Lezione 7

La regressione semplice

Argomenti della lezione:

- **Il modello teorico**
- **Il calcolo dei parametri**

Regressione lineare

Esamina la relazione lineare tra una o più variabili esplicative (o indipendenti, o "predittori") e una variabile criterio (o dipendente)

Duplice scopo:

→ **Esplicativo** → **Predittivo**

Conoscere l'esatta forma della relazione

Trovare un'equazione che permetta di predire quanti incidenti potrebbero capitare ad una persona, conoscendo il suo punteggio di nevroticismo

Regressione ⇒ previsione di un valore sconosciuto di una variabile (Y) in base al valore conosciuto di un'altra variabile (X)

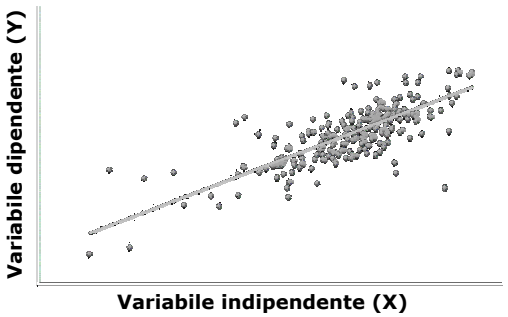
Trovare l'equazione che esprime Y in termini (cioè in funzione) di X

La regressione bivariata (o semplice)

- **Una sola variabile indipendente (VI) sulla quale "regredisce" la variabile dipendente (VD)**
- **Si ipotizza che la VI "determini" o "influenzi" o "predica" la VD**

Individuare la retta che consente di prevedere al meglio i punteggi nella VD da quelli nella VI

Individuare la retta che "interpola" meglio la nuvola di punti (o "scatterplot") della distribuzione congiunta delle due variabili



Forma della relazione: lineare

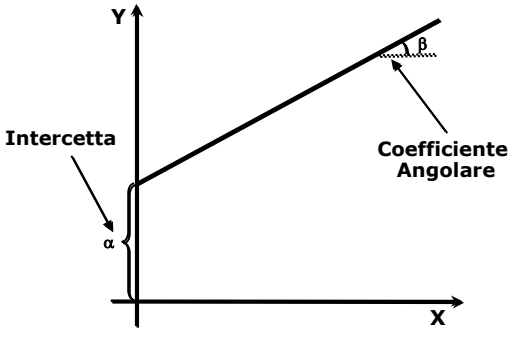
È la relazione più parsimoniosa, e più realistica in moltissimi casi

Equazione che lega Y a X:

$$Y = \alpha + \beta X$$

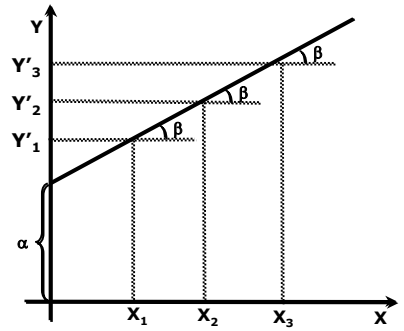
Parametri dell'equazione:

- α **(intercetta)**
- β **(coefficiente angolare)**



Linearità

Per ogni variazione in X si determina sempre la stessa variazione in Y qualunque sia il valore di X sull'asse delle ascisse

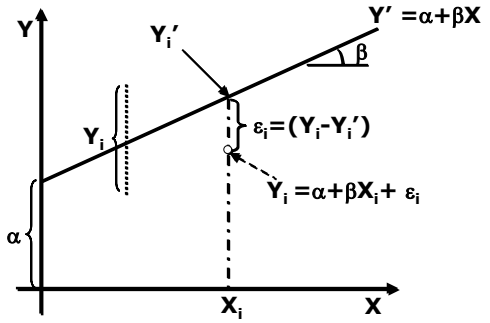


$(X_3-X_2) = (X_2-X_1) \Rightarrow (Y'_3-Y'_2) = (Y'_2-Y'_1)$

Le relazioni tra le variabili non sono perfette. I punti sono dispersi intorno alla retta di regressione. L'equazione deve incorporare un termine di errore (o residuo) per ogni caso

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

- **$Y' = \alpha + \beta X$:**
  valore "teorico" della Y, valore che si ottiene tramite l'equazione di regressione. Parte fissa.
- **"e":**
  deviazione del punteggio osservato Y dal punteggio teorico Y'. Parte variabile.



- **Identificazione della retta di regressione e calcolo dei parametri**
- **Stimare i valori dei parametri della popolazione, α e β, tramite i dati osservati su un campione (a, b)**
- **Identificare la retta che meglio si adatta ai punti che descrivono la distribuzione delle Y sulle X**

Criterio dei minimi quadrati

La retta che interpola meglio il diagramma di dispersione, cioè quella retta che passa più vicina possibile alla nuvola dei punti, è quella che rende minima la somma delle differenze al quadrato tra le Y osservate e le Y' teoriche

Equazione dei minimi quadrati:

$$\Sigma(Y_i - Y_i')^2 = \Sigma(Y - (a + bx))^2 = min$$

Riduce al minimo l'errore commesso nello stimare Y da X

Formule per il calcolo di a e b derivate dall'analisi numerica:

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{\Sigma (X-\bar{X})^2} = \frac{cov(X,Y)}{Var(X)}$$

$$b = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

Calcolo della retta di regressione

Sogg	X	Y	XY	X^2	Y^2
a	1	10	10	1	100
b	2	12	24	4	144
c	5	50	300	35	2500
d	3	30	90	9	900
e	6	62	372	36	3844
f	7	60	420	49	3600
g	4	45	180	16	2025
Tot	29	269	1396	151	13113

Calcolo del coefficiente angolare b:

$$b = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

$$b = \frac{7*1396-29*269}{7*151-29^2} = \frac{9772-7801}{1057-841} = 9.125$$

Calcolo dell'intercetta a:

$$a = \bar{Y} - b\bar{X} \qquad a = 38.4-(9.125)*4.1 = 0.99$$

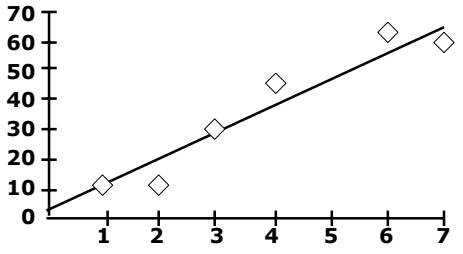
$$Y' = 0.99 + 9.125\ X$$

Rappresentazione grafica della retta

Basta calcolare due soli Y' per due valori X, e tracciare la retta che unisce i due punti (Y'_1, X_1) e (Y'_2, X_2)

Scegliamo $X_1 = 0$ e $X_2 = 7$

$$X_1 = 0 \Rightarrow Y'_1 = 0.99 + 9.125*0 = 0.99$$
$$X_2 = 7 \Rightarrow Y'_2 = 0.99 + 9.125*7 = 63.9$$



Il coefficiente di regressione esprime la relazione tra Y e X nell'unità di misura delle 2 variabili

Per esprimere questa relazione in una scala di misura comune si deve standardizzarlo

Il coefficiente di regressione standardizzato = "peso beta", $\beta^{\wedge}$

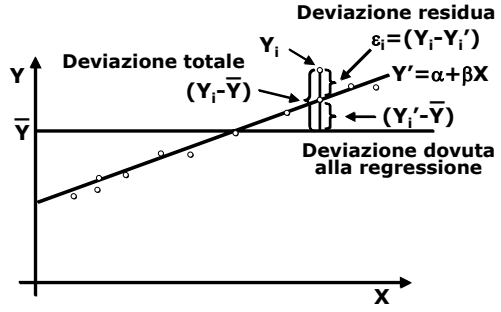
$\beta^{\wedge}$ si ottiene moltiplicando il coefficiente non standardizzato per il rapporto delle deviazioni standard della VI e della VD:

$$\beta^{\wedge} = b\,(s_x/s_y)$$

Nella regressione semplice è uguale al coefficiente di correlazione

$$\beta^{\wedge} = r_{yx}$$

Adeguatezza della equazione di regressione



Dalle deviazioni alle somme dei quadrati

$$(Y_i - \bar{Y}) = (Y_i' - \bar{Y}) + (Y_i - Y_i')$$

$$\Sigma(Y_i - \bar{Y})^2 =$$

$$\Sigma(Y_i' - \bar{Y})^2 + \Sigma(Y_i - Y_i')^2$$

- $\Sigma(Y_i - \bar{Y})^2$ **è la devianza totale delle Y_i dalla loro media**
- $\Sigma(Y_i' - \bar{Y})^2$ **è la devianza di Y_i dalla media che è spiegata dalla regressione**
- $\Sigma(Y_i - Y_i')^2$ **è la devianza di Y_i dalla media che non è spiegata dalle regressione**

È possibile dimostrare che:

$$r^2 = \Sigma(Y_i' - \bar{Y})^2/\Sigma(Y_i - \bar{Y})^2 =$$

Dev. Spiegata/ Dev. Totale

Dividendo i due termini per n:

r^2 = Var. Spiegata/ Var. Totale

L'indice r^2 viene definito coefficiente di determinazione

$(1-r^2)$ indica la proporzione della varianza totale di Y che non è spiegata dalla regressione

$(1-r^2) = \Sigma(Y_i - Y_i')^2/\Sigma(Y_i - \bar{Y})^2$ = Dev. Residua/ Dev. Totale

Dividendo i due termini per n: $(1-r^2)$ = Var. Residua/ Var. Totale

La radice quadrata di $(1-r^2)$ viene definita coefficiente di alienazione

Da $(1-r^2)$ è possibile ricavare il coefficiente che rappresenta la varianza intorno alla retta di regressione per ogni valore di X:

$$S^2_e = (1-r^2)\, S^2_y$$

Deviazione standard degli errori:

"errore standard della stima"

Indice della precisione della retta di regressione

$$S_e = \sqrt{(1-r^2)\, S_y} = \sqrt{\frac{\Sigma(Y-Y')^2}{N-2}}$$

$r = 0$, $S_e = S_y \Rightarrow$ la varianza d'errore coincide con la varianza totale di Y

$r = 1$, $S_e = 0 \Rightarrow$ tutti gli Y cadono sulla retta di regressione Y'

Calcolo dell'errore standard della stima

Sogg	X	Y	Y'	(Y-Y')	$(Y-Y')^2$
a	1	10	10.11	-0.115	0.013
b	2	12	19.24	-0.724	52.42
c	5	50	54.75	-4.75	22.56
d	3	30	28.35	1.635	2.673
e	6	62	54.75	7.25	52.56
f	7	50	54.85	-4.865	23.67
g	4	45	37.49	7.51	56.40
Tot					210.30

$$S_e = \sqrt{\frac{\Sigma(Y-Y')^2}{N-2}} = \sqrt{\frac{210.3}{5}} = 6.48$$

Più S_e è piccolo, meglio la retta di regressione predice i valori Y da quelli di X

CONCLUSIONE

- Regressione bivariata
- Criterio dei minimi quadrati
- Calcolo dei parametri
- Adeguatezza della soluzione