Principi di Econometria

lezione 7

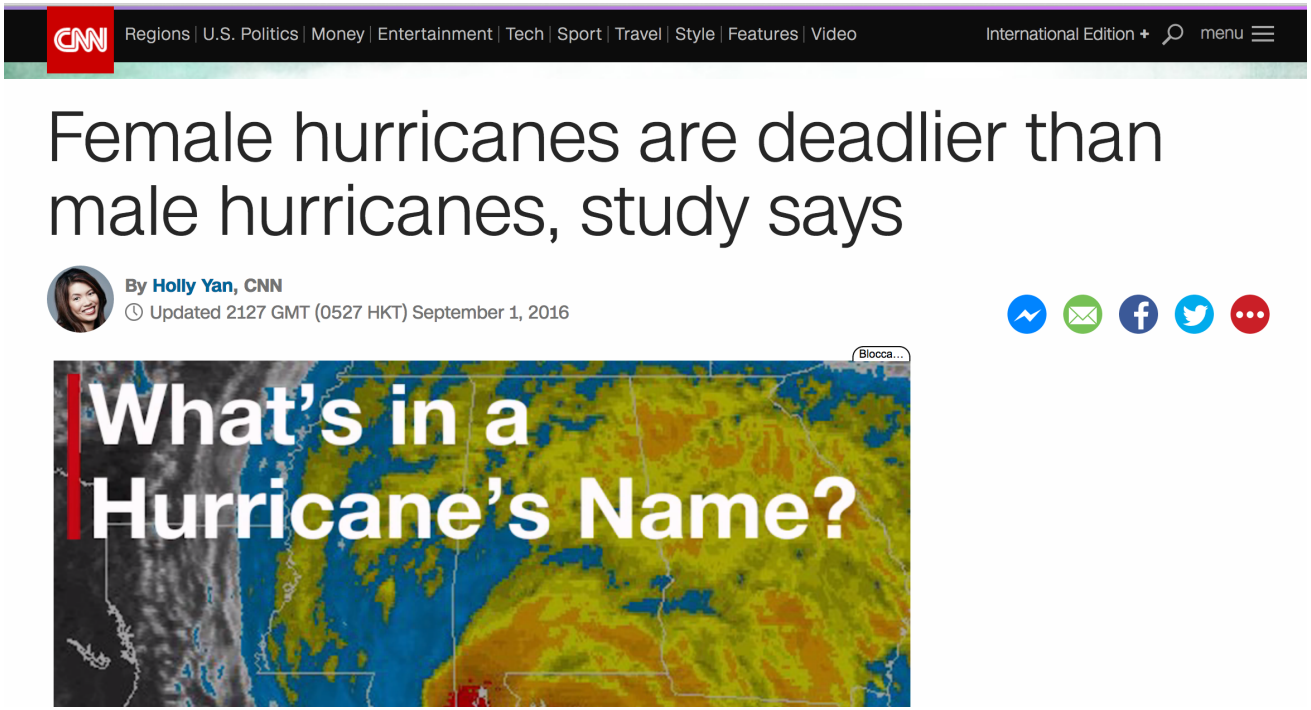
AA 2016-2017

Paolo Brunori

dove siamo arrivati?

- se siamo interessati a studiare l'andamento congiunto di due fenomeni economici
- possiamo provare a misurare i due fenomeni e poi usare la regressione lineare semplice per valutare quanto il legame fra due fenomeni sia ben approssimato da una relazione lineare
- questo metodo ci fornisce: una stima dei due parametri che definiscono la forza della relazione fra i due fenomeni, il livello di incertezza delle stime, una quantificazione della bontà di adattamento dei dati al modello (quanto bene il modello 'spiega' i dati)

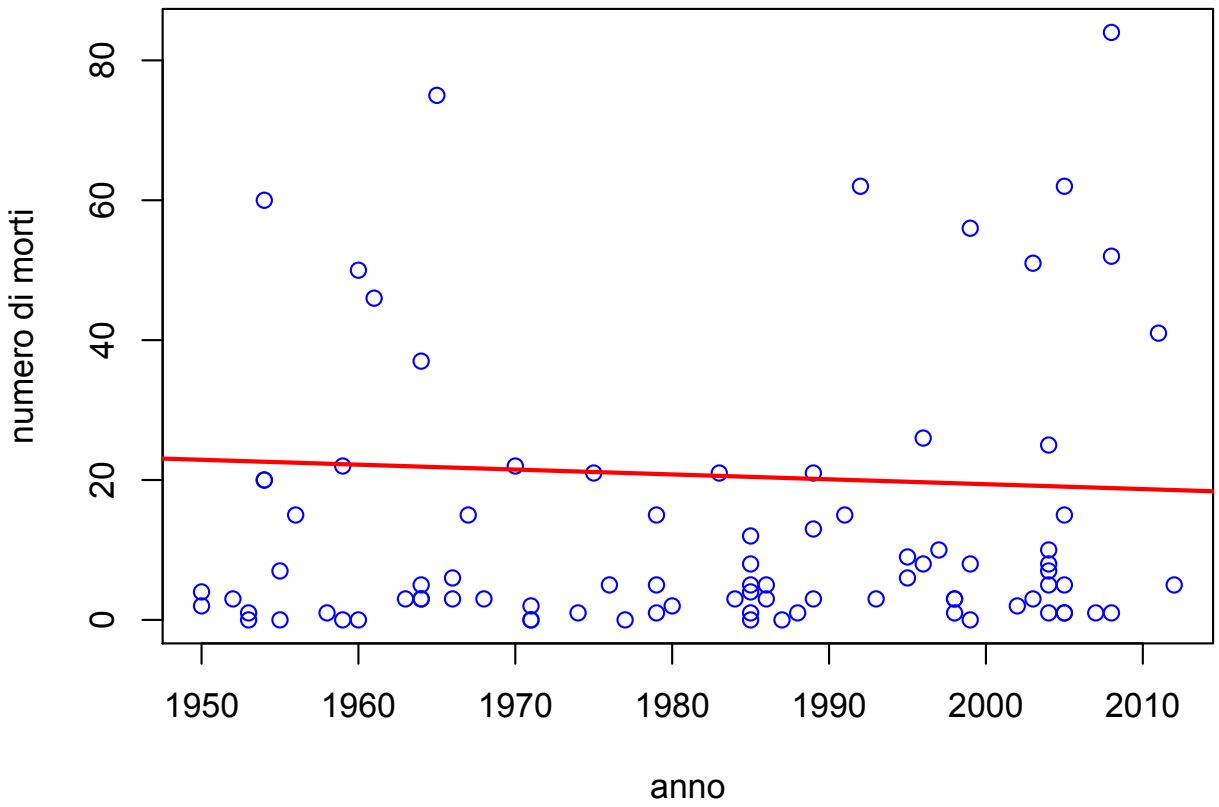
Uragani con nomi femminili sono più pericolosi di quelli con nomi maschili?



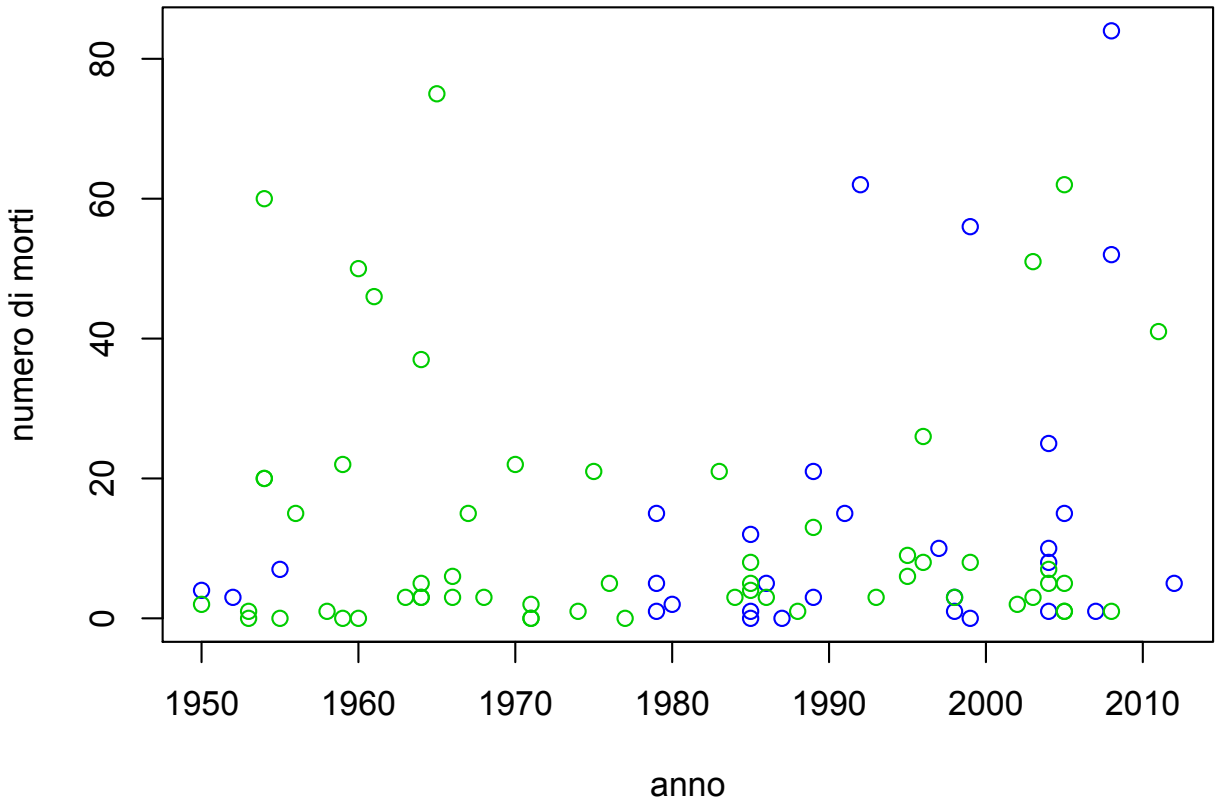
Uragani con nomi femminili sono più pericolosi di quelli con nomi maschili?

Year	Name	MasFem	MinPressure_before	Gender_MF	Category	alldeaths	damage
1950	Easy	6.77778	958	1	3	2	1590
1950	King	1.38889	955	0	3	4	5350
1952	Able	3.83333	985	0	1	3	150
1953	Barbara	9.83333	987	1	1	1	58
1953	Florence	8.33333	985	1	1	0	15
1954	Carol	8.11111	960	1	3	60	19321
1954	Edna	8.55556	954	1	3	20	3230
1954	Hazel	9.44444	938	1	4	20	24260
1955	Connie	8.5	962	1	3	0	2030
1955	Diane	9.88889	987	1	1	200	14730
1955	Ione	5.94444	960	0	3	7	6200
1956	Flossy	7	975	1	2	15	1540
1958	Helene	9.88889	946	1	3	1	540
1959	Debra	9.88889	984	1	1	0	430
1959	Gracie	9.77778	950	1	3	22	510
1960	Donna	9.27778	930	1	4	50	53270
1960	Ethel	8.72222	981	1	1	0	35
1961	Carla	9.5	931	1	4	46	15850
1963	Cindy	9.94444	996	1	1	3	300
1964	Cleo	7.94444	968	1	2	3	6450
1964	Dora	9.33333	966	1	2	5	16260

Uragani con nomi femminili sono più pericolosi di quelli con nomi maschili?



Uragani con nomi femminili sono più pericolosi di quelli con nomi maschili?



in verde uragani con nome femminile

variabili omesse

- come visto nel caso dei dati di consumo di tabacco in Turchia, è difficile ipotizzare che una sola variabile indipendente spieghi il comportamento della dipendete
- in generale infatti u cattura l'effetto di tutte quelle variabili che influenzano Y ma non sono considerate/osservabili
- l'omissione di una variabile Z distorce lo stimatore OLS se si verificano due condizioni :
  1. Z è una delle variabili che determina Y
  2. $corr(X, Z) \neq 0$

esempio: $Y=$ consumo tabacco, $X=$ prezzo, $Z=$ reddito pro capite

- condizione 1: Z ha un impatto su Y?
- condizione 2: Z potrebbe essere correlato con X?
- qual'è la direzione attesa della distorsione?

direzione della direzione da variabile omessa

- is segno della distorsione è pari al segno del prodotto fra le due correlazioni: $\text{corr}(Z, X) \times \text{corr}(Y, Z)$

$\text{corr}(Z, X) < 0$ e $\text{corr}(Y, Z) > 0 \rightarrow \hat{\beta}_1 < \beta_1$

$\text{corr}(Z, X) < 0$ e $\text{corr}(Y, Z) < 0 \rightarrow \hat{\beta}_1 > \beta_1$

$\text{corr}(Z, X) > 0$ e $\text{corr}(Y, Z) > 0 \rightarrow \hat{\beta}_1 > \beta_1$

$\text{corr}(Z, X) > 0$ e $\text{corr}(Y, Z) < 0 \rightarrow \hat{\beta}_1 < \beta_1$

esempio: X ore studiate, Y voto all'esame, Z domande copiate

- condizione 1: Z ha un impatto su Y?
- condizione 2: Z potrebbe essere correlato con X?
- qual'è la direzione attesa della distorsione?
- studenti che non studiano e copiano di più prenderanno un voto alto
- questo diminuirà la mia stima di quanto valga studiare in termini di aumento di voto β_1
- la distorsione è verso il basso $\hat{\beta}_1 < \beta_1$
- infatti il prodotto di $\text{corr}(Z, X) < 0$ e $\text{corr}(Y, Z) > 0$ è negativo

distorsione per variabili omesse espressa in termini di covarianza di X e u

- $lim_{n \to +\infty}(\hat{\beta}_1 - \beta_1) = 0$ sotto l'assunzione:

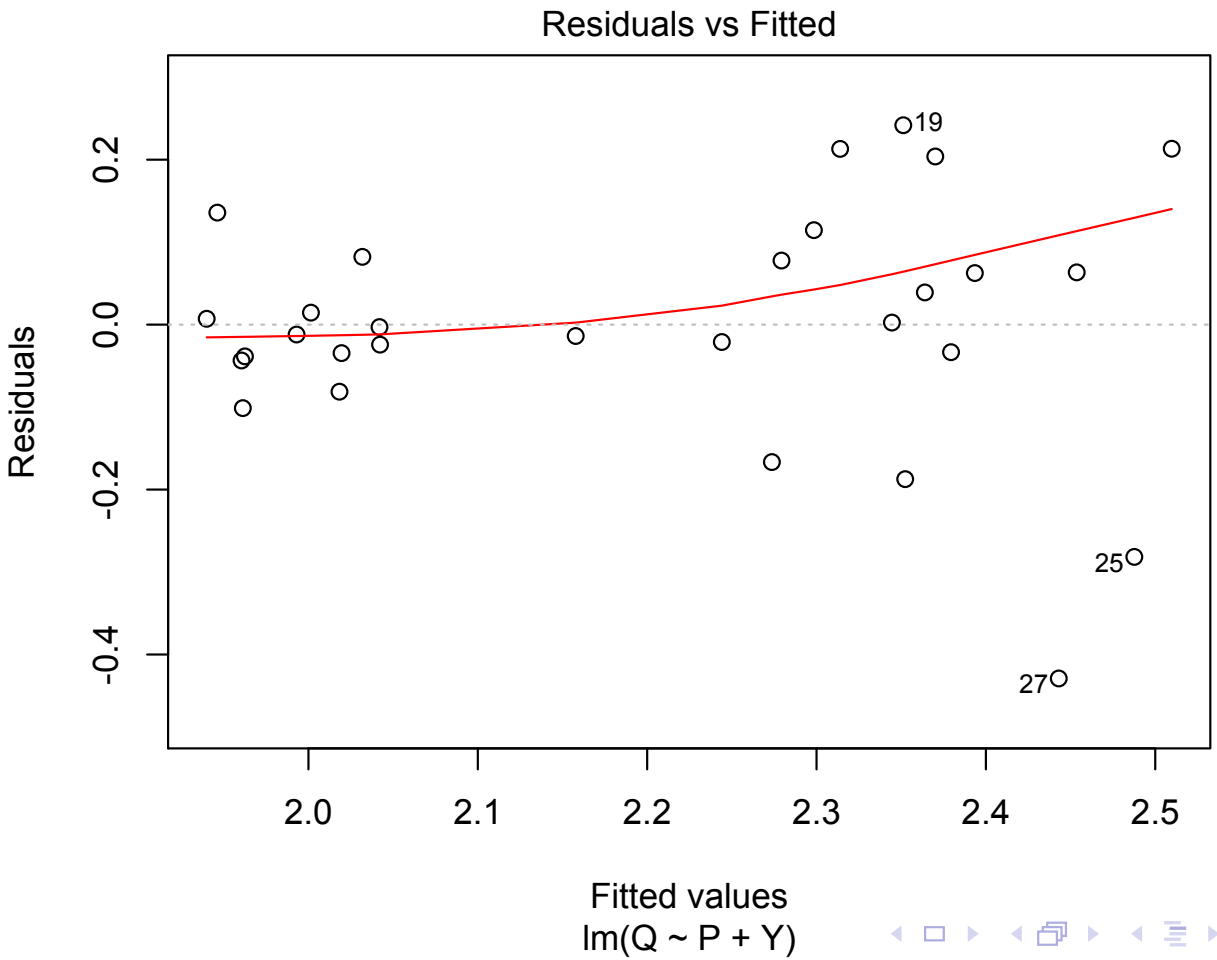
$$cov(X_i, u_i) = 0$$

- in caso contrario:

$$lim_{n \to +\infty}(\hat{\beta}_1 - \beta_1) = \frac{\sigma_u}{\sigma_X} \rho_{u,X}$$

segnali di possibili variabili omesse

- con R possiamo verificare l'assunzione creando un grafico dei residui e dei valori predetti di Y
- $\hat{Y}$ sono una combinazione lineare di X



cosa vogliamo misurare con β_1?

- β_1 può essere semplicemente la pendenza della retta che interpola un grafico a dispersione
- β_1 può invece servire a predirre $Y|X$
- β_1 come effetto causale di X su Y

che cosa si intende per nesso di causalità?

regressione con due regressori

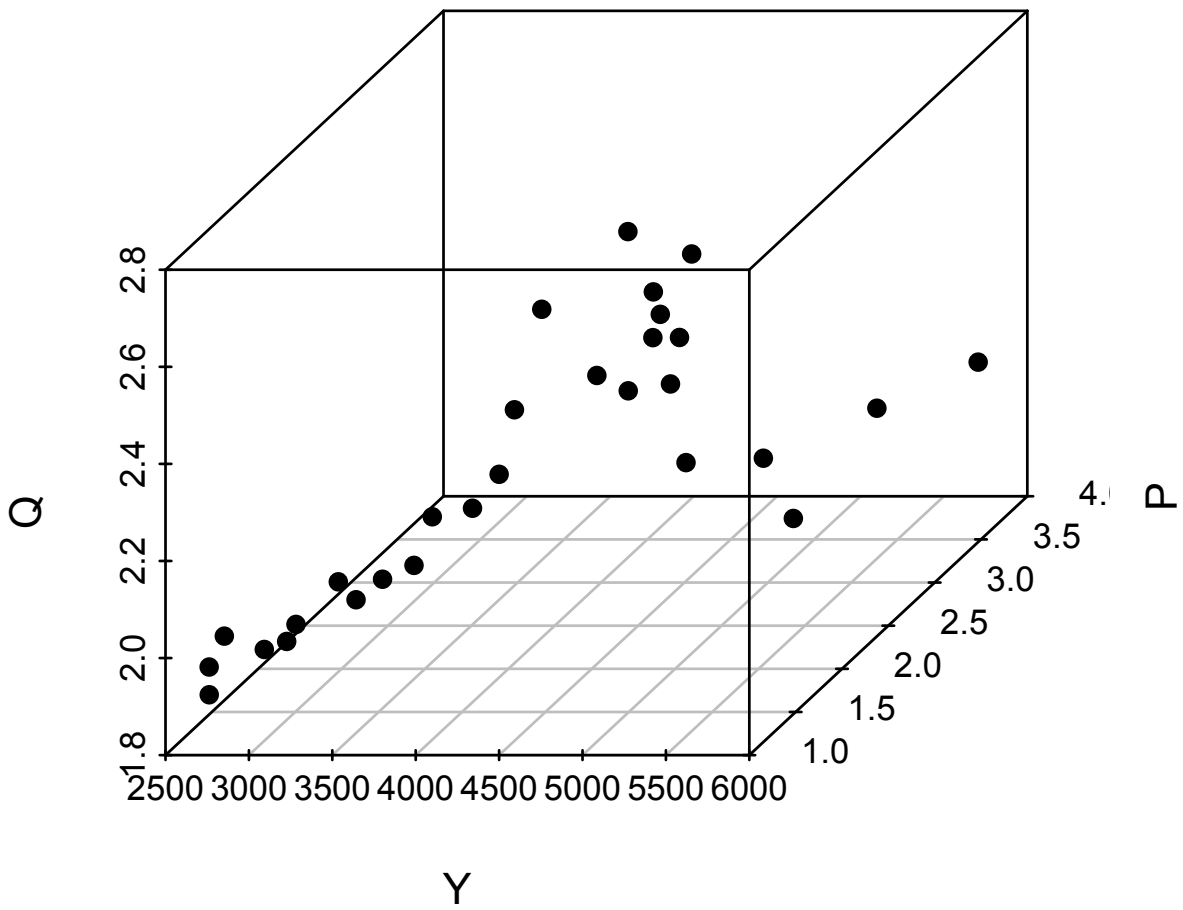
- Y, X_1, X_2
- siamo alla ricerca della relazione:

$$E(Y_i|X_{1,i} = x_i, X_{2,i} = x_2)$$

- se lineare ha la forma generica:

$$E(Y_i|X_{1,i} = x_i, X_{2,i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Consumo di tabacco, prezzo, reddito in Turchia



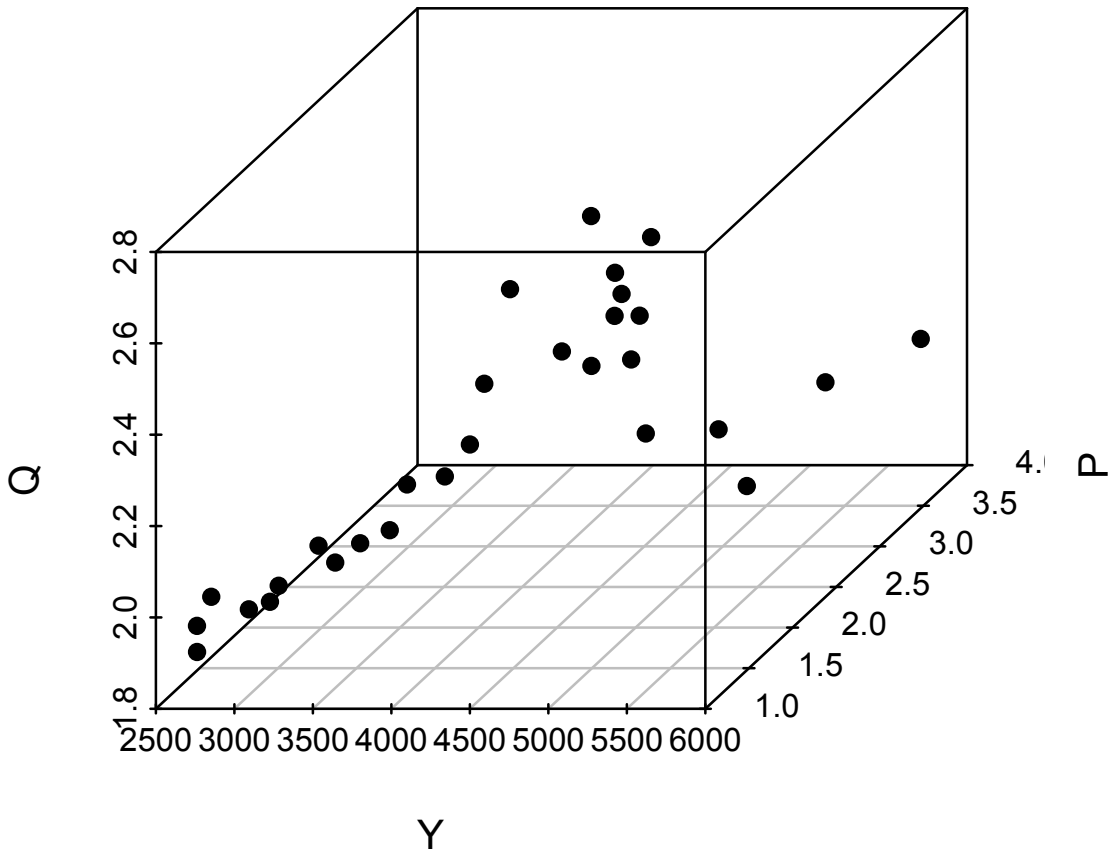
definizioni

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ è la retta (in realtà un piano!) di regressione della popolazione
- β_0 è l'intercetta
- β_1, β_2 sono i coefficienti associati alle variabili X_1, X_2
- analogamente a quanto visto per un solo regressore:

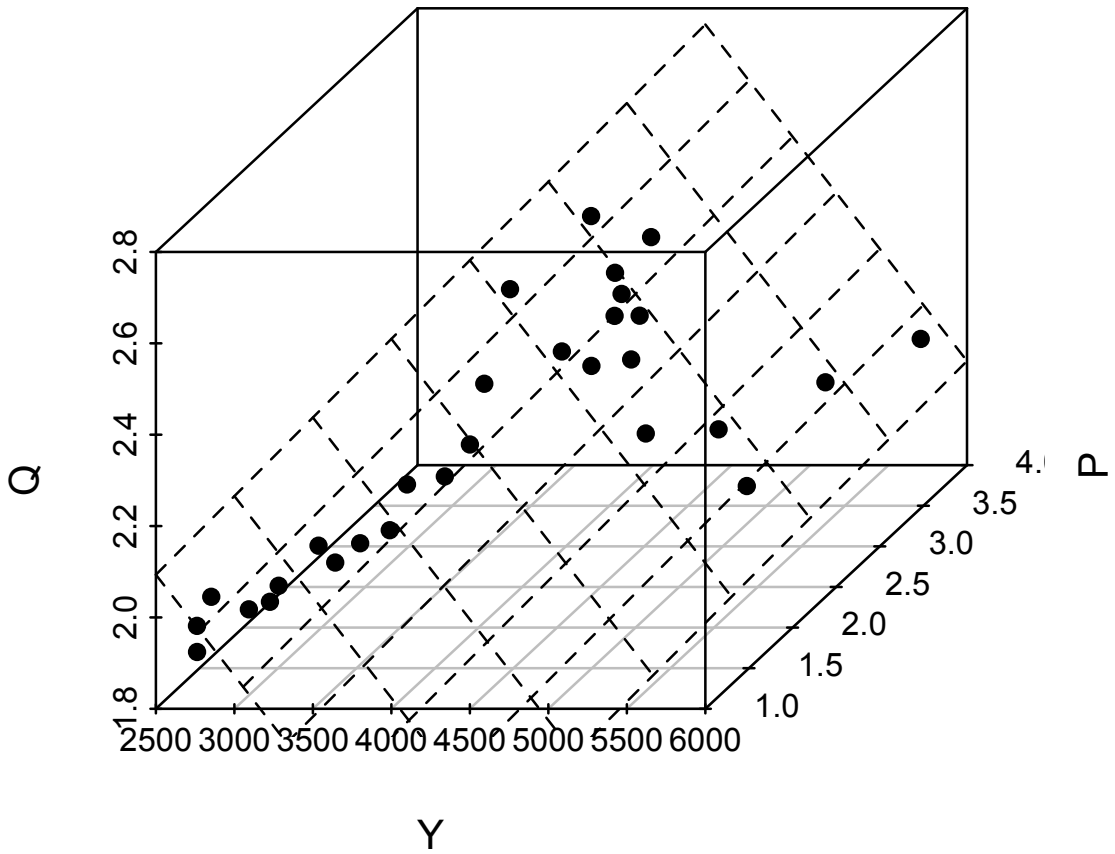
$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

- quindi $\beta_1 = \frac{\Delta Y}{\Delta X_1}$ è l'effetto parziale di X_1 (tenendo costante X_2)

il piano di regressione



il piano di regressione



regressione multipla

- seppur difficile da vedere graficamente Y può essere una funzione di molte variabili
- per un numero generico k di regressori il modello di regressione multipla prende la forma:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i$$

stimatore OLS della regressione multipla

- gli stimatori $\hat{\beta}_i$ che vengono normalmente utilizzati sono quelli che minimizzano:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1,i} - ... - b_k X_{k,1})^2$$

- le formule per ottenerli sono in formula matriciale e potete consultarli qui
- la terminologia è la stessa usata per il modello con un regressore:

  stimatori OLS: $\hat{\beta}_0, \hat{\beta}_1, ...$
  valore predetto: $\hat{Y}_i$
  residuo $\hat{u}_i$

Regressione multipla del consumo di tabacco

	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1237	13.394	0.0000
β_Y	0.0003	0.0000	6.518	0.0000
β_P	-0 .0423	0.0096	-4.3662	0.0001

- come posso interpretare questi coefficienti?

R^2

- è la frazione della varianza campionaria spiegata dai regressori

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- $R^2 \in [0, 1]$
- nella regressione multipla R^2 cresce all'aumentare dei regressori
- $k = n - 1$ regressori mi garantiscono sempre un'interpolazione perfetta (se i regressori sono k oltre l'intercetta β_0)
- nel modello del fumo di tabacco $R^2 = 0.6406$

R^2 corretto

- è possibile tener conto del numero di regressori

$$\bar{R}^2 = R^2 - (1 - R^2)\frac{k}{n - k - 1}$$

- $R^2 > \bar{R}^2$ sempre
- due effetti dell'aggiunta di un regressorie: $\uparrow\downarrow$
- $\bar{R}^2$ può essere minore di zero
- nel modello del fumo di tabacco $\bar{R}^2 = 0.6129$