Variabile X su scala qualitativa (due categorie)

modello di regressione: variabili quantitative misurate almeno su scala intervallo (meglio se Y è di questo tipo e preferibilmente anche le X_i)

variabili esplicative X_i su scala qualitativa $\Longrightarrow$ variabili *dummy*

1. variabili dicotomiche: 2 sole categorie A e B $\begin{cases} X = 0 & \text{se } A \\ X = 1 & \text{se } B \end{cases}$

es sesso: M $\Rightarrow$ $X = 1$
F $\Rightarrow$ $X = 0$

Y = reddito

interpretazione coefficienti:

$\hat{y} = a + bx$

$\hat{y} = a \Longrightarrow$ reddito medio per Femmine ($X = 0$)

$\hat{y} = a + b \Rightarrow$ reddito medio per Maschi ($X = 1$)

differenza tra il reddito medio delle Femmine e dei Maschi

Variabile X su scala qualitativa (più categorie)

2. variabili politomiche $<$ ordinali / nominali

uso variabili dummy: una variabile espressa in C categorie può essere rappresentata in C -1 variabili dummy

Y = contributo in dollari ad una campagna elettorale
X = interesse politico del rispondente
1 = nessun interesse
2 = poco interesse
3 = molto interesse

modello di regressione: $Y = a + b_1 X_1 + b_2 X_2 + u$

$\begin{cases} X_1 = 1 & \text{se } X = 2 \text{ (poco interesse)} \\ X_1 = 0 & \text{altrimenti} \end{cases}$ $\begin{cases} X_2 = 1 & \text{se } X = 3 \text{ (molto interesse)} \\ X_2 = 0 & \text{altrimenti} \end{cases}$

⇨ non serve una terza variabile per X = 1 (nessun interesse)

⇩

definite X_1 e X_2, X_3 è una perfetta combinazione lineare ⇨ **multicollinearità**

3ª modalità (nessun interesse, X_3 = 1) definita da X_1 = 0 e X_2 = 0

Variabile X su scala qualitativa (più categorie)

a ⇨ stima del contributo medio alla campagna elettorale quando X = 1 (nessun interesse politico)

base per confrontare gli effetti della altre 2 categorie su Y

es X = 2 ⇨ poco interesse politico ⇨ X_1 = 1 e X_2= 0

$$Y = a + b_1 X_1 + b_2 X_2 + u$$

$$\hat{Y} = a + b_1(1) + b_2(0)$$

$$\hat{Y} = a + b_1$$

b_1 → differenza nel contributo medio tra la categoria di rispondenti con "poco interesse politico" e quelli con "nessun interesse"

$(a + b_1) - a = b_1$

X = 3 ⇨ molto interesse politico ⇨ X_1 = 0 e X_2= 1

$$\hat{Y} = a + b_1(0) + b_2(1)$$

$$\hat{Y} = a + b_2$$

Inferenza sul coefficiente di regressione

Un coefficiente angolare diverso da 0 indica che c'è dipendenza della *Y* dalla *x.*

Tuttavia non basta il risultato della stima ottenuto da un campione. Occorre condurre il seguente test per inferire il valore del coefficiente angolare della popolazione:

H_0: $\beta_1=0$ → assenza di relazione
H_1: $\beta_1 \neq 0$

Ipotesi per inferenza

$$Y_i = \alpha + \beta X_i + u_i$$

a) *n* osservazioni indipendenti

b) x_i valori prefissati (non variabili casuali)

c) $E(u_i) = 0$ (media nulla)

d) $E(u_i^2) = \sigma^2$ (varianza costante)

e) u_i segue la distribuzione normale $N(0, \sigma^2)$

$$f\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)$$

Distribuzione dei parametri *a* e *b*

Gli stimatori dei minimi quadrati *a* e *b* - entrambi funzioni lineari del termine di errore *u* - hanno anch'essi distribuzione di probabilità normale:

$$a \sim N\left(\alpha;\sigma_u^2 \cdot \left[1/n + \bar{X}^2 \Big/ \sum_{i=1}^{n} x_i^2\right]\right)$$

$$b \sim N\left(\beta;\sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2\right)$$

Per effettuare delle operazioni di inferenza σ^2_u deve essere opportunamente sostituita con un suo stimatore corretto s^2 passando da una distribuzione normale a una distribuzione *t*

$$\frac{a-\alpha}{s \cdot \sqrt{1/n + \bar{X}^2 \Big/ \sum_{i=1}^{n} x_i^2}} \sim t_{(n-2)}$$

$$\frac{b-\beta}{s \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} \sim t_{(n-2)}$$

Stimatore s_{yx}

Una stima corretta di σ^2 è data da:

$$\frac{DEV(E)}{n-2}$$

n. gdl: *n –nr. parametri*

$$S_{YX} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{DEV(E)}{n-2}}$$

Inferenza sul coefficiente di regressione β

In corrispondenza del seguente sistema di ipotesi:

H_0: $\beta = \beta_0$ contro H_1: $\beta \neq \beta_0$

si respinge l'ipotesi nulla se per un certo livello di significatività α si verifica che:

$$\left| \frac{b - \beta_0}{s \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} \right| > t_{\alpha/2,(n-2)}$$

t di Student con n-2 gdl

L'intervallo di confidenza per la stima di β è pari a :

$$b \pm t_{\alpha/2} \cdot s \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}$$

Esempio supermercati

Parametri	Stima	Errore standard	Valore test t	Livello di significatività osservato (*p-value*)
	(1)	(2)	(3) = (1)/(2)	(4)
a (intercetta)	−10,19	22,64	−0,45	0,6646
b	0,67	0,12	5,60	0,0005

p-value: probabilità di ottenere un valore 'più estremo' della statistica *t* se vera H_0

il valore del p-value indica che il test è significativo – il suo valore ha staccato un'area di probabilità pari a 0,0005 sulla coda della distribuzione (regione di rifiuto del test)

Esempio famiglie

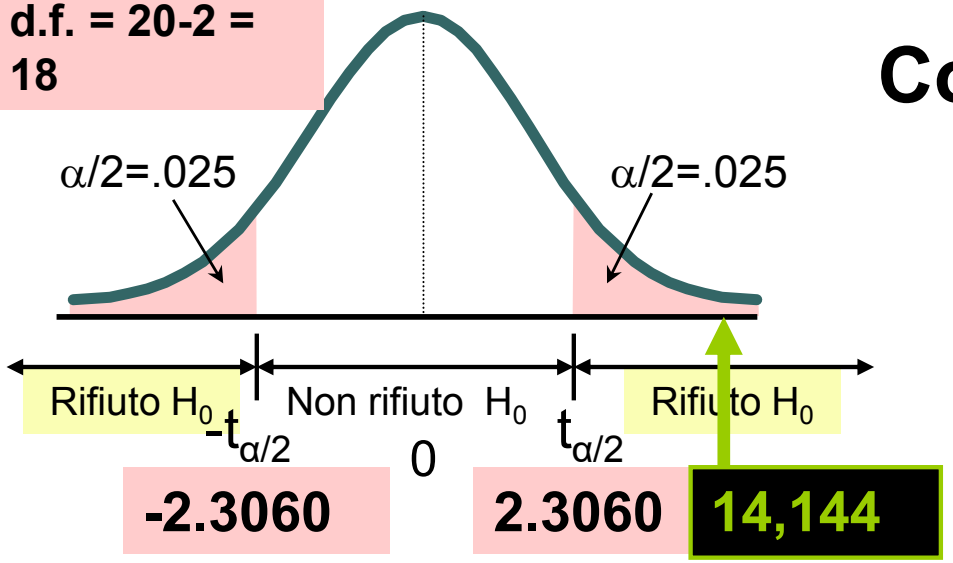
p-value:

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*
Intercetta	796.634744	84.6806	9.4075	3.6415E-10
Variabile X 1	0.482991	0.0341	14.1444	2.8075E-14

$$t = \frac{b_1 - \beta_1}{S_{b1}} = \frac{0{,}483 - 0}{0{,}0341} = 14{,}144$$

Decisione: Rifiuto H_0

d.f. = 20-2 = 18



Conclusioni:

Non ci sono elementi per ritenere che il reddito non influenzi il consumo: rifiuto H0, in quanto alla luce del campione è "troppo inverosimile"

Oltre la regressione lineare semplice

Verosimilmente, effetti molteplici di più variabili indipendenti possono agire sulla variabile dipendente Y

Se la variabile *Y* dipende da due variabili indipendenti X_1 e X_2 , il modello lineare assumerebbe la seguente forma:

$$Y_i = \alpha + \beta X_{1i} + \gamma X_{2i} + u_i$$

con:

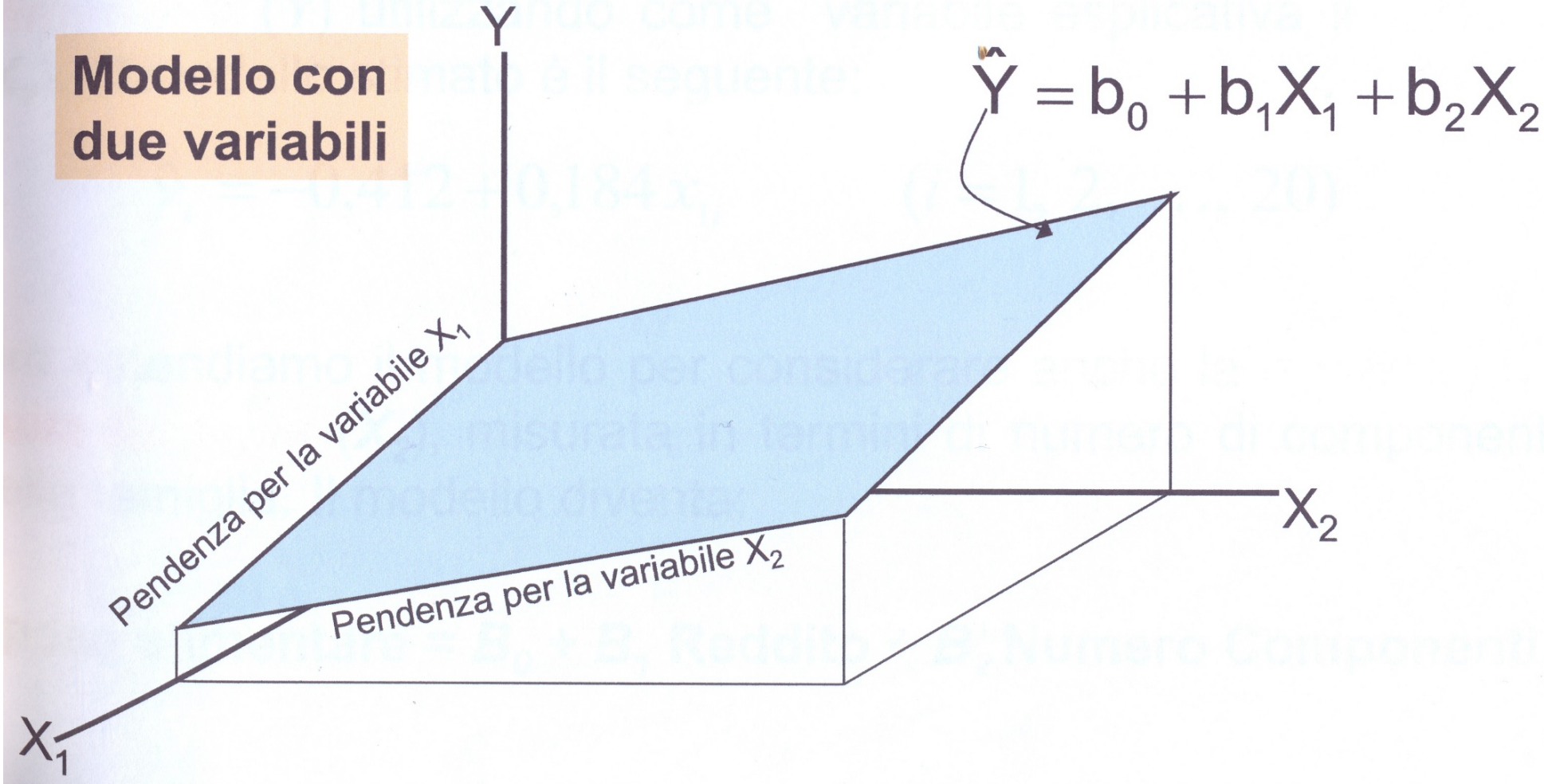
α, β* e *γ costanti che caratterizzano il modello - dette parametri del modello di regressione

u_i termine di errore

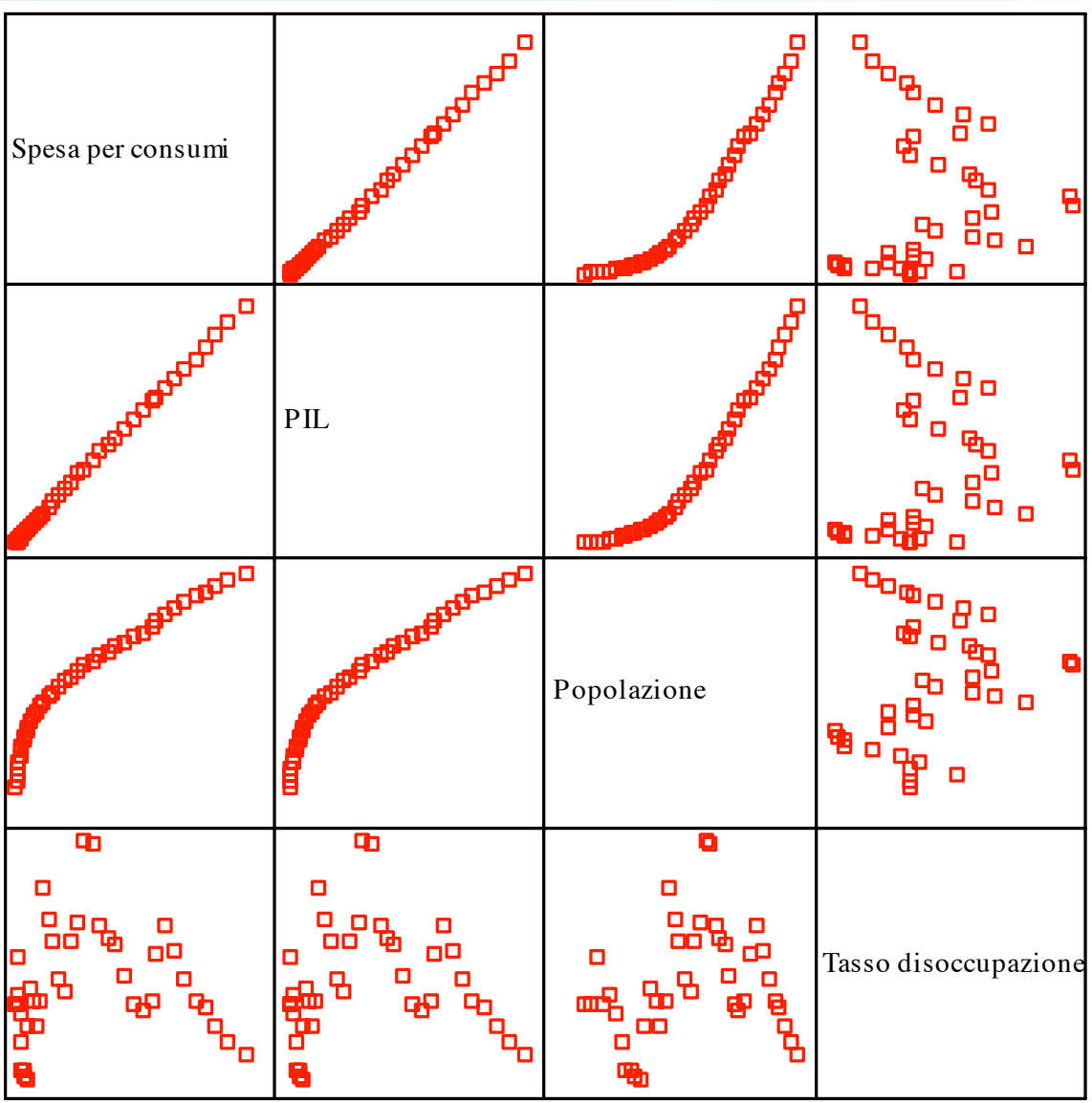
Attraverso le stime *a*, *b* e *c* dei parametri è possibile definire il piano di regressione stimato nello **spazio a 3 dimensioni**:

$$\hat{Y} = a + bX_1 + cX_2$$

Modello con due variabili



Analisi grafica: *matrice* di diagrammi di dispersione



Y = Spesa annuale per consumi
X_1 = Prodotto interno lordo – PIL
X_2 = Popolazione
X_3 = Tasso di disoccupazione – USA 1959-1999

Modello di regressione lineare multipla

- contributo di un maggior numero di variabili indipendenti per l'analisi della variabilità di Y:
- generico modello con ***k*** variabili ***X*** (di cui *k*-1 di interesse)

$$\mathbf{y} = \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \ldots + \beta_k \mathbf{x_k} + \mathbf{u}$$

- **y** un vettore delle osservazioni campionarie di dimensione $n \times 1$
- $\mathbf{x_1}$ è un vettore di dimensione *n* composto da tutti elementi unitari - il parametro β_1 rappresenta quindi l'intercetta del modello
- vettori $\mathbf{x_j}$ (j = 2, 3,..., k) di dimensione *n* dei valori delle *k*-1 variabili esplicative osservati sulle *n* unità campionarie
- **U** vettore degli *n* termini di errore (disturbi)
- β_2, β_3, ..., β_k sono i coefficienti di regressione (parametri) del modello

Notazione matriciale del modello di regressione lineare multipla

modello espresso in forma matriciale:

con

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_i \\ . \\ Y_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & X_{12} & X_{13} & \dots & X_{1k} \\ 1 & X_{22} & X_{23} & \dots & X_{2k} \\ . & . & . & \dots & . \\ 1 & X_{i2} & X_{i3} & \dots & X_{ik} \\ . & . & . & \dots & . \\ 1 & X_{n2} & X_{n3} & \dots & X_{nk} \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_j \\ . \\ \beta_k \end{bmatrix}; \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_n \end{bmatrix}.$$

$$Y_i = \sum_{k=1}^{K} \beta_k X_{ik} + u_i$$

Ipotesi modello di regressione multipla

1. *linearità* del modello
2. matrice ***X***:
   a. valori noti e misurati senza errore
   (b. **a rango pieno** ➔ $\rho(\mathbf{X}) = k$ [rango(**X**)/r(**X**), rank(**X**)/rk(**X**)]
   variabili *X* non perfettamente correlate tra loro)
3. *errore* ***u***
   media 0, varianza costante e non correlati tra loro

$$E(u_i u_j) = \sigma^2 \quad \forall i = j$$

$$E(u_i u_j) = 0 \quad \forall i \neq j$$

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad I.D.$$

Stima coefficienti di regressione

stima del **vettore** dei coefficienti di regressione **b**,
in modo che il modello sia univocamente determinato

$$\hat{\mathbf{y}} = \mathbf{Xb}$$

$\hat{\mathbf{y}}$: vettore delle ordinate teoriche **nello spazio a *k* dimensioni**

Criterio: somma dei residui (= differenza tra **y** osservato e teorico), *al quadrato,* pari al minimo

Stima coefficienti di regressione

vettore *n*-variato dei residui:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{Xb} = \begin{bmatrix} Y_1 - (b_1 X_{11} + b_2 X_{12} + \ldots + b_k X_{1k}) \\ Y_2 - (b_1 X_{21} + b_2 X_{22} + \ldots + b_k X_{2k}) \\ \ldots \\ \ldots \\ Y_n - (b_1 X_{n1} + b_2 X_{n2} + \ldots + b_k X_{nk}) \end{bmatrix}$$

metodo dei minimi quadrati:

i valori del vettore **b** sono individuati in modo da **minimizzare** la **somma dei quadrati dei residui**

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})' (\mathbf{y} - \mathbf{Xb})$$

noti i dati campionari relativi ad **y** *e ad* **X** *la somma dei quadrati dei residui può essere considerata una funzione in k dimensioni delle componenti del vettore* **b**

Minimi quadrati (OLS) in regressione multipla

Somma quadrati dei residui:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} =$$
$$= \mathbf{y}'\mathbf{y} - 2\,\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

Derivata rispetto a **b** posta = **0**

$$\min_{\mathbf{b}}(\mathbf{e}'\mathbf{e}) = \frac{\partial(\mathbf{e}'\mathbf{e})}{\partial \mathbf{b}} = -2\,\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

Risoluzione rispetto a **b** per ottenere:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}$$

Proprietà stimatori a minimi quadrati in regressione multipla

- stimatori **non distorto** $E(\mathbf{b}) = \boldsymbol{\beta}$
- a minima varianza tra **stimatori lineari e non distorti**
- **distribuzione** $\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$

(***b*** *è funzione lineare di* ***y*** *-a sua volta funzione dell'errore* ***u*** *con distribuzione normale*)

- *j*-sima componente $b_j \sim N\left(\beta_j, \sigma^2 a_{jj}\right)$

usualmente non noto, si usa stimatore ***s***

$$t = \frac{b_j - \beta_j}{s\sqrt{a_{jj}}} \sim t_{(n-k)}$$

a_{jj} *= j-esimo elemento sulla diagonale principale della matrice* $(\mathbf{X}'\mathbf{X})^{-1}$