9	234
3	72
4	110

La recta de regressió calculada a partir d'aquestes dades és:

$$\hat{y} = 1{,}659 + 25{,}318x$$

a) Calculeu el coeficient de determinació.

b) Feu una anàlisi dels residus i comenteu-lo.

Solucionari

1.

Calculem el coeficient de determinació a partir de l'expressió:

$$R^2 = \frac{SQR}{SQT}$$

Aquestes dades ens la proporciona l'enunciat del problema, ja que:

- La suma dels quadrats de la regressió és: $SQR = \sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 23.793{,}66$
- I la suma dels quadrats totals és: $SQT = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 43.590{,}0$

De manera que: $R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{10} (y_i - \bar{y})^2} = \frac{23.793{,}66}{43.590{,}0} = 0{,}5458$

Que podem interpretar com que el model de regressió lineal explica el 54,58% de la variabilitat de les vendes.

A partir d'aquest valor, podem calcular el coeficient de correlació tenint en compte que:

$$R^2 = r^2$$

De manera que el coeficient de correlació és l'arrel quadrada del coeficient de determinació amb el mateix signe del pendent de la recta de regressió.

La recta de regressió és: $\hat{y} = 308{,}8 + 3{,}95x$. El pendent és positiu, de manera que tenim una relació positiva entre les despeses en publicitat i vendes. Com més s'inverteix en publicitat, més es ven.

Així, doncs, el coeficient de correlació és:

$$r = +\sqrt{R^2} = +\sqrt{0{,}5458} = 0{,}7388$$

2.

a) El primer que farem serà construir la taula de càlculs:

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
1	3	87	77,61	–41,25	1.701,56	–50,64	2.564,11
2	5	119	128,25	–9,25	85,56	0,00	0,00
3	2	47	52,30	–81,25	6.601,56	–75,96	5.769,16
4	8	195	204,20	66,75	4.455,56	75,95	5.768,86
5	6	162	153,57	33,75	1.139,06	25,32	640,95
6	9	234	229,52	105,75	11.183,06	101,27	10.255,82
7	3	72	77,61	–56,25	3.164,06	–50,64	2.564,11
8	4	110	102,93	–18,25	333,06	–25,32	641,05
Σ		**1.026**			**28.663,50**		**28.204,05**

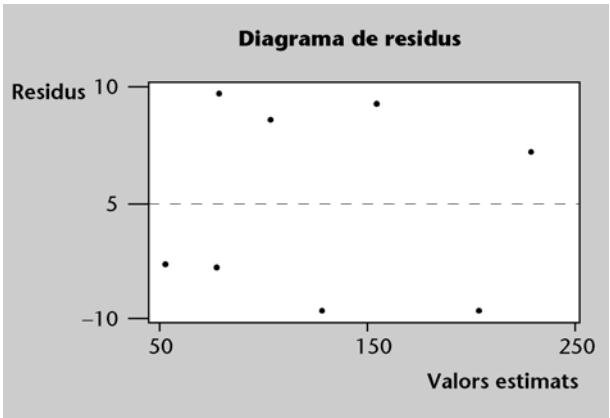
$$SQR = 28.204,05 \quad SQT = 28.663,50$$
$$R^2 = 28.204,05 \,/\, 28.663,50 = 0,9920$$

El model de regressió lineal explica el 99,20% de la variància de la mostra. Tenim bondat en l'ajust.

b) Per a fer l'anàlisi dels residus, en primer lloc calcularem els residus i després farem la representació gràfica.

i	x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	3	87	77,61	9,39
2	5	119	128,25	–9,25
3	2	47	52,30	–5,30
4	8	195	204,20	–9,20
5	6	162	153,57	8,43
6	9	234	229,52	4,48
7	3	72	77,61	–5,61
8	4	110	102,93	7,07

Si representem el valor del residu enfront del valor ajustat, tenim el diagrama dels residus següent:



No observem cap mena de forma determinada en els punts d'aquesta gràfica.

Aquest resultat juntament amb el coeficient de determinació tan alt ens fa concloure que el model lineal és adient per a tractar aquest problema.

Annexos

Annex 1

Descomposició de la suma de quadrats total

A continuació, veurem que la suma de quadrats total de les observacions (*SQT*) es pot expressar de la manera següent:

$$SQT = SQR + SQE$$

on:

- *SQR* és la suma de quadrats de la regressió.
- *SQE* és la suma de quadrats dels residus.

A partir de la definició de residus de la regressió com la diferencia entre els valors observats i els valors estimats per la recta de regressió:

$$e_i = y_i - \hat{y}_i$$

Podem escriure:

$$y_i = \hat{y}_i + e_i$$

I si ara restem al dos membres d'aquesta igualtat la mitjana de les observacions y_i, obtenim una expressió que ens relaciona les desviacions respecte de la mitjana, de les observacions i dels valors estimats:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

Elevant al quadrat i sumant tots els valors:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(\hat{y}_i - \bar{y}) + e_i]^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i + \sum_{i=1}^{n} e_i^2$$

$$2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i = \sum_{i=1}^{n} \hat{y}_i e_i - \sum_{i=1}^{n} \bar{y} e_i = \sum_{i=1}^{n} \hat{y}_i e_i - \bar{y}\sum_{i=1}^{n} e_i = 0 + 0 = 0$$

$$0 \qquad 0$$

Per tant, n'hi ha prou de veure que $\sum_{i=1}^{n} \hat{y}_i e_i = 0$ i $\sum_{i=1}^{n} e_i = 0$.

Observem que a partir de les equacions normals:

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i$$

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^{n} e_i x_i$$

I, per tant:

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 - \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i = 0$$

Hem demostrat així que:

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

Si anomenem:

$\sum_{i=1}^{n} (y_i - \bar{y})^2 = SQT$ **S**uma de **Q**uadrats **T**otals.

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

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

Tenim que: $SQT = SQR + SQE$.

Inferència en la regressió

1. Introducció

En altres sessions ens hem preocupat d'estudiar la relació lineal entre dues variables X i Y a partir dels valors observats en una mostra. Si en el diagrama de dispersió observàvem una relació lineal, aleshores calculàvem la recta que millor s'ajustava a les nostres dades fent que la suma dels quadrats dels residus fos mínima. És l'anomenada *recta de regressió*.

Ara canviarem el punt de vista i pensarem que aquesta mostra d'observacions prové d'una població. Ens preguntem si aquesta relació lineal la podem estendre d'alguna manera a tota la població.

2. El model de regressió en la població

Model de regressió lineal

És molt important tenir present que, per a un mateix valor de la variable X, es poden observar diferents valors de la variable Y, és a dir, associat a cada valor de X no hi ha un únic valor de Y, sinó una distribució de freqüències de Y. Això es deu al fet que Y no solament depèn de X, sinó també d'altres factors difícilment quantificables o simplement desconeguts. La influència d'aquest conjunt de factors és la que determina que la relació entre X i Y sigui estadística i no determinista. Tots aquests factors són els responsables dels errors o residus.

El pes depèn de l'alçada i d'altres factors

En l'exemple de la relació entre el pes i l'alçada de les persones, és evident que hi ha molts factors, com poden ser aspectes genètics, l'activitat física, l'alimentació, etc., que fan que una persona d'una determinada alçada tingui un pes o un altre. Per a una alçada fixa, de per exemple 170 cm, no totes les persones tenen el mateix pes.

Donada una mostra d'observacions (x_i, y_i), $i = 1, \ldots, n$ d'individus d'una població, ja sabem trobar la recta de regressió lineal $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Si tenim en compte que anomenàvem *residu* o *error* la diferencia entre el valor observat i el valor estimat $e_i = y_i - \hat{y}_i$, per a una observació y_i, podem escriure: $y_i = \hat{y}_i + e_i$, és a dir:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x + e_i$$

Això mateix ho podem fer amb diverses mostres d'aquesta mateixa població.

Exemple de les alçades i els pesos

Considerem les observacions dels pesos (kg) i alçades (cm) de tres mostres d'alumnes de la UOC i les rectes de regressió corresponents:

Mostra $j = 1$										
Individus	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
Alçada (x_{ij})	161	152	167	153	161	168	167	153	159	173
Pes (y_{ij})	63	56	77	49	72	62	68	48	57	67

La recta de regressió corresponent és: $\hat{y} = -96{,}112 + 0{,}979x$.

Mostra $j = 2$								
Individus	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
Alçada (x_{ij})	161	152	167	153	161	168	167	153
Pes (y_{ij})	63	56	77	49	72	62	68	48

La recta de regressió corresponent és: $\hat{y} = -82{,}614 + 1{,}029x$.

Mostra $j = 3$									
Individus	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$
Alçada (x_{ij})	161	152	167	153	161	168	167	153	159
Pes (y_{ij})	63	56	77	49	72	62	68	48	57

La recta de regressió corresponent és: $\hat{y} = -98{,}582 + 0{,}94x$.

Observem que els valors obtinguts per cada coeficient són relativament similars:

$\hat{\beta}_0$: –96,112; –82,614; –98,528

$\hat{\beta}_1$: 0,979; 1,029; 0,945

Podem pensar que si agafem més mostres de la mateixa població, anirem obtenint coeficients semblants a aquests.

Ara l'objectiu és donar un model per a tots els individus de la població. Aquest vindrà donat per una expressió anàloga a les trobades per les mostres.

Anomenem **model de regressió lineal per a la població:**

$$y_i = \beta_0 + \beta_1 x + e_i$$

Notació

No posem els "barrets" sobre els paràmetres per a indicar que ara es tracta de la recta de regressió per a la població.

Per a trobar aquest model per a la població, n'hauríem d'estudiar tots els individus. Això és pràcticament impossible, de manera que l'haurem d'estimar a partir dels resultats calculats per a una mostra. És a dir, haurem de fer inferència estadística.

Distribució dels errors en la realitat

La distribució dels errors és diferent per a diferents valors de *X*. Per exemple, les persones que mesuren prop de 160 cm varien menys en pes que les persones que mesuren 185 cm. De totes formes, acceptarem la suposició que sempre són iguals.

Abans de continuar, ens calen fer dues suposicions molt importants:

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

2) Els errors són independents.

Amb aquestes suposicions tenim que:

1) Per cada valor fix x de X obtenim una distribució de valors y de la variable Y. I per cadascuna d'aquestes distribucions en podem calcular la mitjana o esperança matemàtica:

$$\mu_x = E(Y|x) = E(\beta_0 + \beta_1 x + e) = \beta_0 + \beta_1 x + E(e) = \beta_0 + \beta_1 x$$

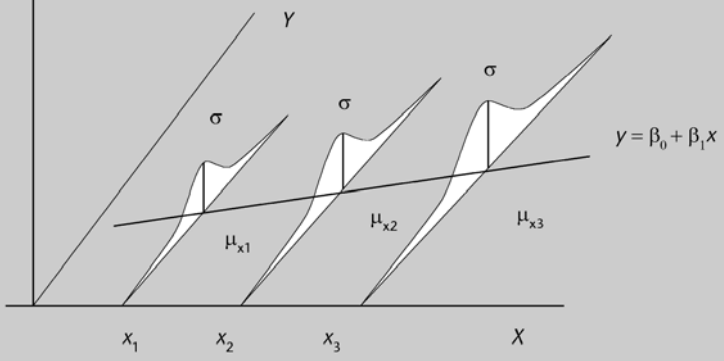
2) També podem calcular la seva variància:

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + e) = Var(\beta_0 + \beta_1 x) + Var(e) = 0 + \sigma^2 = \sigma^2$$

Cada distribució de valors de Y té la mateixa variància σ^2, que és la variància dels residus.

En el gràfic veiem la recta de regressió lineal per a la població.

Recta de regressió lineal per a la població



Distribució de les mitjanes

El primer resultat ens diu que aquestes mitjanes es troben situades sobre una recta.

És important tenir present que per a tenir ben determinat el model de regressió per a la població, hem de conèixer tres paràmetres: β_0, β_1 i σ^2.

Aquests paràmetres desconeguts els haurem d'estimar a partir d'una mostra de la població.

Com es veu en la sessió "El model de regressió simple", els paràmetres de la recta s'estimem pel mètode dels mínims quadrats. Aquest mètode determina aquells valors dels paràmetres que fan mínima la suma dels quadrats dels residus:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\,;\ \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

De manera que $\hat{\beta}_0$ i $\hat{\beta}_1$ són els valors estimats (o "estimadors") dels paràmetres β_0 i β_1 de la població. I la recta que millor s'ajusta a les dades és:

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

Valor mitjà

Hem d'interpretar

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

com l'estimació del valor mitjà de la distribució Y per a un valor fix $X = x_i$.

Encara ens falta estimar la variància dels errors aleatoris, σ^2. Aquest terme reflecteix la variació aleatòria al voltant de la veritable recta de regressió.

Si considerem els residus de la regressió com a estimacions del valors dels errors aleatoris, aleshores en podem estimar la variància a partir de la variància dels residus:

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

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Terminologia

Habitualment, s^2 es denomina *variància residual*.

Hem dividit la suma de les desviacions al quadrat per $n - 2$, no pas per $n - 1$. Això és perquè estimem la mitjana de Y per a un valor donat de X amb una fórmula que conté dos paràmetres estimats a partir de les dades de la mostra ($\hat{\beta}_0$ i $\hat{\beta}_1$). Direm que "hem perdut dos graus de llibertat".

Pèrdua de graus de llibertat

El raonament és el mateix que el que fem en justificar la divisió per $(n - 1)$ en la fórmula de la variància mostral:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$$

Ho fem perquè hem perdut un grau de llibertat en estimar la mitjana a partir de les dades de la mostra.

Exemple de les alçades i els pesos

Considerem les observacions dels pesos (kg) i alçades (cm) d'un conjunt de deu persones:

Individus (i)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Alçada (x)	161	152	167	153	161	168	167	153	159	173
Pes (y_i)	63	56	77	49	72	62	68	48	57	67

La recta de regressió corresponent és:

$$\hat{y} = -96{,}112 + 0{,}979x$$

A la sessió "El model de regressió simple" es dedueix la recta de regressió corresponent a aquest exemple.

Per a fer els càlculs més còmodes, és aconsellable construir la taula de càlculs per la variància dels residus que es mostra tot seguit.

i	x_i	y_i	$\hat{y}_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$e_i = y_i - \hat{y}_i$	e_i^2
1	161	63	61,51	–0,4	0,16	1,49	2,225
2	152	56	52,70	–9,4	88,36	3,30	10,908
3	167	77	67,38	5,6	31,36	9,62	92,498
4	153	49	53,68	–8,4	70,56	–4,68	21,868
5	161	72	61,51	–0,4	0,16	10,49	110,075
6	168	62	68,36	6,6	43,56	–6,36	40,468
7	167	68	67,38	5,6	31,36	0,62	0,381
8	153	48	53,68	–8,4	70,56	–5,68	32,220

i	x_i	y_i	$\hat{y}_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$e_i = y_i - \hat{y}_i$	e_i^2
9	159	57	59,55	–2,4	5,76	–2,55	6,504
10	173	67	73,26	11,6	134,56	–6,26	39,143
Σ	**1.614**	**619**			**476,4**		**356,290**

La vuitena columna té els quadrats dels residus. Sumant totes les dades i dividint pel nombre d'observacions menys 2, és a dir, per 10 – 2 = 8, obtenim la variància dels residus:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{356{,}290}{10-2} = 44{,}536$$

3. Distribució probabilística del pendent ($\hat{\beta}_1$)

L'ordenada a l'origen β_0 ens informa del valor mitjana de la variable *Y* per a un valor de *X* igual a zero. No sempre té interpretació realista en el context del problema: per aquest motiu, únicament considerarem fer inferència estadística sobre el pendent.

Per a poder fer inferència estadística (fer contrastos d'hipòtesis i cercar intervals de confiança), serà necessari conèixer la distribució de probabilitat de $\hat{\beta}_1$.

Del model de regressió lineal, tenim que $\hat{\beta}_1$ és una combinació lineal de les observacions y_i; i si aquestes tenen una distribució normal i són independents (tal com hem suposat en establir el model de regressió), aleshores $\hat{\beta}_1$ també tindrà una distribució normal. Tindrem ben determinada aquesta distribució quan en coneguem l'esperança i la variància.

A partir de l'expressió de $\hat{\beta}_1$ podem trobar el valor esperat i la variància.

Els desenvolupaments matemàtics es mostren en l'annex d'aquesta sessió.

- Valor esperat de $\hat{\beta}_1$:

$$E(\hat{\beta}_1) = \beta_1$$

El pendent estimat de la recta està distribuït segons una distribució normal amb una mitjana igual al valor d'aquest paràmetre per a la població. Encara que aquest valor és desconegut, aquest resultat ens serà molt útil per a tenir informació de la població fent inferència estadística. Això ho veurem una mica més endavant en aquesta sessió.

- Variància de $\hat{\beta}_1$:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

A continuació, veurem que ens farà falta la informació de la mostra, ja que σ^2 és un valor desconegut que haurem d'estimar.

4. L'interval de confiança per al pendent

Acabem de veure que les suposicions del model de regressió lineal simple impliquen que el paràmetre $\hat{\beta}_1$ és una variable aleatòria distribuïda normalment amb:

– Mitjana: β_1

– Variància: $\sigma^2_{\hat{\beta}_1} = \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2}$

Com que aquesta variància σ^2 és desconeguda, l'haurem d'estimar a partir de la variància mostral que ja hem calculat anteriorment:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Definim l'error estàndard del pendent com a:

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

Com que $\hat{\beta}_1$ segueix una distribució normal amb variància desconeguda (ja que no es coneix σ^2), aleshores la variable tipificada:

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

té una distribució *t* de Student amb $n - 2$ graus de llibertat.

Interval de confiança pel pendent amb un nivell significatiu α.

Amb tot això, tenim que un **interval de confiança** de 100 $(1 - \alpha)$% pel pendent β_1 de la recta de regressió poblacional ve donat per:

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

ja que:

$$P\left(-t_{\alpha/2,\,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \leq t_{\alpha/2,\,n-2}\right) = 1 - \alpha$$

Aquest interval està centrat en l'estimació puntual del paràmetre, és a dir, a $\hat{\beta}_1$ i la quantitat en què s'allarga a cada costat de l'estimació depèn del nivell desitjat de confiança, α (mitjançant el valor crític $t_{\alpha/2,\,n-2}$) i de la variabilitat de l'estimador $\hat{\beta}_1$ (mitjançant $s_{\hat{\beta}_1}$).

Exemple de les alçades i els pesos

Considerem una vegada més l'exemple dels pesos i les alçades d'una mostra de deu persones. La recta de regressió corresponent era: $\hat{y} = -96{,}112 + 0{,}979x$, de manera que $\beta_1 = 0{,}979$.

Calcularem un interval de confiança del 95% per al pendent. Per tant, $\alpha = 0{,}05$ i mirant la taula de la t de Student tenim un valor crític de $t_{\alpha/2;\, n-2} = t_{0{,}025;8} = 2{,}3060$.

Per a calcular l'interval de confiança: $[\hat{\beta}_1 - t_{\alpha/2,\, n-2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\, n-2} s_{\hat{\beta}_1}]$, abans hem de calcular:

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

on:

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

Anteriorment, ja hem calculat la variància dels residus:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{356{,}290}{10-2} = 44{,}536$$

De manera que:

$$s_{\hat{\beta}_1}{}^2 = \frac{s^2}{\sum (x_i - \bar{x})^2} = \frac{44{,}536}{476{,}4} = 0{,}093$$

Per tant, l'error estàndard del pendent serà: $s_{\hat{\beta}_1} = \sqrt{0{,}093} = 0{,}306$.

I l'interval de confiança és: [0,979 – 2,3060 · 0,306; 0,979 + 2,3060 · 0,306].

Finalment tenim, [0,274; 1,684]. Així, doncs, tenim un 95% de probabilitat que el pendent de la recta de regressió per a la població es trobi en aquest interval.

5. El contrast d'hipòtesis sobre el pendent

Observem que si en el model de regressió lineal el pendent és zero, aleshores la variable X no té cap efecte sobre la variable Y. En aquest cas, direm que X no és una **variable explicativa** del model.

En aquest apartat farem un contrast d'hipòtesi sobre el pendent de la recta de regressió per a saber si podem afirmar o no que aquest és igual a zero.

Com en tots els contrastos d'hipòtesis, seguirem els passos següents:

1) Establim les hipòtesis nul·la i alternativa:

- Hipòtesi nul·la: H_0: $\beta_1 = 0$, és a dir, la variable X no és explicativa
- Hipòtesi alternativa: H_1: $\beta_1 \neq 0$, és a dir, la variable X és explicativa

No rebutjar la hipòtesi nul·la vol dir que no es pot considerar el paràmetre β_1 significativament diferent de zero. És a dir, la variable X no té influència sobre la variable Y i, per tant, no hi ha una relació lineal entre totes dues variables.

Interpretació geomètrica

No rebutjar H_0 vol dir que la recta estimada té un pendent nul i, per tant, per a qualsevol valor de X la variable Y pren un mateix valor.

2) Fixem un nivell significatiu α.

3) Sota el supòsit de la hipòtesi nul·la certa ($\beta_1 = 0$), tenim l'**estadístic de contrast**:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

que correspon a una observació d'una distribució t de Student amb $n - 2$ graus de llibertat.

4) Finalment, podem actuar de dues maneres:

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

Recordem que...

... el p-valor és la probabilitat del resultat observat o d'un de més allunyat si la hipòtesi nul·la és certa.

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

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

- Si $|t| > t_{\alpha/2,\, n-2}$, es rebutja la hipòtesi nul·la H_0; per tant, hi ha una relació lineal entre les variables X i Y.
- Si $|t| \leq t_{\alpha/2,\, n-2}$, no es rebutja la hipòtesi nul·la H_0, per tant, no hi ha una relació lineal entre X i Y. Diem que la variable X és no explicativa.

Exemple de les alçades i els pesos

Continuant amb l'exemple de les alçades i els pesos, volem contrastar la hipòtesi nul·la que la variable X no és explicativa de la variable Y, és a dir, que el pendent de la recta de regressió és zero.

1) Establim les hipòtesis nul·la i alternativa:

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

2) Calculem l'estadístic de contrast: $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 3{,}202$

Segueix una distribució t de Student amb $n - 2 = 10 - 2 = 8$ graus de llibertat.

3) Establim un criteri de decisió a partir d'un nivell significatiu α fixat: si escollim un nivell significatiu de $\alpha = 0{,}05$:

a) A partir del p-valor: $P(|t| > 3{,}202) = 2P(t > 3{,}202) = 2 \cdot 0{,}0063 = 0{,}0126 < 0{,}05$; per tant, rebutgem la hipòtesi nul·la.

b) A partir del valor crític que és $t_{0{,}025;8} = 2{,}3060$, com que $3{,}202 > 2{,}306$ tenim la mateixa conclusió: rebutgem la hipòtesi nul·la i podem concloure que la variable alçada és explicativa del pes de les persones amb un 95% de confiança.

6. Resum

En aquesta sessió dedicada a la regressió lineal simple hem considerat que les nostres observacions sobre dues variables X i Y són una mostra aleatòria d'una població i que les fem servir per a extreure algunes conclusions del comportament de les variables sobre la població. Hem establert el model de regressió lineal amb les seves hipòtesis bàsiques més importants i hem vist com fer inferència sobre el pendent de la recta obtinguda a partir de la mostra i, en particular, com calcular un interval de confiança i com fer un contrast d'hipòtesi per a decidir si la variable X ens explica realment el comportament de la variable Y.

Exercicis

1.

El departament de personal d'una empresa informàtica dedicada a la introducció de dades ha fet un programa de formació inicial del personal. La taula següent indica el progrés obtingut en mecanografia de vuit estudiants que van seguir el programa, i el nombre de setmanes que fa que el segueixen:

Nombre de setmanes	Guany en velocitat (p.p.m.)
3	87
5	119
2	47
8	195
6	162
9	234
3	72
4	110

La recta de regressió calculada a partir d'aquestes dades és:

$$\hat{y}_i = 1{,}659 + 25{,}318x_i$$

a) Calculeu un interval de confiança del 95% pel pendent de la recta de regressió.

b) Feu un contrast d'hipòtesi amb un nivell de significació $\alpha = 0{,}05$, per saber si la variable "nombre de setmanes" és explicativa de la variable "guany de velocitat".

2.

Una botiga d'ordinadors va fer un estudi per a determinar la relació entre les despeses de publicitat setmanal i les vendes. Es van obtenir les dades següents:

Despeses en publicitat (*x* 1.000 €)	Vendes (*x* 1.000 €)
40	380
25	410
20	390
22	370
31	475
52	450
40	500
20	390
55	575
42	520

Amb aquestes dades s'han obtingut les quantitats següents:

$$\sum_{i=1}^{10} x_i = 347 \qquad \sum_{i=1}^{10} y_i = 4.460 \qquad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 6.018$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1.522{,}1 \qquad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 43.590{,}0$$

$$\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 23.793{,}66 \qquad \sum_{i=1}^{10} (y_i - \hat{y}_i)^2 = 19.796{,}34$$

I la recta de regressió: $\hat{y} = 308{,}8 + 3{,}95\ x$.

A partir de tota aquesta informació, calculeu un interval de confiança del 95% per al pendent.

Solucionari

1.

a) Interval de confiança:

Volem un interval de confiança del 95%, per tant, $\alpha = 0{,}05$ i mirant la taula de la *t* de Student per 6 graus de llibertat, tenim un valor crític de $t_{\alpha/2;n-2} = t_{0{,}025;6} = 2{,}4469$.

Com sempre, el primer que farem és una taula de càlculs adient amb el que ens demanen en aquest problema:

i	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	3	87	–2	4	77,61	9,39	88,116
2	5	119	0	0	128,25	–9,25	85,544
3	2	47	–3	9	52,30	–5,30	28,037
4	8	195	3	9	204,20	–9,20	84,695
5	6	162	1	1	153,57	8,43	71,115
6	9	234	4	16	229,52	4,48	20,061
7	3	72	–2	4	77,61	–5,61	31,506
8	4	110	–1	1	102,93	7,07	49,971
Σ	**40**	**1.026**	**35**	**44**			**459,045**

L'interval de confiança ve donat per:

$$[\hat{\beta}_1 - t_{\alpha/2,\, n-2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\, n-2} s_{\hat{\beta}_1}].$$

I ja estem en condicions de calcular cadascun d'aquests termes:

$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum (x_i - \bar{x})^2} = \frac{76{,}507}{44{,}0} = 1{,}739$$

on $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{459{,}045}{10-2} = 76{,}507$

Per tant: $s_{\hat{\beta}_1} = \sqrt{1{,}739} = 1{,}319$

I l'interval de confiança és:

$$[25{,}318 - 2{,}4469 \cdot 1{,}319;\ 25{,}318 + 2{,}4469 \cdot 1{,}319]$$

És a dir,

$$[22{,}092;\ 28{,}545]$$

b) Contrast d'hipòtesi per $\alpha = 0{,}05$:

1) Establim les hipòtesis nul·la i alternativa:

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

2) Calculem l'estadístic de contrast:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 19{,}200$$

Segueix una distribució t de Student amb $n - 2 = 6$ graus de llibertat.

3) Conclusió: com que per a $\alpha = 0{,}05$ tenim un valor crític $t_{0{,}025;6} = 2{,}4469$ més petit que l'estadístic de contrast $t = 19{,}200$, aleshores rebutgem la hipòtesi nul·la, de manera que el pendent és diferent de zero i la variable "nombre de setmanes" és explicativa del "guany de velocitat".

2.

L'interval de confiança ve donat per:

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

Ens fa falta calcular l'error estàndard del pendent i trobar els valors crítics.

1) Error estàndard del pendent:

Primer calculem:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{19.796{,}34}{10-2} = 2.474{,}54$$

de manera que:

$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum (x_i - \bar{x})^2} = \frac{2.474,54}{1.522,1} = 1,626$$

Per tant, l'error estàndard del pendent val: $s_{\hat{\beta}_1} = \sqrt{1,626} = 1,275$

2) Un interval de confiança del 95% amb $n = 10$, tenim uns valor crítics:

$$t_{0,025;8} = \pm 2,3060$$

3) Per tant, l'interval de confiança és:

$$[3,953 - 2,3060 \cdot 1,275;\ 3,953 + 2,3060 \cdot 1,275]$$

És a dir:

$$[1,013;\ 6,894]$$

Aquest interval de confiança no conté el valor zero, per tant, aquest resultat ens diu que la despesa en publicitat és explicativa de les vendes amb una confiança del 95%.

Annexos

Annex 1

a) Valor esperat de $\hat{\beta}_1$:

$$E(\hat{\beta}_1) = \beta_1$$

Manipulant una mica l'expressió que tenim per a $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}y_i$$

Si fem: $w_i = \dfrac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}$ podem escriure: $\hat{\beta}_1 = \sum_{i=1}^{n} w_i y_i$.

Si ara en calculem el valor esperat:

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} E(w_i y_i) = \sum_{i=1}^{n} w_i E(y_i) =$$

$$= \sum_{i=1}^{n} w_i(\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} w_i\beta_0 + \beta_1 w_i x_i =$$

$$= \sum_{i=1}^{n} w_i\beta_0 + \sum_{i=1}^{n}\beta_1 w_i x_i = \beta_0\sum_{i=1}^{n} w_i + \beta_1\sum_{i=1}^{n} w_i x_i$$

Propietat de la linealitat

La propietat de la linealitat de l'esperança d'una variable $E(kX) = kE(X)$.

Veiem que: $\sum_{i=1}^{n} w_i = 0$ i que $\sum_{i=1}^{n} w_i x_i = 1$

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

Observació

Com que:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

és fàcil veure que:

$$\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

Per a calcular el terme $\sum_{i=1}^{n} w_i x_i$, farem servir la igualtat següent:

$$\sum_{i=1}^{n} w_i(x_i-\bar{x}) = \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i\bar{x} = \sum_{i=1}^{n} w_i x_i - \bar{x}\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i x_i$$

Ja que: $\sum_{i=1}^{n} w_i = 0$

De manera que:

$$\sum_{i=1}^{n} w_i(x_i - \bar{x}) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}(x_i - \bar{x}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 1$$

Així, doncs: $\sum_{i=1}^{n} w_i x_i = 1$

I, finalment, tenim que: $E(\hat{\beta}_1) = \beta_1$

b) Variància de $\hat{\beta}_1$:

$$\sigma_{\hat{\beta}_1} = Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} Var(w_i y_i) = \sum_{i=1}^{n} w_i^2 Var(y_i) =$$

$$= \sigma^2 \sum_{i=1}^{n} w_i^2 = \sigma^2 \left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 = \sigma^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Propietat de la variància

$Var(kX) = k^2 Var(X)$

Tenim que la variància de $\hat{\beta}_1$ és: $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.