Metodi Quantitativi per Economia, Finanza e Management

Lezione n°8

Regressione lineare multipla: le ipotesi del modello, la stima del modello

Il modello di regressione lineare

1. Introduzione ai modelli di regressione
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
6. Commenti

L’obiettivo dell’analisi

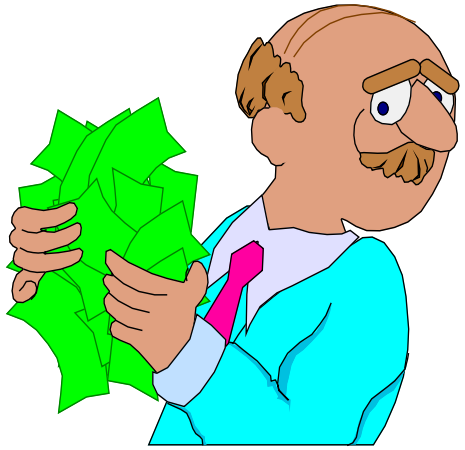
Prevedere la redditivita’
del cliente al tempo t+1

L’impostazione del problema

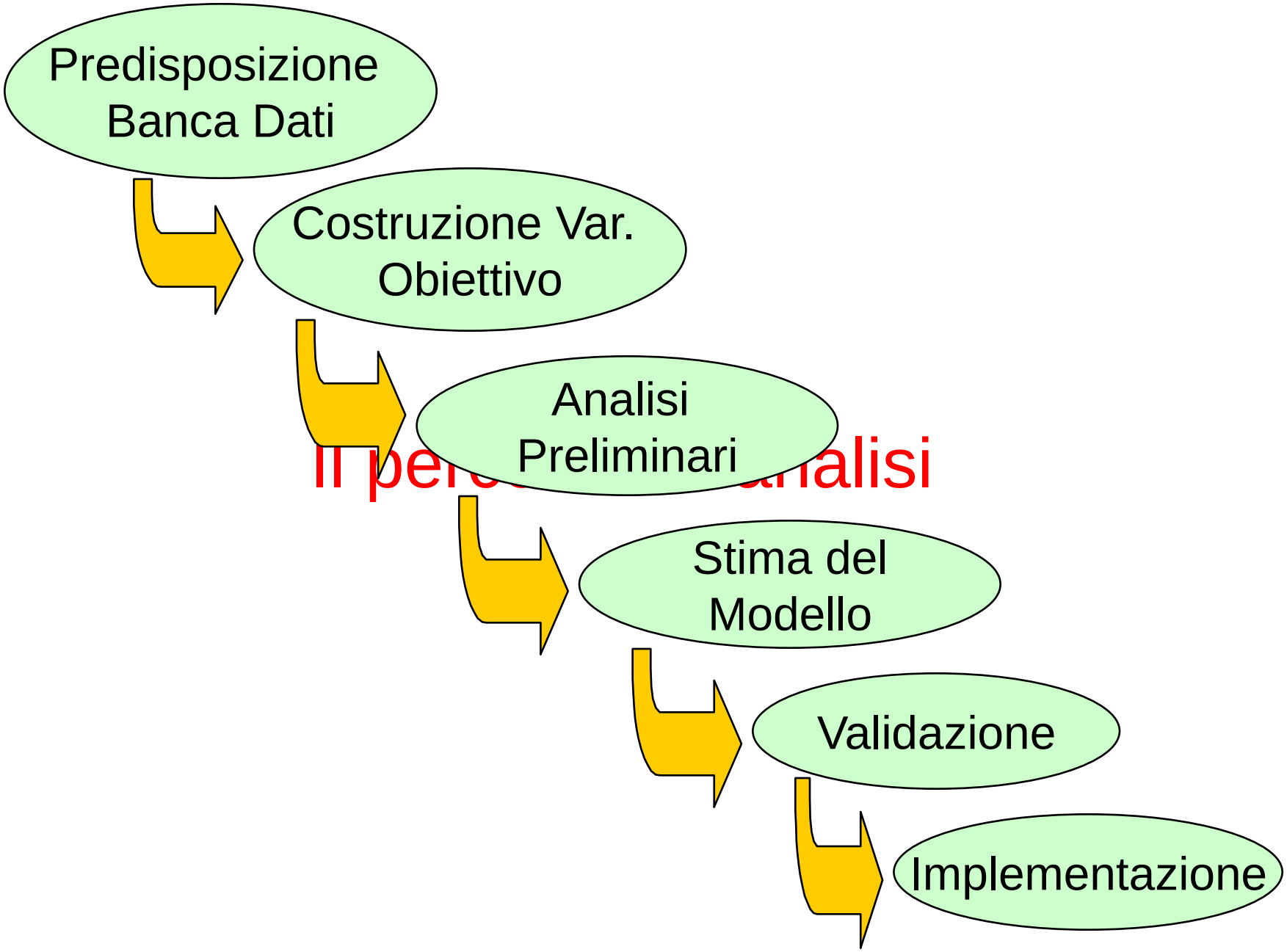
Redditività = ricavi - costi

- redditività var. continua
- classi di redditività (< 0 ; >= 0)

I dati di input



- Y : Redditività consolidata al tempo ‘t’
- X : # ordini al tempo ‘t-1,-2,..’
  pagato ordini al tempo ‘t-1,-2,..’
  pagato rateale mensile …
  sesso (dicotomica)
  area (dicotomiche)
  …..

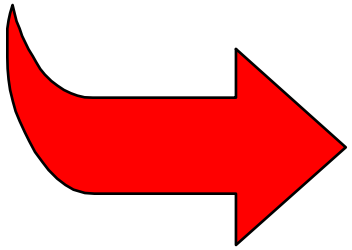

Predisposizione Banca Dati
Costruzione Var. Obiettivo
Analisi Preliminari
Stima del Modello
Validazione
Implementazione

Analisi preliminari

- lo studio della distribuzione di Y
- la struttura di correlazione tra Y e $\underline{X}$

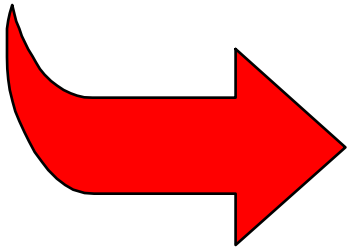
L’impostazione del problema

- Redditività var. continua



Regressione Lineare

- Redditività var. dicotomica



Regressione Logistica

Il modello di regressione lineare

1. Introduzione ai modelli di regressione
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
6. Commenti

I modelli di regressione

Modelli di dipendenza per la rappresentazione di relazioni non simmetriche tra le variabili

- Y “variabile dipendente” (variabile target da spiegare)
- $X_1,\ldots,X_p$ “variabili indipendenti” (variabili esplicative o regressori)

Il modello di regressione lineare

Si vuole descrivere la relazione tra Y e $X_1,\ldots,X_p$ con una funzione lineare

- se p=1 → osservazioni in uno spazio a due dimensioni (i=1,…,n)

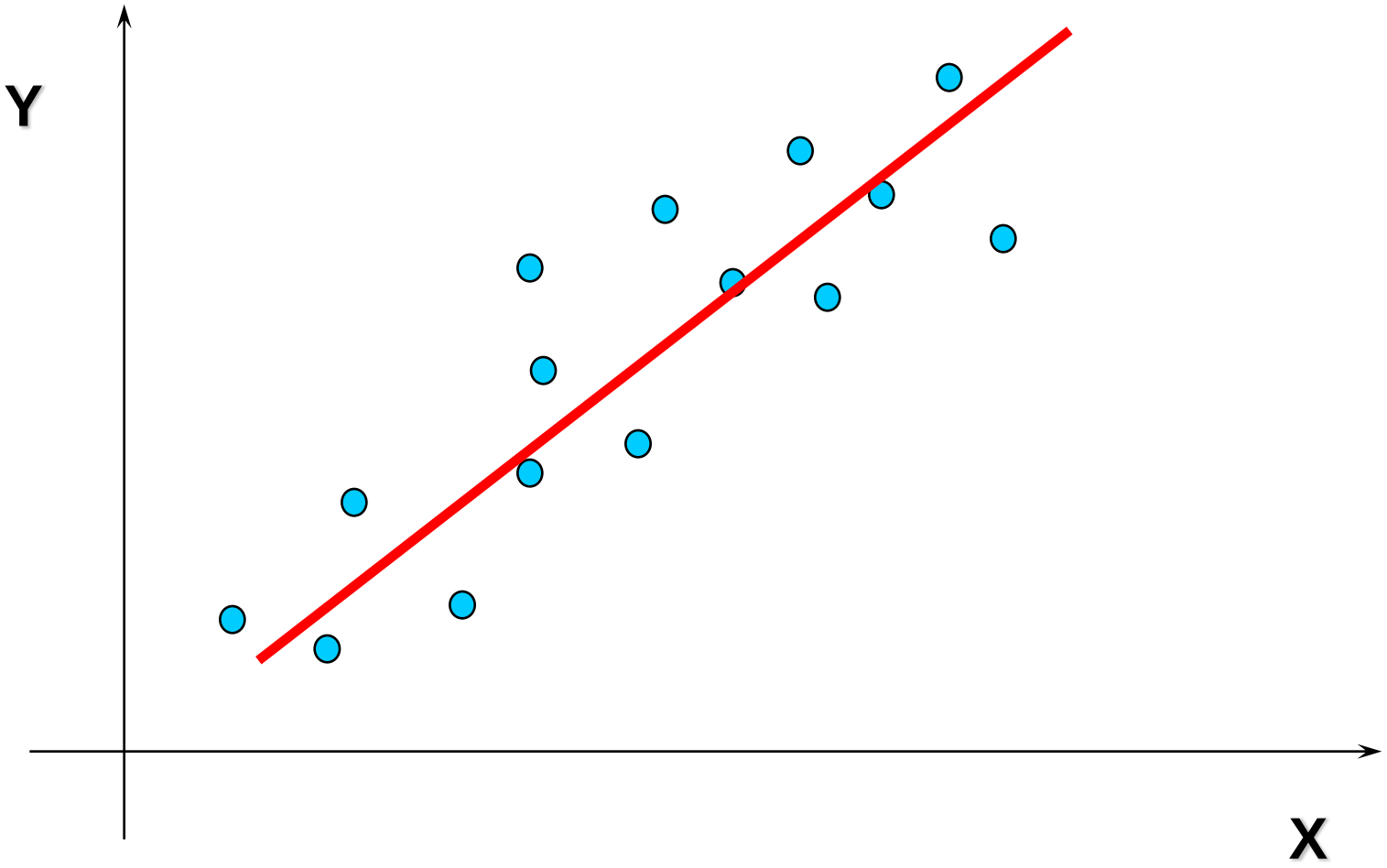
$$Y_i = f(X_{i1})$$

- se p>1 → osservazioni in uno spazio a p+1 dimensioni (i=1,…,n)

$$Y_i = g(X_{i1},\ldots, X_{ip})$$

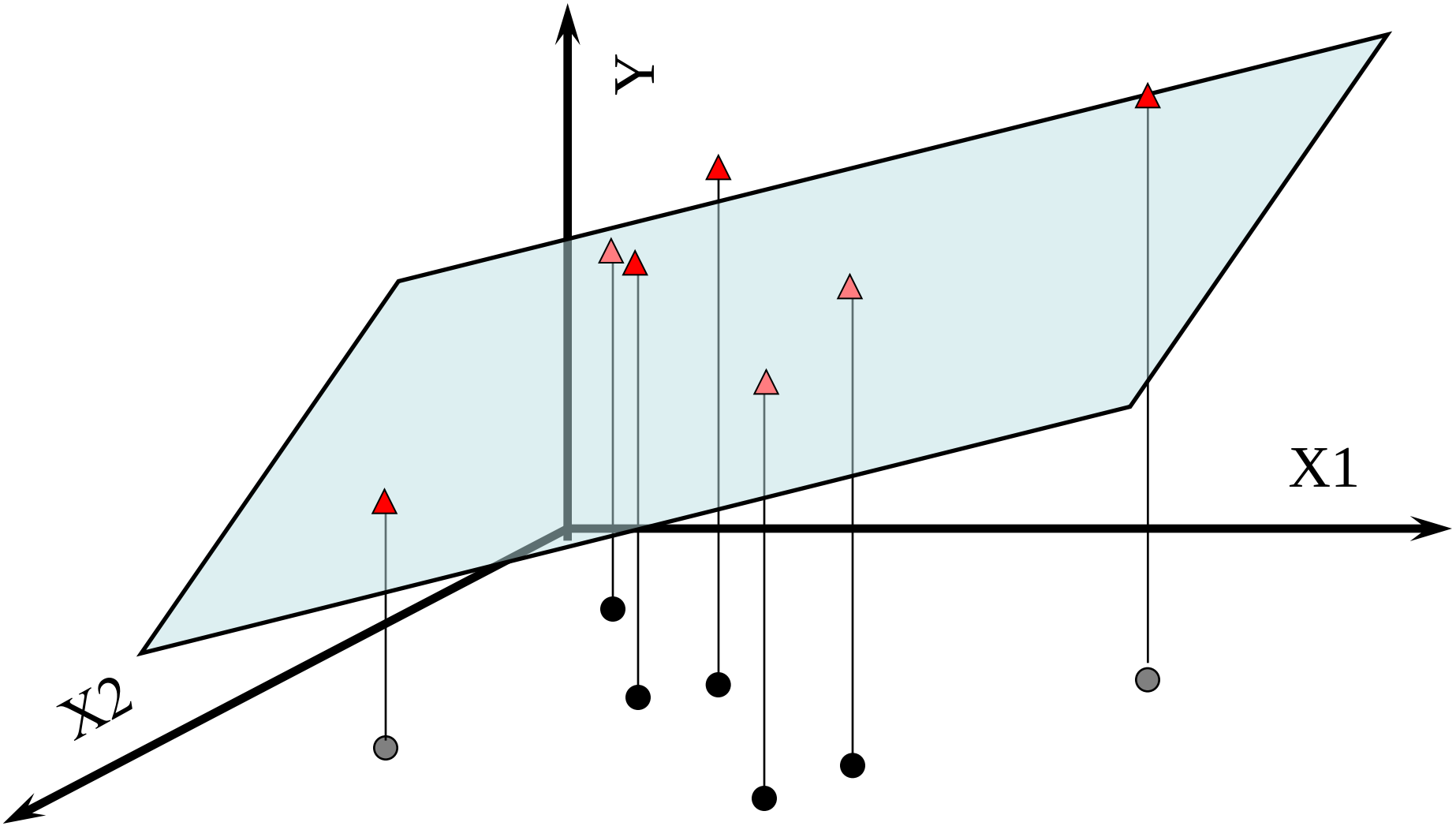
Il modello di regressione lineare

- se p=1 → spazio a due dimensioni → retta di regressione lineare semplice



Il modello di regressione lineare

- se p>1 → spazio a p+1 dimensioni → “retta” di regressione lineare multipla



Il modello di regressione lineare

Obiettivi

- Esplicativo - Stimare l'influenza dei regressori sulla variabile target.

- Predittivo - Stimare il valore non osservato della variabile target in corrispondenza di valori osservati dei regressori.

- Comparativo - Confrontare la capacità di più regressori, o di più set di regressori, di influenzare il target (= confronto tra modelli di regressione lineare diversi).

Il modello di regressione lineare

Le ipotesi del modello

Y	X_1	X_2	X_3	…	…	…	X_p
y_1	x_{11}	x_{12}	x_{13}	…	…	…	x_{1p}
y_2	x_{21}	x_{22}	x_{23}	…	…	…	x_{2p}
y_3	x_{31}	x_{32}	x_{33}	…	…	…	x_{3p}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
y_n	x_{n1}	x_{n2}	x_{n3}	…	…	…	x_{np}
(nx1)				(nxp)			

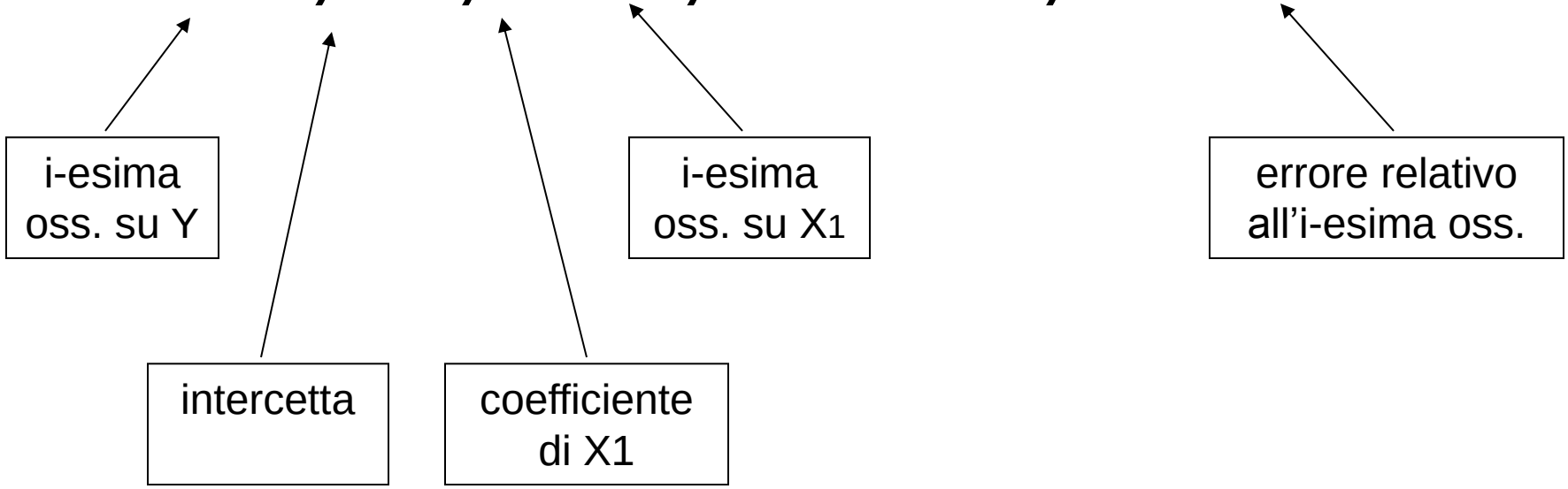
- n unità statistiche
- vettore colonna (nx1) di n misurazioni su una variabile continua (Y)
- matrice (nxp) di n misurazioni su p variabili quantitative ($X_1,\ldots,X_p$)
- la singola osservazione è il vettore riga $(y_i,x_{i1},x_{i2},x_{i3},\ldots,x_{ip})$ $i=1,\ldots,n$

Il modello di regressione lineare

Le ipotesi del modello

Equazione di regressione lineare multipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$



La matrice $X=[1,X_1,\ldots,X_p]$ è detta matrice del disegno.

Il modello di regressione lineare

Le ipotesi del modello

L'errore presente nel modello si ipotizza essere di natura casuale. Può essere determinato da:

- variabili non considerate
- problemi di misurazione
- modello inadeguato
- effetti puramente casuali

Il modello di regressione lineare

Le ipotesi del modello

1. Errori a media nulla $E(\varepsilon) = 0$

2. Errori con varianza costante (omoschedasticità) $Cov(\varepsilon) = \sigma^2 I_n$

3. Errori non correlati (per ogni i≠j) $Cov(\varepsilon_i, \varepsilon_j) = 0$

4. Errori con distribuzione Normale $\varepsilon \sim N(0, \sigma^2 I_n)$

* 1 – 3 → hp deboli
  1 – 4 → hp forti

Il modello di regressione lineare

Le ipotesi del modello

Da un punto di vista statistico

- Y è un vettore aleatorio di cui si osserva una specifica realizzazione campionaria → hp sulla distribuzione
- X è una matrice costante con valore noto → no hp sulla distribuzione
- beta è un vettore costante non noto
- l'errore è un vettore aleatorio di cui si osserva una specifica realizzazione campionaria → hp sulla distribuzione

Il modello di regressione lineare

Le ipotesi del modello

- in media Y può essere rappresentata come funzione lineare delle sole ($X_1,\ldots,X_p$)

$$\mu = E(Y) = X\beta$$

- ogni osservazione di Y è uguale ad una combinazione lineare dei regressori con pesi=coefficienti beta + un termine di errore

$$Y = X\beta + \varepsilon$$

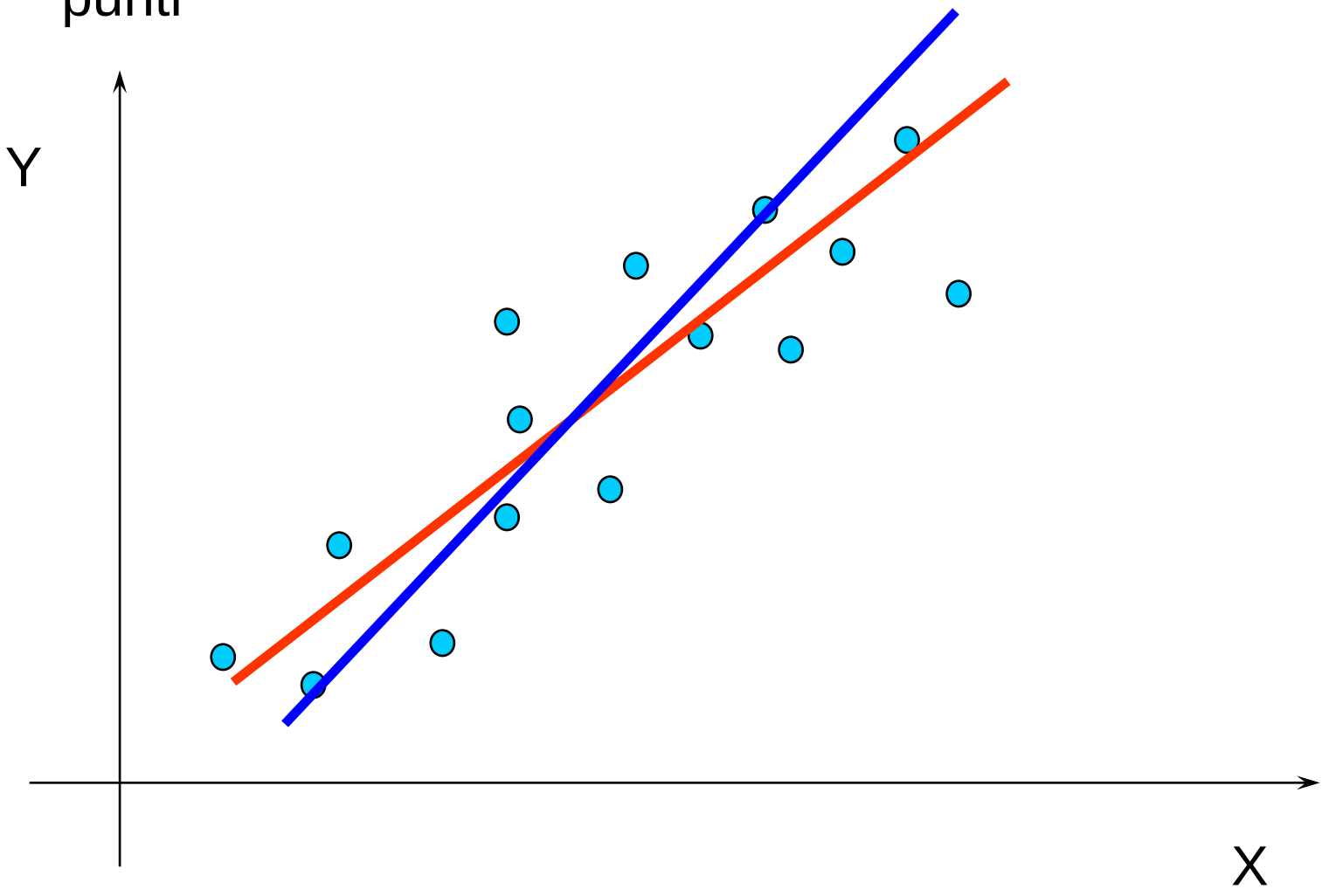
Il modello di regressione lineare

1. Introduzione ai modelli di regressione
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
6. Commenti

Il modello di regressione lineare

La stima del modello

Si vuole trovare la retta lineare migliore data la nuvola di punti



Il modello di regressione lineare

La stima del modello

Equazione teorica → coefficienti non noti

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Equazione stimata → coefficienti stimati (una delle infinite rette possibili)

$$Y = b_o + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + \hat{\varepsilon}$$

stime dei coefficienti

$$Y = \hat{Y} + \hat{\varepsilon}$$

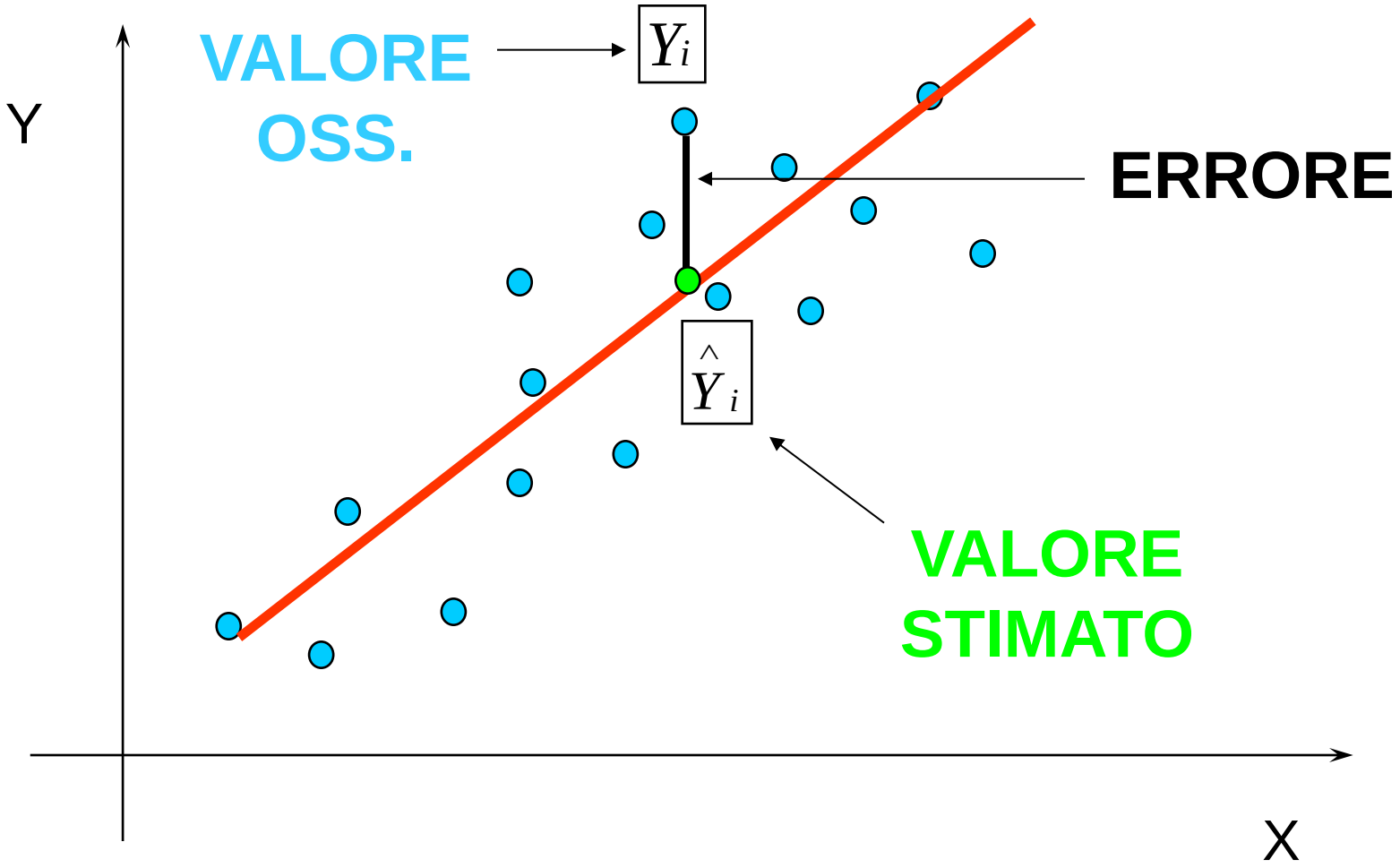
previsione

errore di previsione

Il modello di regressione lineare

La stima del modello

Stimando la retta di regressione si commette un errore di previsione: Metodo dei Minimi Quadrati



Il modello di regressione lineare

La stima del modello

Obiettivo → trovare la miglior approssimazione lineare della relazione tra Y e $X_1,\ldots,X_p$ (trovare le stime dei parametri beta che identificano la "migliore" retta di regressione)

Metodo dei minimi quadrati → lo stimatore LS è la soluzione al problema

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - X_i\beta\right)^2 \Leftrightarrow \min_{\beta} \varepsilon'\varepsilon$$

Il modello di regressione lineare

La stima del modello

Lo stimatore dei Minimi Quadrati: LS

- è funzione di Y e X $\hat{\beta}_{LS} = (X'X)^{-1} X'Y$

- ha media $E(\hat{\beta}_{LS}) = \beta$

- ha varianza $Var(\hat{\beta}_{LS}) = (X'X)^{-1}\sigma^2$

Il modello di regressione lineare

La stima del modello

Proprietà dello stimatore LS

- non distorto
- consistente (se valgono certe hp su X'X)
- coincide con lo stimatore di max verosimiglianza sotto hp forti

→ BLUE (Best Linear Unbiased Estimator)

Il modello di regressione lineare

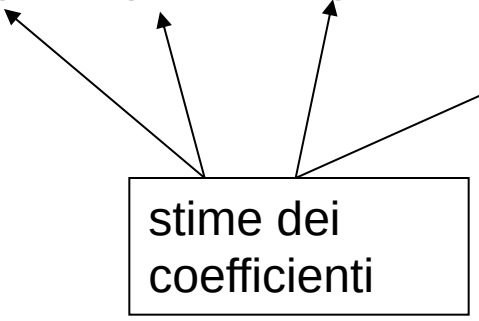
La stima del modello

Equazione teorica → coefficienti non noti

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Equazione stimata → coefficienti stimati (una delle infinite rette possibili)

$$\hat{Y} = b_o + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$



Il modello di regressione lineare

La stima del modello

Interpretazione dei coefficienti

- impatto di X_j su Y posto che nel modello sono presenti altre variabili

- tasso di variazione di Y al variare di X_j

- come varia Y al variare di una unità di X_j se gli altri regressori non variano

Il modello di regressione lineare

La stima del modello

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	-15016	2324.86370	-6.46	<.0001
PAG_ORD	Pagato in contrassegno	**1**	1.19433	0.05485	21.78	<.0001
PAG_MES	Pagato con rate mensili	**1**	2.52341	0.10102	24.98	<.0001
TOT_ORD	Totale ordini	**1**	14881	683.88703	21.76	<.0001
LISTA	Numero di liste di appartenenza	**1**	603.36550	1110.84778	0.54	0.5871
SESSO	Sesso	**1**	3453.14705	1994.83468	1.73	0.0835
CEN	Residenza Centro	**1**	-6431.88493	2597.25872	-2.48	0.0133
SUD	Residenza Sud	**1**	-18390	2077.96317	-8.85	<.0001

Il modello di regressione lineare

La stima del modello

Segno del coefficiente

• indica la direzione dell'impatto del regressore a cui è associato

Valore del coefficiente

• indica l'incremento marginale di Y
• dipende dall'unità di misura di X_j

Il modello di regressione lineare

La stima del modello

Valore del coefficiente

• per valutare l'impatto relativo dei singoli regressori è necessario considerare i coefficienti standardizzati

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	30935	869.69238	35.57	<.0001	0	0
Factor1		**1**	61162	869.81092	70.32	<.0001	0.71583	1.00000
Factor3		**1**	24154	869.81092	27.77	<.0001	0.28269	1.00000
Factor4		**1**	3446.48124	869.81092	3.96	<.0001	0.04034	1.00000
Factor6		**1**	-13861	869.81092	-15.94	<.0001	-0.16223	1.00000