

LINEAR REGRESSION: LEAST SQUARES

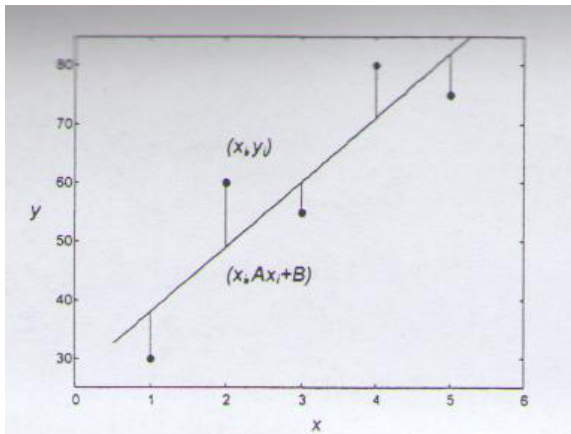
Se il coefficiente di correlazione r è prossimo a 1 o a -1 e se il diagramma di dispersione suggerisce una relazione di tipo lineare, ha senso determinare **l'equazione di una retta che approssimi “nel modo migliore” i dati.**

Sia dato un insieme di n punti $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, si vuole determinare la retta

$$y = Ax + B$$

che meglio “approssima” questi punti.

In questa figura si osserva la retta $y = Ax + B$ ed i punti $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. I segmenti in figura hanno lunghezza $|Ax_i + B - y_i|$ e rappresentano l'errore che si fa a voler rappresentare i dati con una singola retta (e quindi con i soli due parametri A e B).



Se i punti fossero perfettamente allineati, si potrebbe scegliere l'unica retta che passa per tutti i punti e si renderebbe quindi nullo l'errore.

In generale si può quantificare l'errore totale E che si commette sommando i quadrati delle lunghezze di tutti i segmenti ossia

$$E(A, B) = \sum_{i=1}^n (Ax_i + B - y_i)^2$$

Per ogni possibile scelta della retta, ossia della scelta dei parametri A e B , l'errore è differente.

Il criterio che generalmente viene usato per definire il “modo migliore” è detto **metodo dei minimi quadrati** e consiste nel minimizzare la quantità $E(A, B)$ rispetto ad A e B .

MINIMI QUADRATI

Definizione. La **retta di regressione lineare** è la retta di equazione

$$y = Ax + B$$

che rende minimo l'errore $E(A, B)$ dove

$$E(A, B) = \sum_{i=1}^n \left(Ax_i + B - y_i \right)^2$$

In altre parole, si devono determinare A e B in modo che $E(A, B)$ sia il minimo possibile.

Si noti che $E(A, B)$ è una quantità non negativa che può assumere il valore nullo soltanto quando tutti i punti sono perfettamente allineati.

Se i dati sono perfettamente allineati ossia **esattamente** linearmente correlati:

$$y_i = ax_i + b,$$

si avrà:

$$E(A, B) = \sum_{i=1}^n \left(Ax_i + B - y_i \right)^2 = \sum_{i=1}^n \left(Ax_i + B - ax_i + b \right)^2$$

scegliendo

$$A = a, \quad B = b$$

si ottiene $E = 0$ che è necessariamente il minimo di E .

In altre parole, se i dati sono esattamente disposti su una retta, questa **coincide** con la retta di regressione lineare.

In generale, per trovare il minimo di $E(A, B)$ basta calcolare le due derivate parziali rispetto ad A e B e imporre che siano nulle.

Imponendo che la derivata rispetto ad A sia nulla si ottiene:

$$A \sum_{i=1}^n x_i^2 + B \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i = 0$$

mentre imponendo che la derivata rispetto ad B sia nulla si ottiene:

$$A \sum_{i=1}^n x_i + nB - \sum_{i=1}^n y_i = 0$$

If we use the following four equalities (that the student should easily find out from the definitions of mean, variance and covariance)

$$\sum_{i=1}^n x_i = n\bar{x}, \quad \sum_{i=1}^n y_i = n\bar{y},$$

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

the previous equations can be rewritten as

$$\mathbf{A(n-1)s_x^2 + An\bar{x}^2 + Bn\bar{x} - (n-1)s_{xy} - n\bar{x}\bar{y} = 0},$$

$$\mathbf{B = \bar{y} - A\bar{x}}$$

By substitution of the second equation in the first, we obtain

$$\mathbf{A(n-1)s_x^2 - (n-1)s_{xy} = 0},$$

da cui si ottiene:

$$A = \frac{s_{xy}}{s_x^2} = r \frac{s_y}{s_x}$$

dove s_{xy} = covarianza, s_x = deviazione standard di x , s_y = deviazione standard di y e r = coefficiente di correlazione. Abbiamo inoltre già visto che

$$B = \bar{y} - A\bar{x}$$

dove $\bar{x}$ = media campionaria di x e $\bar{y}$ = media campionaria di y .

CASO LINEARE

Abbiamo visto che nel caso in cui i punti siano del tutto allineati, ossia $y_i = ax_i + b$, si ha $r = \text{sign}(a)$, si ha inoltre $s_y^2 = a^2 s_x^2$ e quindi $s_y = |a|s_x$.

In definitiva:

$$A = r \frac{s_y}{s_x} = \text{sign}(a) \cdot |a| = a$$

mentre

$$B = \bar{y} - A\bar{x} = \bar{y} - a\bar{x} = b$$

La retta di regressione è quindi $\bar{y} = a\bar{x} + b$ ossia la retta che passa per i tutti dati.

Esempio (dati non allineati).

Acquisto di un'automobile usata.

Anni (x)	9	12	8	15	6
Costo (y)	18	8	17	7	20

Avevamo già calcolato le medie campionarie

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = \frac{9 + 12 + 8 + 15 + 6}{5} = 10$$

$$\bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = \frac{18 + 8 + 17 + 7 + 20}{5} = 14$$

Avevamo già calcolato anche la covarianza

$$\begin{aligned}s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\&= \frac{(-1)(4) + (2)(-6) + (-2)(3) + (5)(-7) + (-4)(6)}{4} = \\&= \frac{-4 - 12 - 6 - 35 - 24}{4} = -20,25\end{aligned}$$

Si ricordi che la covarianza è negativa perché i dati sono ordinati in modo approssimativamente decrescente.

Calcoliamo ora

$$\begin{aligned}s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \\&= \frac{(-1)^2 + (2)^2 + (-2)^2 + (5)^2 + (-4)^2}{4} = = \mathbf{12,5}\end{aligned}$$

e, per completezza, calcoliamo anche

$$\begin{aligned}s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \\&= \frac{(4)^2 + (-6)^2 + (3)^2 + (-7)^2 + (6)^2}{4} = = \mathbf{36,5}\end{aligned}$$

Abbiamo quindi tutti i dati necessari

$$\bar{x} = 10, \quad \bar{y} = 14, \quad s_{xy} = -20,25$$

ed anche

$$s_x^2 = 12,5, \quad s_x = 3,54, \quad s_y^2 = 36,5, \quad s_y = 6,04$$

E per completezza possiamo anche ricalcolare il coefficiente di correlazione

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-20,25}{3,54 \cdot 6,04} = -0,948$$

La correlazione negativa è molto forte (il coefficiente è prossimo a -1), ha quindi senso interpolare linearmente (cercare la retta di regressione lineare).

Abbiamo quindi

$$A = \frac{s_{xy}}{s_x^2} = \frac{-20,25}{12,5} = -1.62$$

oppure possiamo usare

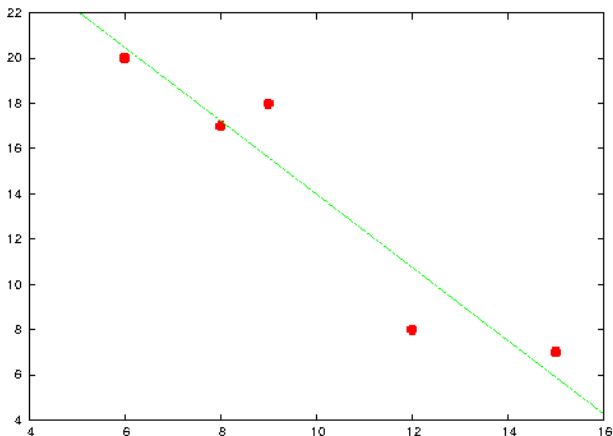
$$A = r \frac{s_y}{s_x} = -0,948 \cdot \frac{3,54}{6,04} = -1.62$$

Il parametro B è invece

$$B = \bar{y} - A\bar{x} = 14 + 1,62 \cdot 10 = 30,2$$

Quindi la retta di regressione lineare è

$$y = -1.62x + 30,2$$



La retta di regressione lineare è $y = -1.62x + 30,2$.
In generale, più $|r|$ è prossimo al valore 1, più la retta di regressione lineare *rappresenta* in modo corretto i dati.

LEAST SQUARES. NON-LINEAR REGRESSION

In many cases, after plotting experimental data points $(x_1, y_1), \dots, (x_n, y_n)$ on a scatter plot, a correlation between the two observed variables may be evident, yet not of a linear nature,. One can try to use higher-degree polynomial as the approximating function. The simplest case (after a straight line), is a second-degree polynomial:

$$y = Ax^2 + Bx + C,$$

with the three parameters A , B and C determined by minimizing

$$E(A, B, C) = \sum_{i=1}^n \left(Ax_i^2 + Bx_i + C - y_i \right)^2.$$

The minimization can be achieved by derivation of the above expression with respect each of the three parameters and it yields

$$\sum_{i=1}^n \left(Ax_i^2 + Bx_i + C - y_i \right)^2 x_i^2 = 0$$

$$\sum_{i=1}^n \left(Ax_i^2 + Bx_i + C - y_i \right)^2 x_i = 0$$

$$\sum_{i=1}^n \left(Ax_i^2 + Bx_i + C - y_i \right)^2 = 0$$

which has a unique solution (except all x_i are equal, which means all data on a vertical straight line).

Let us consider the following five points: $(-3, 3)$, $(-2, 2)$, $(0, 1)$, $(2, 1)$, $(4, 3)$, the system of equations for the three parameters is

$$369A + 37B + 33C = 87$$

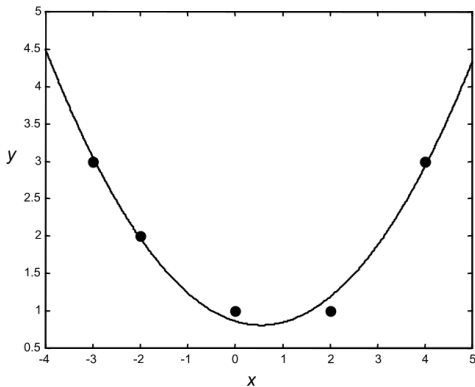
$$37A + 33B + C = 1$$

$$33A + B + 5C = 10$$

whose solution is

$$A = .0,178, \quad B = -0.196, \quad C = 0.863.$$

The scatterplot and the regression parabola can be seen in the figure below



LEAST SQUARES. LINEARIZATION METHODS

Suppose again that the scatterplot of the experimental data points $(x_1, y_1), \dots, (x_n, y_n)$ show a non-linear correlation between the two observed variables and also suppose that you strongly suspect that the relation between the two variables has a specific nature as, for example,

$$y = Bx^A,$$

taking the logarithm of both side you obtain

$$\ln y = \ln B + A \ln X.$$

then you can define $y' = \ln y$, $x' = \ln x$, $B' = \ln B$ and you get

$$y' = Ax' + B'.$$

Then, the strategy is the following; you define for all points $(x'_i, y'_i) = (\ln x_i, \ln y_i)$ and apply to these redefined variables the ordinary linear regression method in order to obtain the optimal values A and B' . Then you obtain $B = \exp B'$ and then you plot the original data (x_i, y_i) against the function $y = Bx^A$.

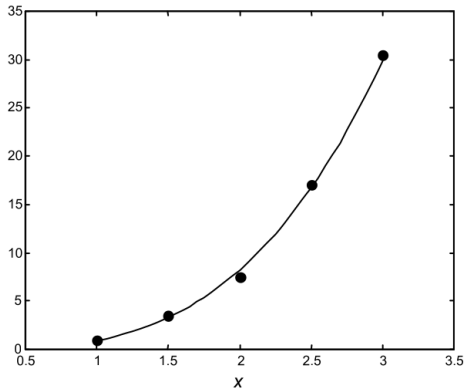
For example, if you use the following data

$(1.0, 0.9), (1.5, 3.5), (2, 7.5.0), (2.5, 17.0), (3.0, 30.5),$

you obtain the following values for the parameters:

$$A = 3.181, B' = -0.0979, \rightarrow A = 3.181, B' = 0.907.$$

The scatterplot and the regression function can be seen in the figure below



Suppose now that you strongly suspect that the relation between the two variables has the following nature

$$y = B \exp(Ax),$$

if you take the logarithm of both sides you obtain

$$\ln y = \ln B + Ax,$$

which become

$$y' = Ax + B'.$$

if you define $y' = \ln y$ and $B' = \ln B$. You can procede as in previous example and obtain the scatterplot. Try, for example, with the following data:

$$(-1.0, 6.7), (0.1, 3.9), (1.0, 2.1), (2.0, 1.3), (3.0, 0.9).$$