DEFINIZIONE DI COVARIANZA CAMPIONARIA (di un insieme di dati accoppiati)

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

dati raggruppati: $s_{xy} \approx \frac{1}{n-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} f_a(i,j)(\overline{x}_i - \overline{x})(\overline{y}_j - \overline{y})$

DEFINIZIONE DI COEFFICIENTE DI CORRELAZIONE (di un insieme dati accoppiati)

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{k=1}^{n}(y_k - \overline{y})^2}}$$

x_i *predittore* $\qquad$ y_i *responso*

DEFINIZIONE DI VALORE PREVISTO

Data la retta dei minimi quadrati $y=b_0+b_1 x$, il **valore previsto** di y per $x = x_i$ è

$$\hat{y}_i = b_0 + b_1 x_i.$$

DEFINIZIONE DI RESIDUO

Se y_i sono i valori osservati e $\hat{y}_i$ quelli previsti, le quantità seguenti sono dette **residui**

$$r_i = y_i - \hat{y}_i.$$

DEFINIZIONE DI R^2

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_n)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_n)^2}.$$

devianza totale $DT = \sum_{i=1}^{n}(y_i - \overline{y})^2$
devianza spiegata $DS = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$
devianza dei residui $DR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Vale la proprietà $DT = DS + DR$. $\qquad R^2 = \frac{DS}{DT} = 1 - \frac{DR}{DT}$

Si noti che nel caso della regressione semplice $R^2 \equiv \rho_{xy}^2$.

Modello

Si pensa ogni Y_i come una variabile dipendente, funzione lineare affine di X_i (detto predittore) più una perturbazione casuale W_i:

$$Y_i = \beta_0 + \beta_1 X_i + W_i.$$

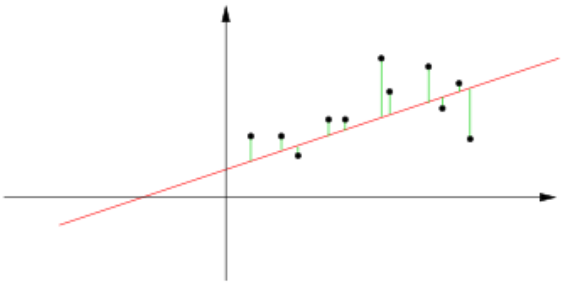
Si suppone che:

- β_0 e β_1 sono numeri reali (incogniti);
- le W_i sono v.a. indipendenti e tutte $\mathcal{N}(0, \sigma^2)$.

Il modello potrebbe comunque non valere perché:

- il legame fra le x e le y potrebbe essere più complesso;
- le W_i potrebbero non essere normali;
- la varianza delle W_i potrebbe dipendere da i (non essere dunque identicamente distribuite).

l'idea è che la retta $y = \beta_0 + \beta_1 x$ sarà quella più vicina ai punti dello *scatterplot*.



La retta dei minimi quadrati ha equazione

$$y = \left(\overline{y}_n - \frac{s_{xy}}{s_x^2}\overline{x}_n\right) + \frac{s_{xy}}{s_x^2}x,$$

e quindi gli stimatori per β_0 e per β_1 sono rispettivamente

$$B_0 = \left(\overline{Y}_n - \frac{S_{xy}}{S_x^2}\overline{X}_n\right) \qquad \text{e} \qquad B_1 = \frac{S_{xy}}{S_x^2}.$$

Stima della varianza delle W_i

Uno stimatore per la varianza σ^2 delle perturbazioni W_i è

$$T = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2.$$

Per stimare la qualità di una regressione possiamo utilizzare i seguenti criteri:

- Il coefficiente di correlazione ρ_{xy} deve essere vicino a 1 o a -1 (ρ^2_{xy} vicino ad 1).
- L'esame visivo dello scatterplot delle due variabili: i dati devono essere vicini alla retta di regressione.
- L'esame del grafico dei residui: in ascissa i valori previsti, in ordinata i valori dei residui. La nuvola dei punti deve avere un aspetto omogeneo, senza segni di curvatura, allargamenti o restringimenti.
  - Un grafico dei residui che presenti curvatura è un indizio che una dipendenza lineare non spiega bene i dati. Si può tentare di correggere questo difetto con trasformazioni di x e/o y, oppure si può provare a passare a una regressione multipla (che definiremo più avanti).
  - Un allargarsi/restringersi della nuvola di punti è un indizio che gli errori non sono tutti dello stesso tipo al variare di i. Si scelga quella combinazione di trasformazioni che danno la nuvola dei residui più omogenea possibile.

Segno della covarianza

Se $s_{xy} > 0$ si hanno variabili positivamente correlate (a valori piccoli di x corrispondono valori piccoli di y e idem per i valori grandi).
Se $s_{xy} < 0$ si hanno variabili negativamente correlate (a valori piccoli di x corrispondono valori grandi di y e viceversa).
Se $s_{xy} = 0$ si hanno variabili scorrelate.

Modello

Si pensa ogni Y_i come una variabile dipendente, funzione lineare affine dei predittori $Z_1, \ldots, Z_k$ più perturbazioni casuali W_i. Si hanno n osservazioni $Y_1, \ldots, Y_n$:

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 Z_{11} + \beta_2 Z_{21} + \cdots + \beta_k Z_{k1} + W_1 \\
Y_2 &= \beta_0 + \beta_1 Z_{12} + \beta_2 Z_{22} + \cdots + \beta_k Z_{k2} + W_2 \\
\cdots &= \cdots \\
Y_n &= \beta_0 + \beta_1 Z_{1n} + \beta_2 Z_{2n} + \cdots + \beta_k Z_{kn} + W_n
\end{aligned}
$$

dove Z_{ij} è la osservazione j del predittore Z_i, $\beta_0, \ldots, \beta_k$ sono k parametri e le $W_1, \ldots, W_n$ sono v.a. i.i.d. con legge $\mathcal{N}(0, \sigma^2)$.

L'equazione $y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)}$ è l'equazione di un iperpiano. Esso rappresenta quell'iperpiano che rende minima la somma dei quadrati delle lunghezze d_i dei segmenti congiungenti i punti osservati all'iperpiano stesso