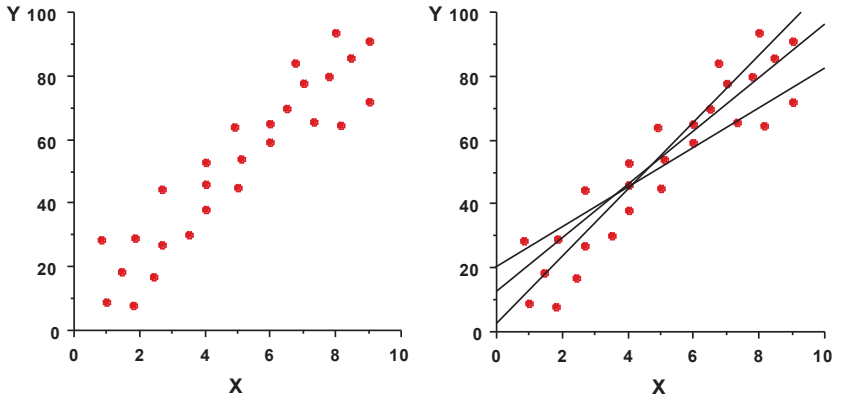
LA REGRESSIONE LINEARE SEMPLICE

Se due variabili X e Y sono tra loro correlate, e sono entrambe su scala a intervalli o rapporti equivalenti, la tecnica statistica della regressione lineare consente di calcolare il valore di Y previsto (o stimato), dato un certo valore di X. Si parla di regressione semplice quando i punteggi di Y (Variabile Dipendente) sono stimati sulla base di un'unica Variabile Indipendente X (vedi Nota).

Un esempio che tutti conosciamo è quello delle tabelle del peso ideale, previsto in base all'altezza della persona. Tra altezza e peso esiste una correlazione positiva, per cui le persone più alte pesano in genere di più e quelle più basse pesano di meno. La correlazione ovviamente non è perfetta, perché le variazioni nel peso delle persone non sono determinate soltanto dall'altezza, ma anche da altri fattori quali il genere, l'età, le abitudini alimentari, le differenze individuali ecc. (se la relazione fosse perfetta tutte le persone di una certa altezza avrebbero esattamente lo stesso peso). Comunque una buona percentuale della variabilità nel peso delle persone è spiegata dall'altezza (questa percentuale è indicata dal coefficiente di determinazione); l'esistenza di questa relazione permette, sulla base della misurazione di peso e altezza di un elevato numero di persone, di individuare una formula (detta equazione di regressione) che consente di stabilire qual è l'incremento medio del peso, in etti, per ogni centimetro di altezza in più. Sulla base di questa formula, quindi, si può calcolare il peso "previsto" per persone di altezza diversa.

N.B. La regressione delle Y sulle X si riferisce alla previsione dei punteggi stimati Y' a partire dai punteggi X. In questo caso X è considerata la Variabile Indipendente e Y la variabile Dipendente. Dal punto di vista statistico è irrilevante quale delle due variabili sia considerata indipendente e quale dipendente, tuttavia non sempre lo è dal punto di vista concettuale. Ad esempio nel caso della relazione tra peso e altezza, è sensato dire che il peso dipende dall'altezza, ma non ha senso sostenere il contrario.

L'equazione di regressione, nel caso della regressione lineare (l'unica di cui tratteremo), è l'equazione di una retta, detta retta di regressione. Quando la correlazione tra X e Y è perfetta, ad ogni valore osservato di X corrisponde un solo valore osservato di Y, quindi i punti del diagramma di dispersione si dispongono in modo da formare esattamente una retta.
Questo accade molto raramente nella realtà ed in genere se esiste una correlazione lineare tra X e Y i punti si dispongono intorno ad una retta ipotetica che rappresenta l'insieme dei valori di Y stimati o previsti (indicati con Y') a partire dai valori di X. Ma, come si può osservare nel grafico che segue, le rette che passano attraverso i punti di un diagramma di dispersione relativo a due variabili X e Y tra loro correlate, possono essere numerose (nel grafico ne sono state tracciate tre), qual è tra le tante possibili quella che meglio descrive i dati?



La retta "migliore", quella che meglio descrive la relazione tra X e Y è ovviamente quella che passa mediamente più vicina a tutti i punti del diagramma di dispersione, o in altre parole, quella che rende minime le distanze tra i valori Y (punti del diagramma) e i valori Y' (punti sulla retta, stimati sulla base dalla regressione delle Y sulle X). La retta di regressione è quindi chiamata retta dei minimi quadrati, in quanto rende minima la somma dei quadrati delle differenze Y-Y'.
Le procedure di calcolo della regressione consentono di

individuare la retta che possiede questa caratteristica.
L'equazione della retta di regressione è la seguente:

$Y' = a + bX$ dove X è un valore qualsiasi della Variabile Indipendente (V.I.), Y' è il corrispondente valore di Y stimato (previsto) in base alla regressione delle Y sulle X, e *a* e *b* sono i parametri della retta:

b è il coefficiente di regressione ed indica di quante unità aumenta (o diminuisce) Y all'aumentare di una unità di X;
b determina l'inclinazione della retta, cioè l'angolo che la retta forma con l'asse delle ascisse, infatti a valori elevati di *b* corrisponde un maggiore angolo di inclinazione della retta, e viceversa. Se la correlazione tra X e Y è negativa, anche *b* è negativo. La formula di calcolo per *b* è la seguente:

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

a è detta intercetta, valore predetto Y' per X=0.
Si ottiene con la formula: $a = Y - bX$

Un altro indice molto importante è il coefficiente di determinazione r^2 che si calcola elevando al quadrato il coefficiente di correlazione r di Pearson. Moltiplicato per 100 fornisce la percentuale di **devianza** di Y spiegata da X. La devianza non è altro che il numeratore della varianza, cioè:

$$\sum (Y - \bar{Y})^2$$

La formula si riferisce alla devianza totale della variabile Y, che si scompone in devianza spiegata dalla regressione e devianza residua. Nell'esempio relativo alla relazione tra altezza (X) e peso (Y) un coefficiente di determinazione $r^2 = 0.49$ indicherebbe che il 49% della **devianza totale** del peso è determinato dall'altezza (**devianza spiegata**), mentre il rimanente 51% (**devianza residua**) sarebbe determinato da altri fattori (i valori riportati sono solo ipotetici).

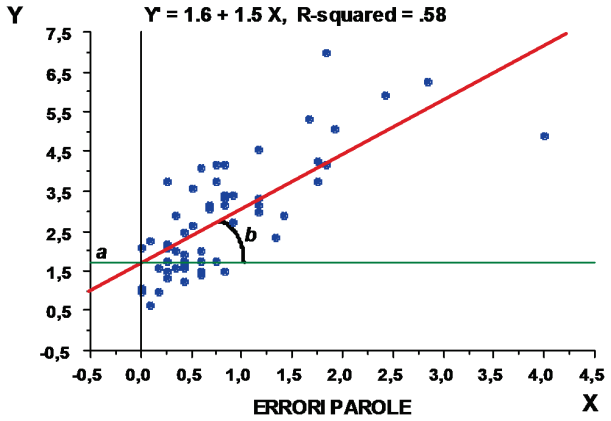
Calcolo dell'equazione di regressione - Esempio

Le Non Parole sono stringhe di lettere senza significato che somigliano a parole reali, per esempio LIPOZIA, costruita anagrammando la parola POLIZIA. Se si chiede a dei bambini di leggere a voce alta una lista di Parole ed una lista di Non Parole con caratteristiche simili, i bambini tenderanno a commettere più errori sulle Non Parole rispetto alle Parole, sia che commettano in genere pochi errori di lettura, sia che ne commettano molti. Una percentuale di errori sulle Non Parole molto elevata rispetto alle Parole, è tuttavia considerata indice di un disturbo di lettura **di tipo fonologico**, dovuto ad una difficoltà ad usare le regole di conversione grafema-fonema. Si vogliono quindi determinare i criteri entro i quali una differenza negli errori di lettura tra Parole e Non Parole può essere definita "normale", in altri termini, si vuole rispondere alla seguente domanda:
dato un numero X di errori commessi sulle Parole, qual'è il numero Y di errori sulle Non Parole previsto in condizioni di normalità?

Per rispondere a questa domanda un gruppo di ricercatori somministra a 60 bambini di età compresa tra i 6 e gli 8 anni una prova di lettura a voce alta di Parole e Non Parole, costituita da 6 liste di Parole e 6 liste di Non Parole; ogni lista comprende 12 stimoli. Viene registrato il numero medio di errori per lista, calcolato separatamente per le Parole (X) e le Non Parole (Y). Viene condotta una analisi della regressione lineare, considerando X come Variabile Indipendente e Y come Variabile Dipendente.

Risultati (dati reali):
I parametri della retta sono:
a (intercetta) = 1.6; *b* (coefficiente di regressione) = 1.5.
L'equazione della retta di regressione (Y' = *a* +*b* X) è quindi:
Y' = 1.6 + 1.5 X
Il valore ***b*** = 1.5 indica che per ogni errore in più commesso sulle Parole, il numero stimato di errori sulle Non Parole aumenta di 1.5 unità. Infatti:
per X = 0 ---> Y' = 1.6 + 1.5 x 0 = 1.6 (valore di ***a***);
per X = 1 ---> Y' = 1.6 + 1.5 x 1 = 3.1
per X = 2 ---> Y' = 1.6 + 1.5 x 2 = 4.6 e così via;
dato un qualsiasi valore di X, è possibile calcolare il corrispondente valore Y' previsto in base alla regressione.

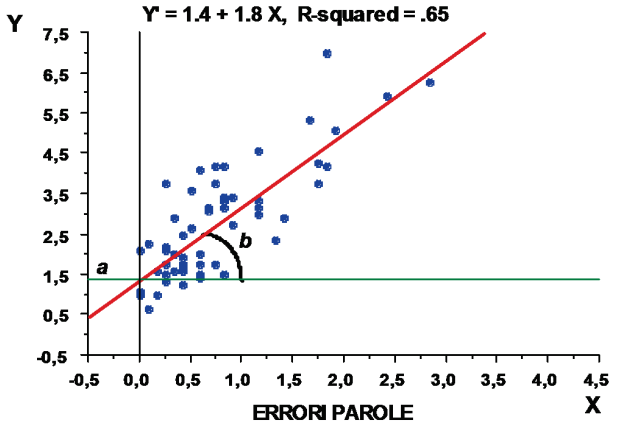
Diagramma di dispersione e retta di regressione (è riportato anche il coefficiente di determinazione = 0.58):



La retta identifica i valori Y' previsti, o stimati, per i diversi valori di X, mentre i punti del grafico rappresentano i valori osservati di Y per i diversi valori di X.
Il valore di ***a*** è contrassegnato da una linea orizzontale, viene anche indicato l'angolo ***b*** che la retta forma con questa linea (o con qualsiasi altra linea parallela all'asse delle ascisse).

Nel grafico si evidenzia un punto all'estrema destra che si discosta dal resto del campione (dati di questo genere vengono chiamati *outliers* o *valori erratici*).

Il grafico successivo si riferisce alla regressione calcolata sugli stessi dati, dopo aver tolto i valori X e Y del bambino cui si riferisce il punto in questione:



L'esclusione del dato *erratico* produce una maggiore inclinazione della retta, infatti ***b*** passa da 1.5 a 1.8, inoltre aumenta anche il coefficiente di determinazione, da 0.58 a 0.65.
La capacità di previsione dei valori Y' è migliorata, in quanto è aumentata la percentuale di devianza spiegata dalla regressione (dal 58% al 65%). La nuova equazione di regressione è la seguente: **Y' = 1.4 + 1.8 X**
A questo punto è possibile rispondere alla domanda posta inizialmente: dato un numero X di errori commessi sulle Parole, qual'è il numero Y di errori sulle Non Parole previsto in condizioni di normalità?
Se un bambino commette, ad esempio, 10 errori sulle Parole, il numero di errori che ci aspettiamo commetta sulle Non Parole può essere calcolato sostituendo il valore 10 a X nell'equazione di regressione, quindi: **Y'** = 1.4 + 1.8 x 10 = 1.4 + 18 = **19.4**.

L'errore standard della stima s_e

Il valore Y' calcolato mediante l'equazione di regressione è il valore previsto per la variabile Y, dato un certo valore di X. Ovviamente la previsione effettuata comprende una componente di errore (a meno che la correlazione tra X e Y sia esattamente 1 o –1). L'entità di questa componente di errore dipende dalla dispersione dei valori Y osservati, intorno alla retta di regressione, cioè dalla distanza che esiste, in media, tra i valori Y (punti del diagramma di dispersione) ed relativi i valori Y' (i punti sulla retta). Se si sommano i quadrati delle differenze Y-Y' si ottiene la **devianza residua**, cioè quella parte di variabilità di Y che non è spiegata dalla regressione. In formula:

Devianza residua = $\Sigma(Y-Y')^2$

Dividendo la devianza residua per N (numerosità del campione), si ottiene un indice medio di dispersione dei punteggi intorno alla retta, cioè la **varianza residua**; estraendo la radice quadrata di questo indice si ottiene l'**errore standard della stima s_e**, in formula:

$$S_e = \sqrt{\frac{\sum(Y - Y')^2}{N-2}}$$

L'errore standard della stima indica di quanti punti in media si può sbagliare prevedendo Y' a partire da X. Maggiore è la dispersione dei punteggi Y intorno alla retta, maggiore sarà la devianza residua e maggiore quindi l'errore di previsione.

Nella prima analisi effettuata su 60 bambini, il valore dell'errore standard della stima era **s_e = 0.93**; nella seconda analisi, su 59 bambini, si è ottenuto, invece, **s_e = 0.84**. Eliminando il valore erratico è diminuita la dispersione dei dati intorno alla retta, ciò ha provocato un aumento della devianza spiegata (dal 58% al 65%) una conseguente diminuzione della devianza residua (dal 42% al 35%) ed anche l'errore di previsione s_e è diminuito di conseguenza.

Uso dei risultati di una regressione lineare per valutare la prestazione di un singolo soggetto

Nell'esempio descritto è stata utilizzata la tecnica della regressione lineare per ottenere una equazione di regressione che consentisse di calcolare il numero di errori sulle Non Parole previsto, in condizioni di normalità, dato un certo numero di errori sulle Parole.
Vediamo ora, riferendoci sempre allo stesso esempio, come possono essere utilizzati i risultati di analisi condotte da altri autori (riportate su articoli di ricerca o manuali di test) per esaminare la prestazione di un bambino che abbia le stesse caratteristiche del campione di riferimento.
Immaginiamo di esaminare mediante un test di lettura di Parole e Non Parole un bambino di 7 anni, ed ammettiamo che il manuale del test riporti i risultati della regressione così come descritti per il campione di 59 bambini.
Il bambino da noi esaminato commette 4 errori nella lettura di Parole (X) e 10 errori nella lettura di Non Parole (Y). Osserviamo che il numero di errori sulle Non Parole è superiore a quello sulle Parole, ma non sappiamo se questa differenza può essere considerata nella norma o può suggerire la presenza di un problema di lettura di tipo fonologico.
Utilizzando l'equazione di regressione riportata nel manuale del test, calcoliamo il numero di errori previsto per le Non Parole per X = 4, utilizzando l'equazione di regressione
Y' = 1.4 + 1.8 x 4 per cui Y' = 1.4 + 7.2 = **8.6**
Il numero di errori previsto sulle Non Parole per un bambino che commette 4 errori sulle parole è 8.6; il bambino in questione ha commesso 10 errori sulle Non Parole, quindi un numero superiore rispetto al previsto. La differenza (detta **residuo**) tra il valore osservato Y ed il valore atteso Y' può essere considerata significativa?

Poiché i residui relativi a variabili con distribuzione normale si distribuiscono anch'essi normalmente, il residuo può essere standardizzato e trasformato in punto z, il punto z calcolato sarà quindi confrontato con il valore critico di z al 5% che corrisponde a **1.64**. Per standardizzare la differenza Y - Y' è sufficiente dividerla per l'**errore standard della stima**, che, ricordiamo, è un indice di variabilità (dispersione) dei punteggi osservati Y intorno alla retta di regressione (valori Y'). Dalla precedente analisi il valore dell'errore standard della stima (s_e) risultava uguale a 0.84.
Si procede quindi al calcolo del punto z:

$$z = \frac{Y - Y'}{s_e} = \frac{10 - 8.6}{0.84} = 1.67$$

Poiché z calcolato (1.67) risulta superiore a z critico (1.64), si può affermare che la differenza è significativa. Il numero di errori commessi dal bambino sulle Non Parole è quindi **significativamente** superiore al numero di errori previsto in base alla sua prestazione sulle Parole. Questo dato suggerisce un disturbo di lettura di tipo fonologico.