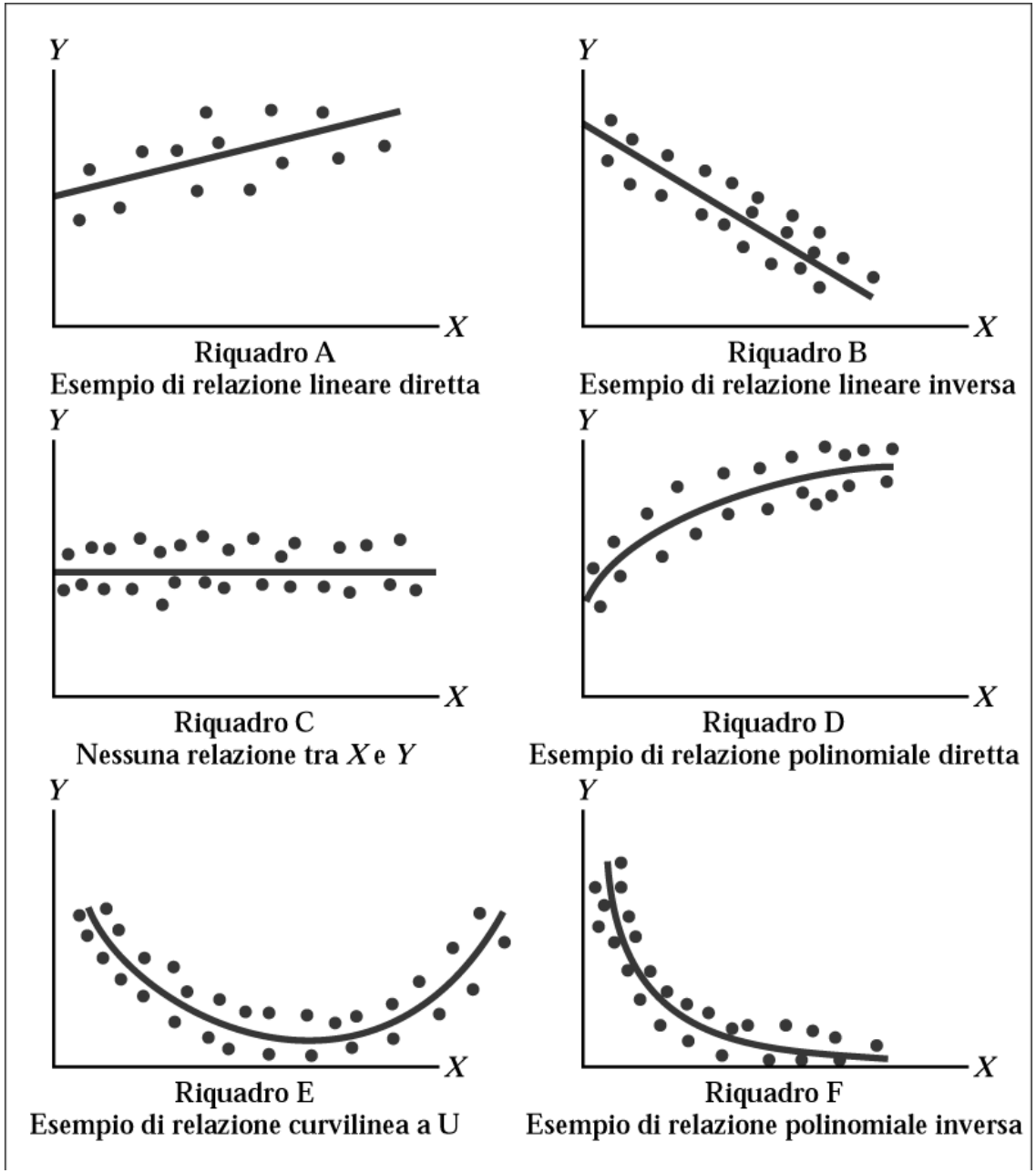
Regression



Esempi di relazioni tra variabili evidenziate dai diagrammi di dispersione

Regression Model with Error Terms

L'analisi di regressione ha per obiettivo quello di sviluppare un modello statistico in grado di prevedere i valori di una variabile dipendente sulla base di una variabile esplicativa.

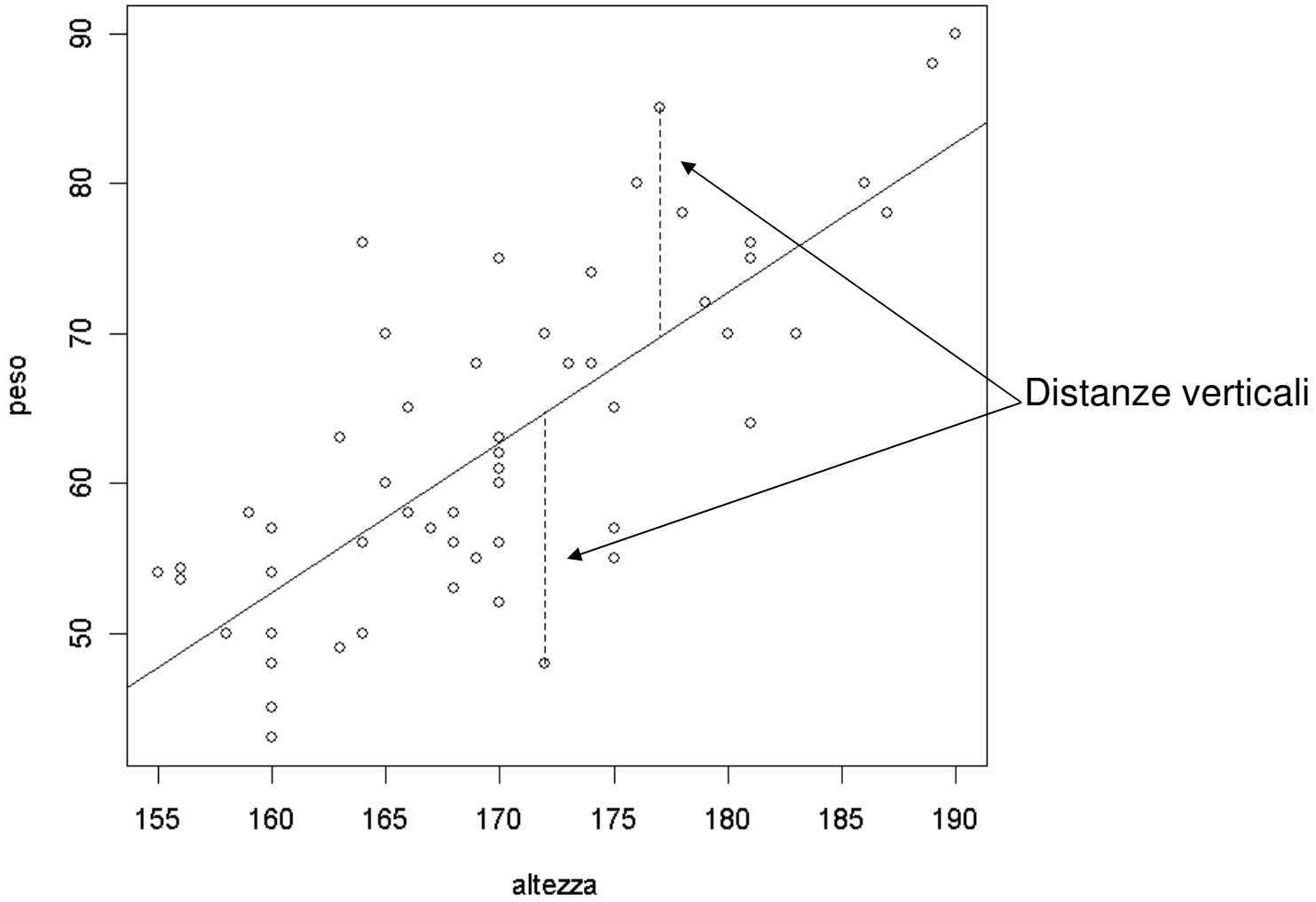
$$y_i = b_0 + b_1 x_i + e_i$$

risposte osservate

intercetta

inclinazione (coefficiente angolare)
coefficiente di regressione lineare

residui

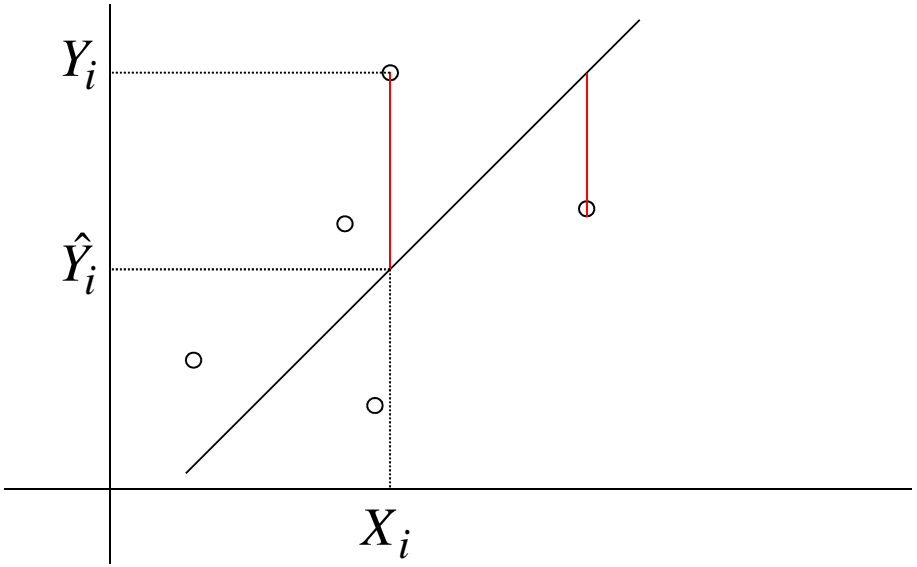


Alcune distanze sono “positive”

Alcune sono “negative”

Come individuare la linea di regressione ottimale (“best fit”)
per un determinato insieme di osservazioni?

In altri termini, come individuare (stimare) i valori di *a* e di *b*
in un’equazione lineare in modo ottimale
(“equazione di regressione”)?

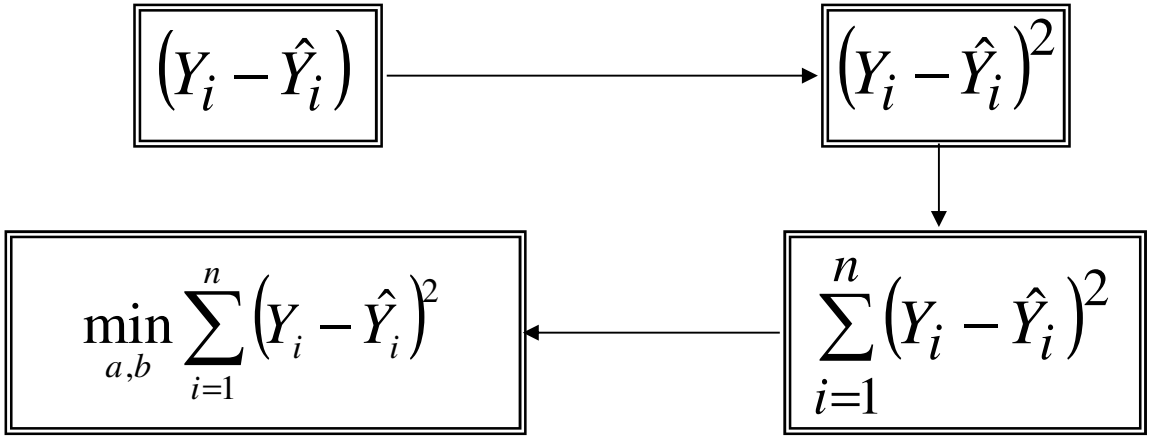


Stimiamo:

$$\hat{Y}_i = b_0 + b_1 X_i$$

$\hat{Y}_i$: valori previsti per Y per l'osservazione i–ma

con il metodo dei minimi quadrati ordinari
(OLS: *ordinary least squares*)



OLS

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\min_{a,b}$$

$$\Updownarrow$$

$$\sum_{i=1}^{n}\left[Y_i-\left(b_0+b_1X_i\right)\right]^2=\min_{a,b}$$

$$\Downarrow$$

$$\hat{Y}_i=b_0+b_1X_i$$

Intercetta: valore di *Y* per *X*=0

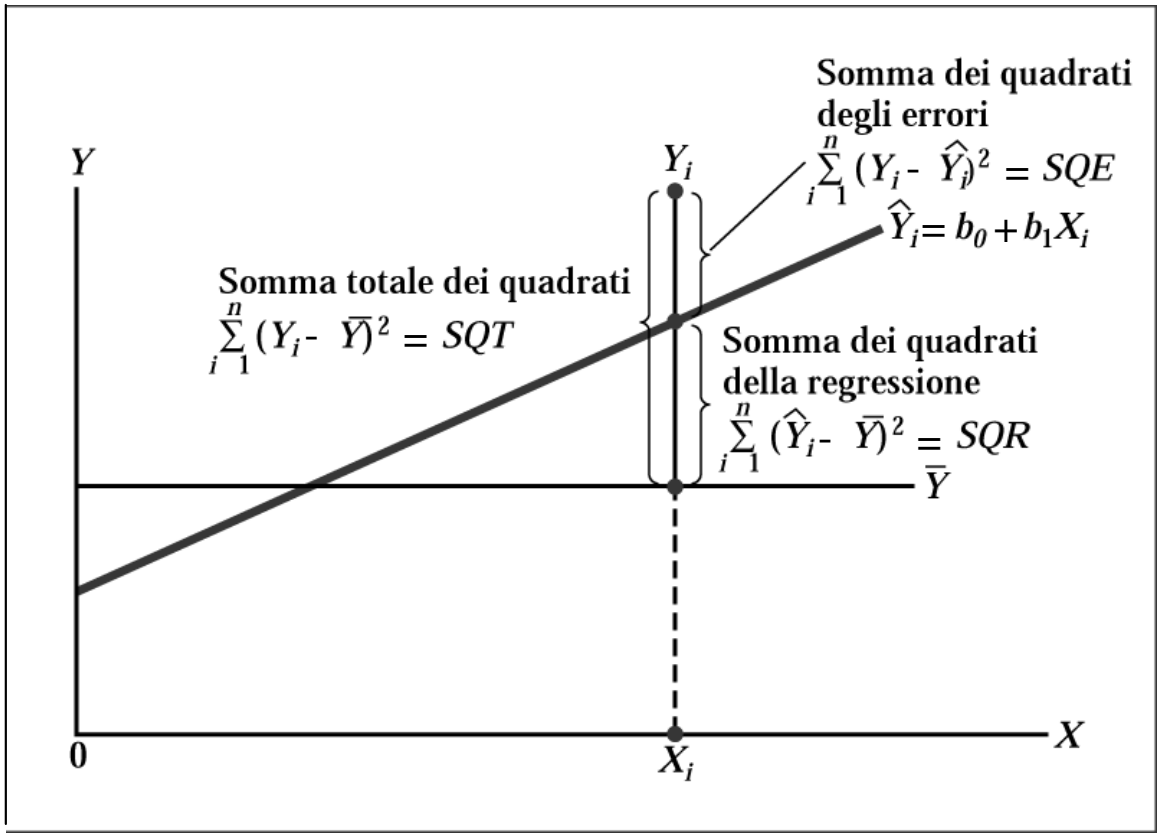
"Inclinazione" della linea di regressione

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \times (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Scomposizione della devianza

È importante per valutare la bontà previsiva (nel *range* osservato) del modello stimato.



Le misure di variabilità nel modello di regressione

Devianza totale

SQT *Somma Totale dei Quadrati* (in inglese *Total sum of squares*): misura di variabilità di Y_i attorno a $\bar{Y}$;

$$SQT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SQT=SQR+SQE

Devianza spiegata

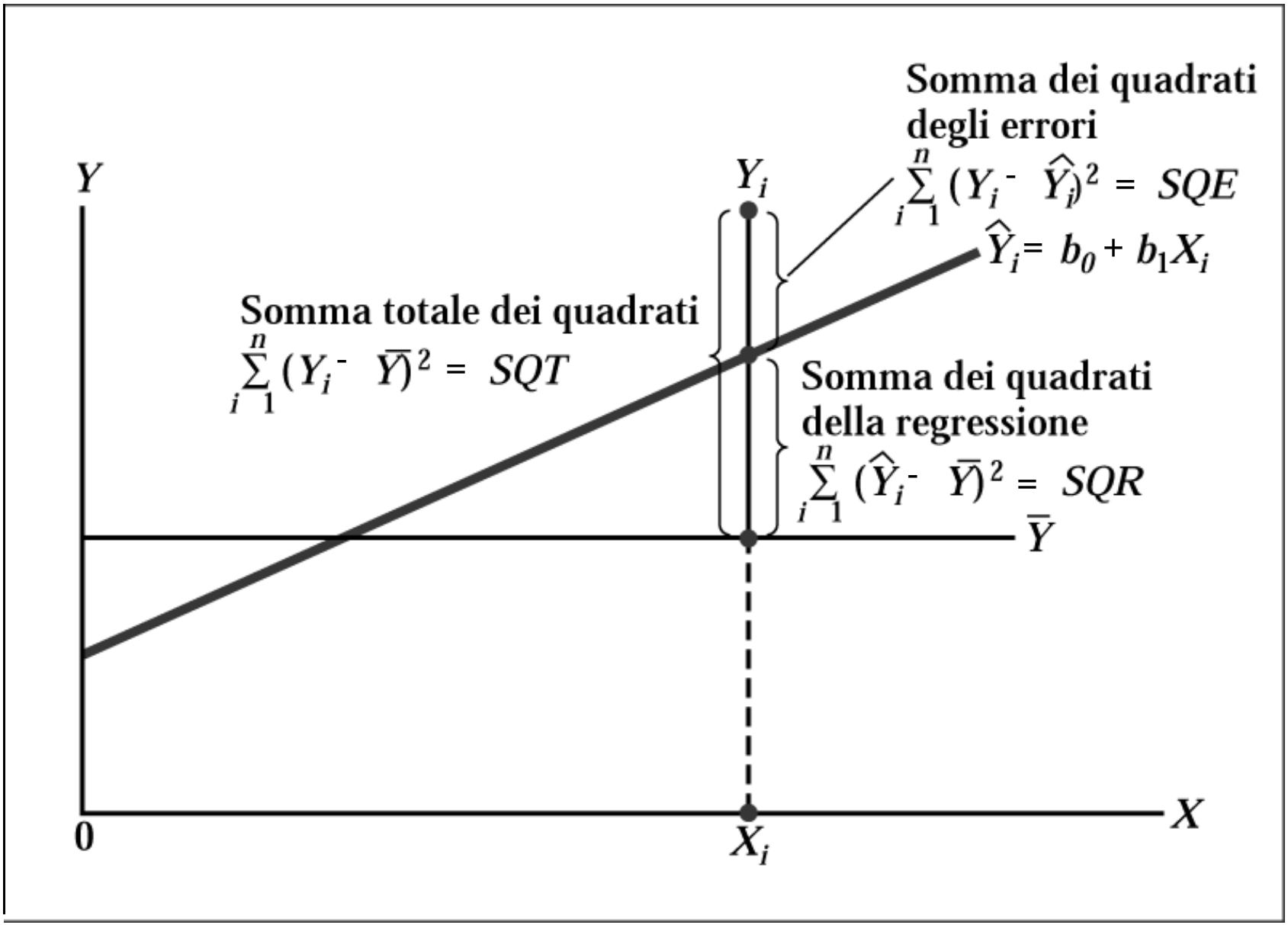
SQR *Somma dei Quadrati della regressione* (in inglese *Regression sum of squares*): attribuita alla relazione lineare tra X e Y;

$$SQR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2$$

Devianza residua

SQE *Somma dei Quadrati degli errori* (in inglese *Error sum of squares*): variabilità non spiegata dalla regressione;

$$SQE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$



Le misure di variabilità nel modello di regressione

$\Downarrow$

Coefficiente di determinazione lineare

$$r^2 = \frac{SQR}{SQT} = \frac{\text{Devianza Spiegata}}{\text{Devianza Totale}}$$

r-quadro: riduzione proporzionale nell'errore di previsione

✓ Una misura di associazione sintetizza quanto bene x può prevedere y.

✓ Se sostituendo i valori di x nell'equazione di previsione $\hat{y} = a + bx$ è possibile prevedere y molto meglio di quanto si può fare senza usare x, allora le variabili sono da considerare come fortemente associate.

✓ Una tale misura di associazione è costituita da quattro aspetti:

- Regola 1 per prevedere y senza usare x.
- Regola 2 per prevedere y utilizzando i valori x.
- Una misura di sintesi dell'errore di previsione per ciascuna regola, E_1 per quelli commessi con la regola 1 e E_2 per quelli commessi con la regola 2.
- La differenza tra l'ammontare di errori con le due regole è $E_1 - E_2$. Convertendo questa riduzione nell'errore in una proporzione si ha:

$$\text{Riduzione proporzionale nell'errore} = \frac{E_1 - E_2}{E_1}$$

✓ Per la regola 1, i valori previsti sono tutti uguali a $\bar{y}$. L'errore di previsione totale è:

$$E_1 = \sum (y - \bar{y})^2$$

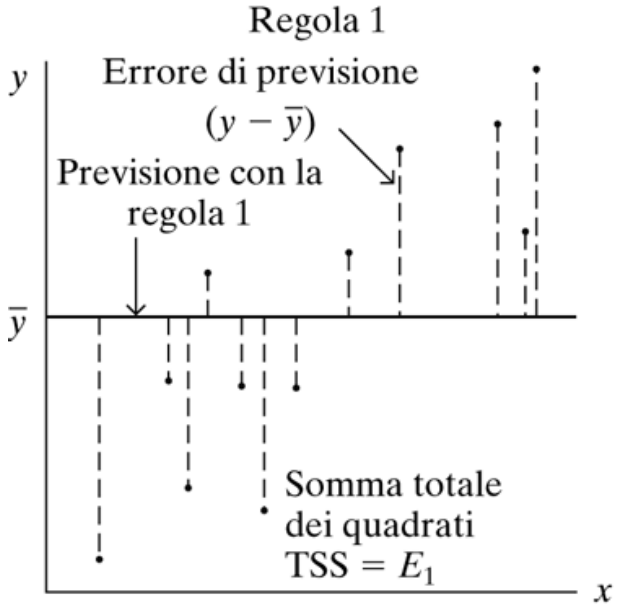
Questa è la *somma totale dei quadrati* dei valori di *y* intorno alla media (TSS).

✓ Per la regola 2, i valori previsti sono i valori $\hat{y}$ derivanti dall'equazione di previsione. L'errore di previsione totale è

$$E_2 = \sum (y - \hat{y})^2$$

Tale quantità è la *somma degli errori al quadrato* (SSE).

✓ Quando *x* e *y* hanno una forte associazione lineare, l'equazione di previsione fornisce previsioni ($\hat{y}$) che sono molto meglio che $\bar{y}$, nel senso che la somma degli errori di previsione al quadrato è notevolmente più piccola.





✓ La riduzione proporzionale nell'errore che si ha impiegando l'equazione di previsione lineare invece che $\bar{y}$ per prevedere y è

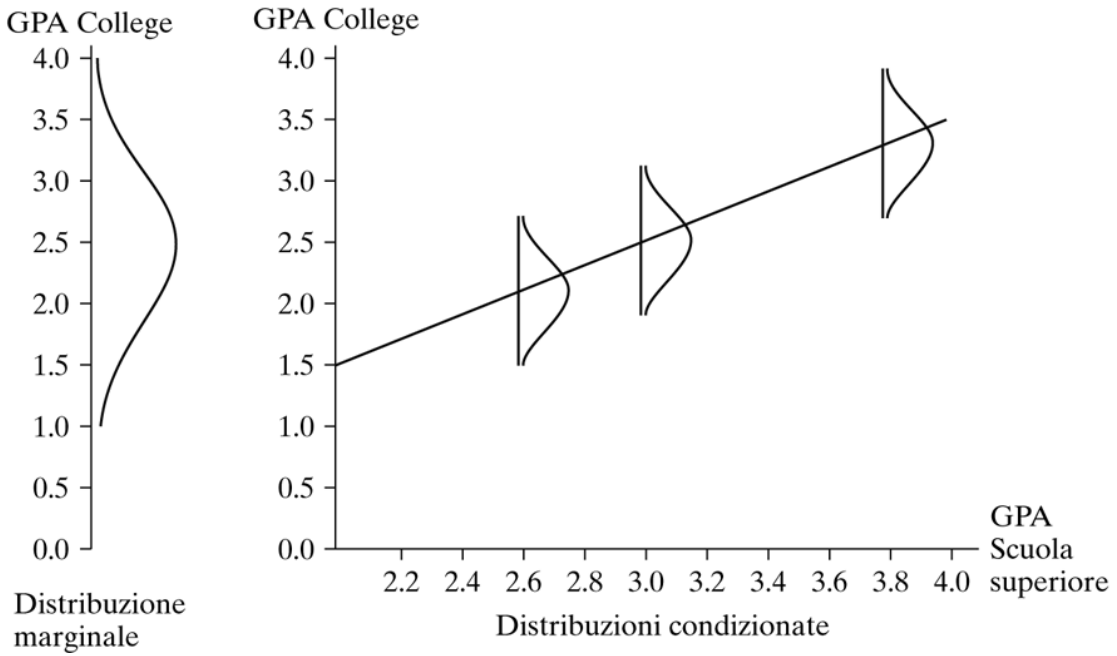
$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{TSS - SSE}{TSS} = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

✓Tale quantità è chiamata ***r*-quadro** o anche **coefficiente di determinazione.**

Inferences

- Assumptions
- Significance test and confidence intervals for the regression coefficient
- Standard error of the regression coefficient

✓ Maggiore è l'associazione, minore è la variabilità condizionata rispetto a quella marginale.



Mean Square Error
(Residual Standard Error)

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

Descrivere le variazioni intorno alla retta di regressione

✓ Il modello di regressione lineare ha un altro parametro σ che descrive la deviazione standard di ciascuna distribuzione condizionata.

✓ σ misura la variabilità dei valori di y per tutti i soggetti che hanno uno stesso valore di x.

✓ Definiamo σ come la **deviazione standard condizionata**.

✓ Per variabili quantitative, l'assunzione di base è che la distribuzione di y sia normale per ogni valore fissato di x.

✓ Il modello ordinario di regressione lineare assume che la deviazione standard σ della distribuzione condizionata di y sia identica per i differenti valori di x.

✓ La stima di σ usa il valore numerico di $SSE = \sum (y - \hat{y})^2$ che misura la variabilità campionaria intorno alla retta dei minimi quadrati:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

La variabilità condizionata tende a essere inferiore alla variabilità marginale

✓ La stima puntuale della deviazione standard della popolazione della variabile *y* è

$$\sqrt{\frac{\sum (y-\bar{y})^2}{n-1}}$$

✓ Questa è la deviazione standard della distribuzione *marginale* di *y* poiché usa soltanto i valori di *y* e ignora i valori di *x*.

✓ Per evidenziare che tale deviazione standard dipende soltanto dai valori di *y* la indicheremo con s_y.

✓ La somma dei quadrati $\sum (y-\bar{y})^2$ nel numeratore di s_y è chiamata la **somma totale dei quadrati**.

✓ Tipicamente, si verifica una minore variabilità nei valori di *y* in corrispondenza di un *x* fissato piuttosto che complessivamente per tutti i valori di *x*.

Inferences

- Assumptions
- Significance test and confidence intervals for the regression coefficient
- Standard error of the regression coefficient

$$\mathrm{SE}(b_1) = \frac{\sqrt{\mathrm{MSE}}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}}$$

Confidence intervals for average values

A confidence interval for $E(y \mid x^\star)$, the average (expected) value of y for a given $x^\star$, is

$$\hat{y} \pm t^\star_{n-2}\, s_y \sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s_y is the standard deviation of the residuals

Prediction intervals for specific predicted values

$$\hat{y} \pm t^{\star}_{n-2}\, s_y \sqrt{1 + \frac{1}{n} + \frac{(x^{\star} - \bar{x})^2}{(n-1)s_x^2}}$$

ASPETTI DELLA REGRESSIONE

- La distinzione fra variabili esplicative e di risposta è fondamentale nella regressione
- Esiste una stretta connessione fra la correlazione e il coefficiente di regressione
- La retta di regressione (a m.q.) passa sempre per i punti che hanno come coordinate la media della variabile di risposta e la media della variabile esplicativa
- Il quadrato della correlazione è la frazione della variabilità nei valori della variabile di risposta (*y*) spiegata dalla retta di regressione di *y* su *x*.

La regressione verso la media

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

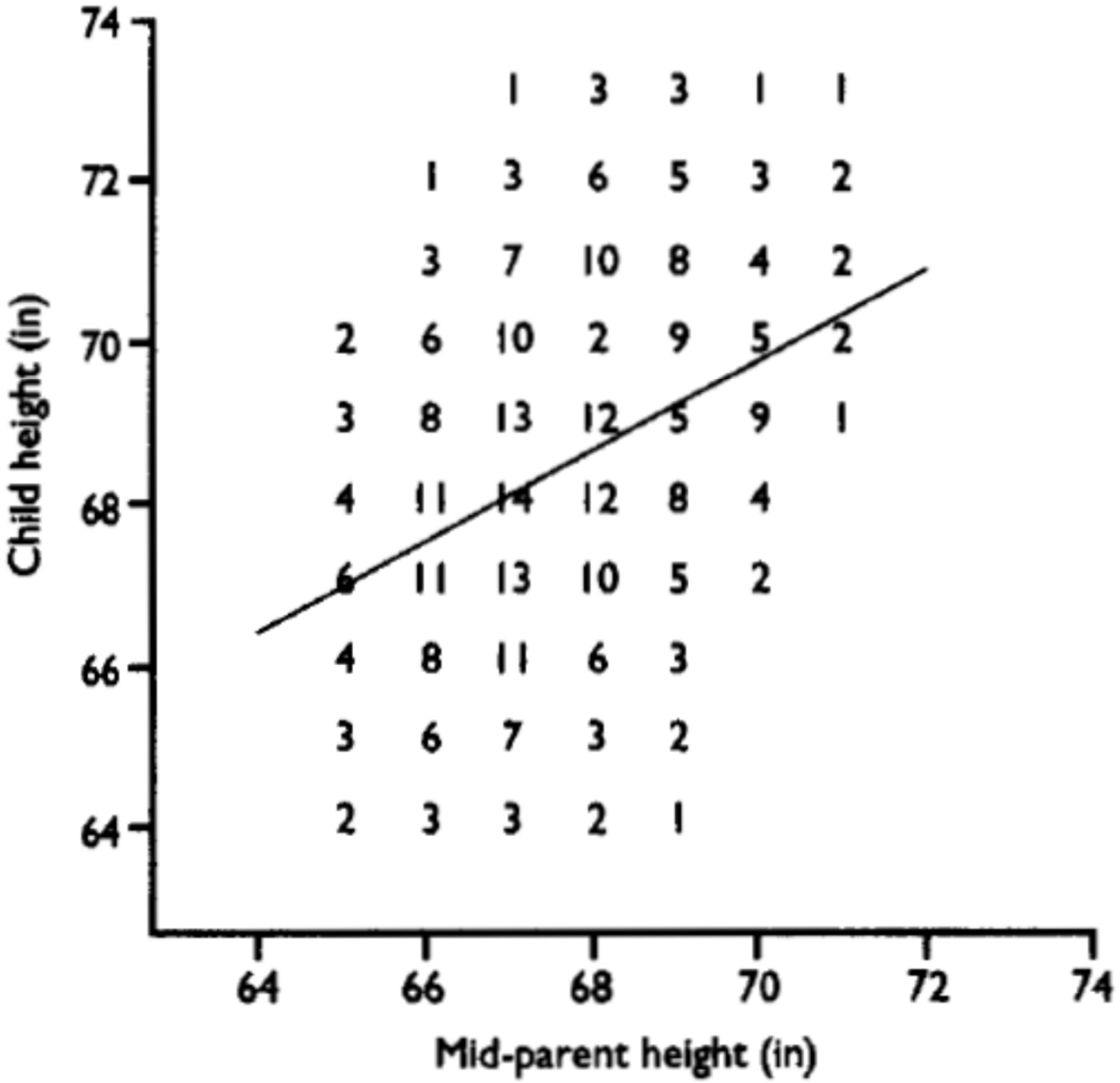
1 Galton F. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* 1886;15:246-63.

Statistics Notes

Regression towards the mean

J Martin Bland, Douglas G Altman

BMJ VOLUME 308 4 JUNE 1994 1499



Galton's original data showing the relation between the heights of children and their parents, with regression line[1]

$$b_1 = r\frac{\sigma_y}{\sigma_x}$$

$$b_1 = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Inoltre, dalla relazione precedente che lega b_1 ad r, segue che, qualora le deviazioni standard σ_X e σ_Y delle due variabili siano uguali, il coefficiente di regressione b_1 sarà esattamente uguale a r e, quindi, minore di 1 in valore assoluto (abbiamo già detto che il caso $|r| = 1$ non è quasi mai un caso reale). A maggior ragione, nel caso in cui $\sigma_Y < \sigma_X$, si avrà $|b_1| < |r| < 1$. Soltanto se $\sigma_Y > \sigma_X$ sarà possibile che $|b_1| > 1$, ma questo dipenderà sia dall'entità di $|r|$ sia dal valore del rapporto σ_Y / σ_X.

Nel caso descritto da Galton le varianze (e le deviazioni standard) delle altezze di genitori e figli sono alquanto diverse fra loro.

```
> apply(galton,2,sd)
     mph      cah
1.666310 2.371068
> apply(galton,2,var)
     mph      cah
2.776589 5.621963
```

```
> apply(galton,2,var)
     mph      cah
2.776589 5.621963
```

In particolare la varianza delle altezze dei genitori (2.78) è circa la metà della varianza delle altezze dei figli (5.62). Dobbiamo però ricordare che le altezze dei genitori sono medie di due osservazioni e che le medie hanno una variabilità più bassa di quella dei valori che la costituiscono; in particolare la varianza della media di due osservazioni è la metà della varianza dei dati originali. Questo spiega il risultato osservato.

Grazie a questo il rapporto σ_Y/σ_X fra le deviazioni standard delle due variabili risulta ben maggiore di 1:

```
> sd(galton$cah)/sd(galton$mph)
[1] 1.422945
```

(è quasi uguale a $\sqrt{2}$) ma, data la bassa correlazione fra le due variabili (circa 0.4), il loro prodotto non riesce a superare il valore 1. Per ottenere questo risultato sarebbe stata necessaria una correlazione almeno pari a 0.7.

Come fanno notare Bland e Altman, il fenomeno della regressione verso la media è presente in numerose circostanze. Fra i vari esempi proposti, i due Autori citano il caso di uno studio in cui era stato chiesto ad un campione di persone di riferire (senza pesarsi) il proprio peso. I soggetti erano stati successivamente pesati ed il peso misurato era stato confrontato con quello riferito impiegando un modello di regressione lineare in cui la variabile dipendente era il peso riferito e la variabile indipendente era il peso effettivamente misurato. La pendenza della retta di regressione era risultata minore di uno. Quindi, il peso riferito da un soggetto "pesante" era (mediamente) inferiore al peso reale, mentre il peso riferito da un soggetto "leggero" era (mediamente) superiore al peso reale. È immediata (ma superficiale) l'interpretazione "psicologica": le persone sovrappeso tendono a sottostimare il loro peso, mentre quelle sottopeso tendono a sovrastimarlo. Infatti, se scambiando la variabile dipendente e quella indipendente, trovassimo ancora una volta una pendenza minore di uno, l'interpretazione sarebbe opposta: le persone sovrappeso sovrastimano il loro peso, mentre quelle sottopeso lo sottostimano. Il risultato osservato può essere invece attribuito (almeno in parte) alla regressione verso la media.

Lo psicologo Daniel Kahneman, premio Nobel per l'Economia, afferma che la regressione verso la media potrebbe spiegare la teoria secondo cui i rimproveri migliorerebbero la *performance*, mentre le lodi la peggiorerebbero.

Anche l'interpretazione dei risultati di test che valutano il livello di conoscenza deve tener conto della regressione verso la media. Uno studente che supera un test conseguendo un punteggio molto elevato tenderà (in media) a conseguire un voto più basso quando il test viene ripetuto e viceversa. Per valutare se un intervento produce un reale miglioramento è fondamentale la presenza di un gruppo di controllo.

Anche nello sport è presente la regressione verso la media. Un atleta che in un certo anno ottiene delle ottime prestazioni, tenderà (in media) a non ripeterle l'anno successivo (e viceversa un giocatore con basse prestazioni tenderà, in media, a migliorarle). Basta guardare le statistiche sui giocatori americani di *baseball*.

- È possibile verificare quanto abbiamo appena detto mediante una simulazione. Supponiamo che un test sia costituito da 100 domande di tipo Vero/Falso e che ogni studente risponda a caso a ciascuna domanda. Quello che ci aspettiamo è che, *in media*, uno studente risponda correttamente a (circa) 50 domande.

- Prendiamo ora il 10% degli studenti che hanno riportato il punteggio peggiore e osserviamo che la loro media è, del tutto comprensibilmente, alquanto inferiore a 50.

- Supponiamo ora di programmare un intervento educativo su questi 20 studenti per cercare di migliorare il loro risultato, ma che l'intervento non abbia alcun esito; vale a dire che gli studenti sceglieranno ancora a caso la risposta.

- Ripetiamo a questo punto il test su questi studenti alla fine dell'intervento educativo e osserviamo che il loro punteggio medio è migliorato, anche se gli studenti hanno continuato a rispondere a caso.

- La spiegazione è data dalla regressione verso la media.

- Per verificare concretamente che l'intervento educativo non ha avuto alcun esito sarebbe stato necessario dividere a caso i 20 studenti in due gruppi, sottoponendo all'intervento solo uno dei due gruppi, lasciando l'altro come controllo. Si sarebbe allora visto che il punteggio medio dei due gruppi sarebbe stato simile (con un valore intorno a 50) e migliore del punteggio iniziale.

- Come illustra questo esempio, è molto importante tener conto della regressione verso la media quando si programma un esperimento. In particolare è fondamentale avere un gruppo di controllo.