15 SET 2020, 1° LEZIONE ECONOMETRIA, a. a. 2020-2021
(Bocconi, Pavia, Cattolica, Genova)

1° LEZIONE di ECONOMETRIA - Corso "QUANTITATIVO"

(PERSONE: 88 → REGISTRA)
- SUFFICIENTI gli APPUNTI! (+ PAPER)
- + PROBLEM SET (+ DIFFICILI!)
- PARTE TEORICA + EMPIRICA (Commento dei risultati)
- UTILIZZO di "STATA" (software)

COSA è l'ECONOMETRIA? (METODI STATISTICI)

Applicazione del metodo scientifico in ambito economico.

"Metodo utilizzato per cercare la verità, per capire se le ipotesi sulla realtà che abbiamo formulato sono vere o false".

→ Permette di verificare se i dati, la nostra rappresentazione della realtà, sono adeguati alla realtà.

- Partiamo da un'idea che è un modello teorico (EQUAZIONE) [IPOTESI]
  - Vediamo se rappresenta bene la realtà che sono i dati.

Ovviamente, parlando di statistica, parliamo sempre in base di probabilità.

MODELLO TEORICO → lo giustifico
↓
EQUAZIONE
↓
ADATTO a RAPPRESENTARE la REALTÀ? → APPLICO TECNICHE STATISTICHE

Avere un modello teorico, cosa vuol dire? Scriverlo? Spiegarlo?

IMMAGINIAMO il MODELLO TEORICO del SALARIO.

$wage_i$ = education (GRADO di ISTRUZIONE e ANNI di SCUOLA), unemployment (SUSSIDI di DISOCCUPAZIONE) (DISOCCUPAZIONE (TAX)), age (ETÀ), genere (M/F) (-12% degli uomini!), livello dei prezzi del Paese in cui si vive, etnia, domanda e offerta di lavoro, settore lavorativo, experience, ore lavoro

La mia produttività marginale è più alta se ho fatto quel lavoro per un po' di anni.
↓
Come scrivo il mio modello? ⇒ EQUAZIONE

$$wage_i = \beta_1 ETNIA + \beta_2 EXPER + \beta_3 EDUC + \beta_4 unempl + \dots$$

COEFFICIENTE: peso di ogni variabile

Dobbiamo cercare di quantificare il coefficiente (POS./NEG.) e dargli una misura, cioè stimare la dimensione di quel coefficiente.

Se un coeff. fosse = 0, vuol dire che ... non impatta nell'equazione, cioè che non impatta nel mio salario e quindi il mio modello non è corretto. Non posso essere sicura che la variabile con coeff = 0 giochi un ruolo nel definire quale è il salario.

Ma il modello teorico, quindi la mia equazione, è in grado di spiegare perfettamente la realtà? → NO! → Devo aggiungere una variabile che rappresenta un errore e chiamerò ε (EPSILON).

Avere una componente d'errore nel modello, cosa vuol dire? → Mi rappresenta

tutto ciò che io non sono in grado di spiegare.
Un modello, infatti, è una semplificazione della realtà. La realtà è sempre più complessa del modello che noi costruiamo.
È rappresenta "qualcosa" che mi sono scordato di considerare, il fatto che le relazioni non siano lineari, la presenza di errori di rappresentazione ...
ε "SPAZZATURA", contiene tutto ciò che noi non siamo in grado di studiare.
Dovrebbe avere 2 caratteristiche di fondo:
- che sia il più piccolo possibile
- che sia casuale, ossia non collegato con le mie variabili esplicative (determinanti)

NORMALE (GAUSSIANA) con media e varianza

ESEMPIO:

COMPONENTE COSTANTE

VETTORE

DIFF. tra OSSERV. e quello che mi dice il modello

• AGGIUNGO PEDICE SU VARIABILI

$$Wage_i = B_0 + B_1 EDUC_i + \varepsilon_i$$

SI RIFERISCE al SINGOLO INDIVIDUO.

VETTORE di DATI, ossia è una COLONNA di NUMERI (VETTORE di dimensione n)

$$\begin{array}{l} \text{CICCIO} \\ \text{TOPOLINO} \\ \text{PAPERONE} \\ \text{MINNIE} \\ \text{AMELIA} \\ \vdots \end{array} \begin{bmatrix} 10 \\ 9 \\ 25 \\ 13 \\ 6 \\ \vdots \end{bmatrix} \left.\vphantom{\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}}\right\} i = 1 \ldots n$$

$\Rightarrow$ la dimensione del mio vettore è $= n$.

• VETTORE di DATI = COLONNA di NUMERI

$$Wage_i = B_0 + B_1 EDUC + \varepsilon_i$$

$$\begin{array}{l} \text{CICCIO} \\ \text{TOPOL.} \\ \text{PAPER.} \\ \text{MINNIE} \\ \text{AMELIA} \\ \vdots \end{array} \underbrace{\begin{bmatrix} 10 \\ 9 \\ 25 \\ 43 \\ 6 \\ \vdots \end{bmatrix}}_{N \times 1} \left.\vphantom{\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}}\right\} i = 1 \ldots n \qquad \underbrace{\begin{bmatrix} 9 \\ 13 \\ 1 \\ \vdots \end{bmatrix}}_{N \times 1}$$

REALTÀ = CAMPIONE
↓
Insieme di dati disponibili per ogni individuo.

• SCALARE = numero unico uguale per tutti

VARIABILI = VETTORI = COLONNE di NUMERI che riportano le rilevazioni riguardo ad ogni soggetto.

Il VETTORE è uno scalare, un valore unico costante per tutti gli individui che per noi corrisponderà ad un livello marginale.

MODELLO BIVARIATO $\rightarrow$ MODELLO MULTIVARIATO (matrici)
(intuizione)

TEORIA → EQUAZIONE → CALCOLO COEFFICIENTE e TESTO SE È ≠ da ZERO o NO

1. DEFINIRE un METODO di STIMA che MI FORNISCA il COEFFICIENTE

COME FACCIO?

CHE CARATTERISTICHE dovrà AVERE il COEFFICIENTE?

VALORE ATTESO $E(X)$ = VALORE del COEFFICIENTE ← ▷ SIA CORRETTO, cioè che l'IMPATTO della MIA VARIABILE sul MODELLO sia CORRETTO.

▷ ABBIA la VARIANZA il più piccolo possibile.

(Visto che il coeff è stimato si porta dietro una variabilità, una VARIANZA)

COME POSSO CALCOLARE il MIO COEFFICIENTE?

▷ OLS: il mio coeff deve essere tale da rendere $\mathcal{E}$ il più piccolo possibile (SOMMATORIA ERRORI al QUADRATO)

PERCHÉ $\sum_i \mathcal{E}_i^2$?

Perché l'errore può essere anche negativo, quindi elevando al quadrato evito che gli errori si compensino e do importanza agli errori grandi!

LEZIONE 16 SETTEMBRE 2020, 2° LEZIONE.

$$Y = B_1 + B_2 x + \mathcal{E}$$

Al di là che l'econometria è il modo per cercare la verità (modello teorico, lo scrivo in equazione essendo in grado di motivare il perché scrivo → verifico se è vero quello che scrivo attraverso metodi di stima...)

1) Do valore alle mie incognite, i coefficienti β_1 e β_2.

2) Verifico se i coefficienti sono $\neq$ o $= 0$...

COME FACCIO a STIMARE questi COEFFICIENTI? → Definisco una metodologia di stima.

1) Definisco il perimetro (DELIMITO): definire le assunzioni che ottengo per fare sì che i risultati che ottengo abbiano determinate proprietà.

Pongo (6) assunzioni → Se sono soddisfatte → Pratico OLS.

Prima devo introdurre il metodo di studio.

Vediamo

▷ MODELLO DETERMINISTICO (CAMPIONE = INTERA POPOLAZIONE): i ~~modelli~~ dati che ho a disposizione rappresentano l'intera popolazione. (es. ESPERIMENTO)

▷ MODELLO STOCASTICO (CAMPIONE = PARTE di POPOLAZIONE presa in modo CASUALE): i miei dati sono una distribuzione della popolazione, (componenti casuali) quindi non l'intera popolazione. Dipendono da una DISTRIBUZIONE + AMPIA.

LE 6 ASSUNZIONI:

MODELLO LINEARE

$E(Y|X) = B_1 + B_2 X$ [VALORE ATTESO di Y DATO X]

2 EFFETTI

- B_1, B_2 = costanti ⇒ EFFETTO MARGINALE è COSTANTE
- B_2 = ME(marginale) = costante : IMPATTO di una variazione della x sulla y

||

"la variazione che ottengo nella Y data da una variazione UNITARIA che ho nella x è costante ed è uguale per tutti".

↓

NO nel MODELLO di MASSIMA VEROSIMIGLIANZA

VARIABILITÀ

: "NON TUTTI i VALORI di X sono UGUALI"

"Data la mia distribuzione $\{x_i, y_i\}$, $i = 1, \dots, n$ (NUMERO di INDIVIDUI che ho = DIMENSIONE del MIO SEMPLE), esiste un $i \neq j$ tale che $x_i \neq x_j$"

$\{x_i, y_i\}\ i = 1, \dots, n,\ \exists\, i \neq j \,/\, x_i \neq x_j$

↓

La VARIABILITÀ, ossia il fatto che i dati non siano tutti uguali, è la "ricchezza" che mi permette la stima. → Se non ho VARIABILITÀ, NON POSSO STIMARE NULLA!

Se $x_i \neq x_j$ → HO VARIABILITÀ! [DEVO AVERE ...]

⟹ 1° EFFETTO

$\underbrace{\frac{1}{N-1} \sum_i (x_i - \bar{x})^2}_{\text{VARIANZA CAMPIONARIA}}$: VARIANZA CAMPIONARIA deve essere > 0 ↓ HO VARIABILITÀ dei DATI

2° EFFETTO : non ho MULTICOLLINEARITÀ : due variabili non sono la combinazione lineare una dell'altra

↓ ogni variabile mi aggiunge delle informazioni.

NON POSSO AVERE MULTICOLLINEARITÀ ↓ ogni variabile mi aggiunge delle informazioni

↓ 1 VETTORI di DATI (variabili) può essere ottenuto come comb. line. tra gli altri vettori

Ho MULTICOLLINEARITÀ quando io posso ottenere un vettore di dati come COMBINAZIONE LINEARE degli altri VETTORI.

↓

Se NON HO MULTICOLLINEARITÀ

↓

le mie variabili ~~sono~~ NON SONO RIDONDANTI

ES:

Vettore 1 : ETÀ

Vettore 2 : ETÀ + 5

RIDONDANTI ⇒ implicano che non posso invertire la matrice.

3 $\forall i \neq j$ (2 individui ≠) sto parlando di, la covarianza fra y_i e y_j (della mia variabile dipendente) dato X è $= 0$.

$cov(Y_i, Y_j | X) = 0$ ⇒ Il mio campione deve essere CASUALE

Al di là delle variabili indipendenti non deve esserci relazione tra i 2 individui. → Ci deve essere CASUALITÀ

$\Rightarrow cov(\varepsilon_i, \varepsilon_j | X) = 0$: le componenti di errore (individuali) tra gli individui devono essere incorrelate.

Non ci deve essere CORRELAZIONE tra gli ERRORI

Deriva dal fatto che il mio sample deve essere casuale!

4 Riprendendo la distinzione tra MODELLO DETERMINISTICO (Intera popolazione) e STOCASTICO (Parte di popolaz.)

▷ **DETERMINISTICO**

$E(\varepsilon) = 0 \Rightarrow E(Y) = \beta_1 + \beta_2 X \Rightarrow E(Y) = \beta_1 + \beta_2 X$

$E(\beta_1) + E(\beta_2 X)$

▷ **STOCASTICO**

DATI PROVENGONO da DISTRIBUZ. CASUALE

Ho un'IMMAGINE della REALTÀ ⇒

$E(\varepsilon | X) = 0$: condizione di ESOGENEITÀ

$\Downarrow \quad \not\Uparrow$

$cov(\varepsilon, X) = 0$, ma non è vero il contrario.

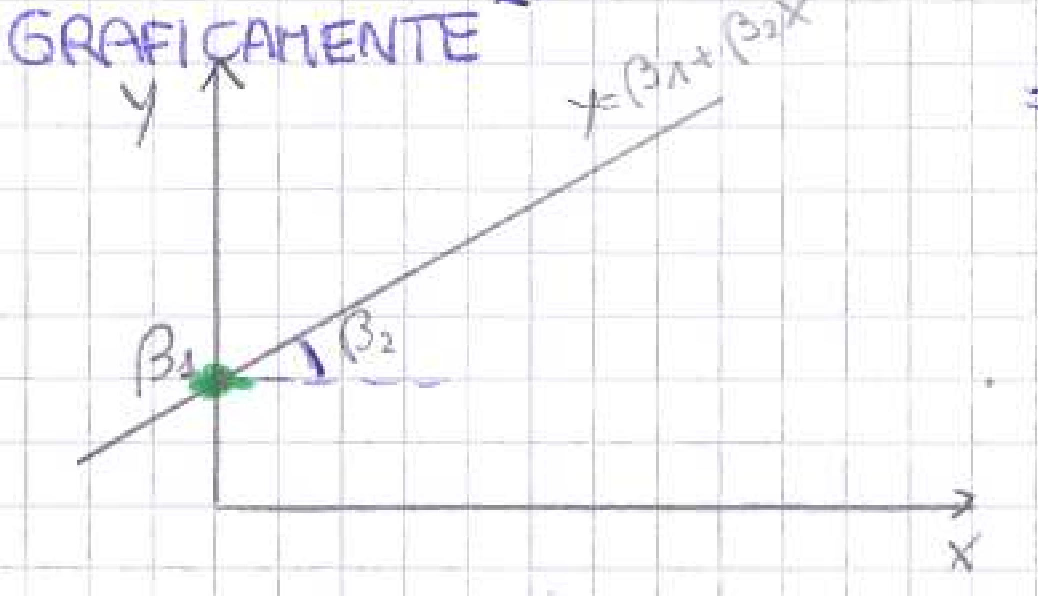
(covarianza tra errori e variabili indipendenti $= 0$).

INCORRELAZIONE tra ε e VAR. INDIPENDENTI

da questo ottengo che ⇒ $\beta_1 = E(Y|X)$ quando $X = 0$

β_1 è l'INTERCETTA; $E(Y|X)$ retta

GRAFICAMENTE



$\Rightarrow \beta_2 = \partial E(Y|X) / \partial X$ (DERIVATA di...)

β_2 = PENDENZA! = COEFFICIENTE ANGOLARE = costante indipendentemente dal livello delle X (A1)

β_1 e β_2 = ~~COEFF~~ INTERCETTA e COEFF. ANGOLARE

"FOC: First order condition" = Sono le derivate prime, ossia quelle VALORI che massimizzano o minimizzano la mia funzione obiettivo.

5 $V(\varepsilon | X) = E(\varepsilon^2 | X) = \sigma^2$ → la VARIANZA del mio ERRORE deve essere COSTANTE

VARIANZA dell'ERRORE dato X; VALORE ATTESO; SIGMA², VALORE GENERICO → Non c'è nessun pedice → È COSTANTE

Questo per la condizione 4 ($E(\varepsilon|X) = 0$). (VAR ≠ = ETEROSCHEDASTICITÀ)

⇒ La varianza del mio errore deve essere costante : OMOSCHEDASTICITÀ $V(\varepsilon|X) = \sigma^2$

↓ LA VARIANZA dell'ERRORE è COSTANTE tra gli INDIVIDUI

▷ $V(\beta_1) = 0$ perché β_1 costante

▷ $V(\beta_2) = 0$ perché costante

▷ ε è l'unica componente che ha con variabilità (ALEATORIA)

$V(Y|X) = \sigma^2$ * è l'unica componente aleatoria che ha variabilità

$V(\beta_1 + \beta_2 X + \varepsilon | X) = \sigma^2$

OMOSCHEDASTICITÀ: la varianza degli errori è costante tra gli individui.

ATTENZIONE!

⚠ Le ASSUNZIONI da A_1 ad A_5 le chiamiamo ASSUNZIONI DEBOLI.
Se aggiungiamo la A_6 diventano ASSUNZIONI FORTI.

ESEMPIO: Ho MASCHIO / FEMMINA → non inserisco ERRORE ⇒ Dovrei avere 2 varianze ≠, quindi avrei ETEROSCHEDASTICITÀ.

6) Se mio errore dato X si distribuisce come una normale con media 0 e varianza σ^2 (SIGMA 2)

$$\varepsilon | X \sim N(0, \sigma^2)$$

A_6, A_4, A_5

A6: Se fatto che la distribuzione sia normale

$\sim$: si distribuisce secondo...

$$\Rightarrow Y|X \sim N(\underbrace{\beta_1 + \beta_2 X}_{\text{MEDIA}}, \underbrace{\sigma^2}_{\text{VARIANZA}})$$

Componente che conosco → A4

Componente aleatoria del modello → A5

ORA ↓

Per calcolare i coefficienti devo costruire un metodo di STIMA, che mi permetta di verificare che i coefficienti sono corretti. (OTTENERE dei VALORI dei COEFFICIENTI TALI PERCUI QUESTI RISULTATI che OTTENGO ABBIANO CERTE PROPRIETÀ) → COEFFICIENTI: GIUSTI, VARIANZA MINIMA

Quindi, partendo dalle 6 assunzioni che ci delimitano i metodi di stima, andiamo a determinare il metodo di stima.

CRITERIO di STIMA: **MODELLO OLS**: (ORDINARY LEASE SQUARE) (basato su $Y = \beta_1 + \beta_2 X + \varepsilon$)

• Voglio trovare quei coefficienti $\hat{\beta}_1, \hat{\beta}_2 \ldots$ che rendano il più piccolo possibile $\sum \varepsilon_i^2$. (TUTTO CIÒ che NON RIESCO a SPIEGARE sia ...)

METODO: $\min_{\beta_1, \beta_2} \sum_{i=1}^{n} \varepsilon_i^2$ METODO di STIMA : calcolerò i miei parametri β_1 e β_2 → $[Y = \beta_1 + \ldots]$ in modo che rendano la sommatoria del quadrato degli errori il più piccolo possibile.

Rispetto a cosa? → MINIMIZZO rispetto a β_1, β_2

β_1, β_2 = PARAMETRI → calcolo → STIMATORI = $\hat{\beta}_1, \hat{\beta}_2$

β_1, β_2 = COSTANTI. VETTORI di DATI → serve PEDICE

COSA VUOL DIRE?

$$\min_{\beta_1, \beta_2} \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_i)^2 = D(\beta_1, \beta_2)$$

FUNZIONE OBIETTIVO da MINIMIZZARE

ottengo ⇒ $\hat{\beta}$ = VALORE STIMATO (VETTORE) = $\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ = argomento che minimizza la mia funz. obiettivo $D(\beta_1, \beta_2)$ rispetto a β_1 e β_2.

→ Vettore che si porta dietro 2 valori

$$\arg\min_{\beta_1, \beta_2} D(\beta_1, \beta_2)$$

STIMATORE OLS ↳ $\hat{\beta}$ OLS

COSTANTI x MODELLO LINEARE

VALORI che MINIMIZZANO $\sum_{i=1}^{n} \varepsilon_i^2$

Per calcolare β_1 e β_2 devo fare una minimizzazione $\Rightarrow \min_{\beta_1,\beta_2} \underbrace{\sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 x_i)^2}_{D(\beta_1,\beta_2)}$

minimizzazione ↓ mi servono le FOC ↘ le FOC sono le derivate.

$FOC_1 = \dfrac{\partial D(B_1,B_2)}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$

↓ $= 0$

FOC ↙ DERIVATA della mia funzione obiettivo rispetto a ... ↖

$FOC_2 = \dfrac{\partial D(B_1,B_2)}{\partial B_2} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)\cdot x_i$

⇓ $= 0$

⚠ β_1: INTERCETTA

β_2: COEFF. ANGOLARE

↘ Risolvendole, si ottiene la formula del modello OLS.

LEZIONE del 22 SETTEMBRE 2020

MODELLO TEORICO con IPOTESI

↓

COSTRUISCO un MODELLO di STIMA calcolando i coefficienti che facciano sì che il quadrato dei miei errori sia il più piccolo possibile.

1) COSTRUISCO FUNZ. OBIETTIVO
2) MINIMIZZO → FOC₁ e FOC₂

Per ottimizzare una

MASSIMIZZAZIONE / MINIMIZZAZIONE

↓

DERIVATA (FOC)

della funzione OBIETTIVO rispetto a β_1 e β_2.

La mia FOC_2 altro non è che $FOC_1 \cdot x_i$.

Risolviamo ora le FOC per cercare di dare un valore a β_1 e β_2. → Cerco di isolarle!

$FOC_1 = \dfrac{\partial D(B_1,B_2)}{\partial B_1} = -2\dfrac{\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)}{n} = \dfrac{0}{n}$

• TOLGO il -2
• DIVIDO per n

↓ $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$

$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$

! $\sum \dfrac{\beta_1}{n} = \beta_1$ perché costante

$FOC_2 = \dfrac{\partial D(B_1,B_2)}{\partial B_2} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)\, x_i = 0$

• TOLGO -2
• DIVIDO per n → $\underbrace{\dfrac{1}{n}\sum y_i x_i = \hat{\beta}_1 \cdot \bar{x} + \hat{\beta}_2 \sum \dfrac{x_i^2}{n}}$

da risolvo con alcuni passaggi matematici.

$$\begin{cases} \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} & [FOC_1] \\ \dfrac{1}{n}\sum y_i x_i = \cancel{\hat{\beta}_1}\bar{x} + \hat{\beta}_2 \dfrac{1}{n}\sum x_i^2 & \rightsquigarrow \text{Posso riscriverla in altri termini. } [FOC_2] \end{cases}$$

• Moltiplichiamo a dx e sx il risultato di FOC_1 per $\bar{x}$ → $\bar{y}\bar{x} = \cancel{\hat{\beta}_1}\bar{x} + \hat{\beta}_2 \bar{x}^2$

• Prendo FOC_2 e ⇒ $FOC_2 - FOC_1\,\bar{x}$ (prima parte sx, poi la dx!)

$\Rightarrow \dfrac{1}{n}\sum y_i x_i - \bar{y}\bar{x} = \hat{\beta}_2\left(\dfrac{1}{n}\sum x_i^2 - \bar{x}^2\right)$

COME FACCIO a RISOLVERLA?

$$\frac{1}{n}\sum_i^1 y_i x_i - \underbrace{\bar{y}\bar{x}}_{\substack{\text{Posso scriverla come} \\ \bar{y}\frac{1}{n}\sum_i x_i}} = \hat{\beta}_2\left(\frac{1}{n}\sum_i^1 x_i^2 - \underbrace{\bar{x}^2}_{\frac{1}{n}\sum_i x_i \cdot \bar{x}}\right)$$

$$\Rightarrow \cancel{\frac{1}{n}}\sum_i^1 y_i x_i - \bar{y}\cancel{\frac{1}{n}}\sum_i^1 x_i = \hat{\beta}_2\left(\cancel{\frac{1}{n}}\sum_i x_i^2 - \cancel{\frac{1}{n}}\sum_i^1 x_i \cdot \bar{x}\right)$$

$$\Rightarrow \sum_i^1 y_i x_i - \bar{y}\sum_i^1 x_i = \hat{\beta}_2\left(\sum_i^1 x_i^2 - \sum_i^1 x_i \bar{x}\right)$$

Il mio obiettivo, ora, è calcolarmi i coefficienti che minimizzino la sommatoria dei quadrati.

$\sum_i^1$ variaz. della media $= 0$.

$\sum_i (x_i - \bar{x}) = 0$

$\Rightarrow$ Risolvo rispetto a β_2

$$\Rightarrow \beta_2 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})(x_i - \bar{x})}$$

$\equiv$ COVARIANZA STIMATA fra x e y.

$\wedge$ = STIMATO.

$\equiv$ NUMERATORE della VARIANZA

$$\Rightarrow \hat{\beta}_2 = \frac{\hat{cov}(x,y)}{\hat{V}(x)}$$

EFFETTO MARG.

($\hat{V}(x)$ la vogliamo grande, basta che sia $\neq 0$)

Perché stimati? → Perché ha a che fare col campione! NON È REALTÀ COMPLESSIVA!

Come lo si interpreta?

"RAPPORTO fra VARIABILITÀ $\frac{\text{Come si muovono insieme x e y}}{\text{Variabilità dei dati} = \text{VARIANZA}}$"

RAPPORTO fra VARIABILITÀ → Una delle ASSUNZIONI

"Quanto la x e la y si muovono insieme dato quanto si muove nel complesso la x, cioè la variabilità totale della x".

• $\hat{V}$ grande → Poca capacità della x di spiegare il modello.

$\Rightarrow$ Nel modello OLS, la pendenza, il mio coefficiente β_2 (EFFETTO MARGINALE) altro non mi dice quanto la x e la y si muovono insieme rispetto alla variabilità della x.

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$$

$$\hat{\beta}_2 = \frac{\hat{cov}(x,y)}{\hat{V}(x)}$$

CHE ASSUNZIONI ABBIAMO UTILIZZATO?

• II° – A_2 – VARIABILITÀ.

• I° – A_1 – MODELLO LINEARE.

Introduciamo ora un nuovo concetto

DIFFERENZA tra RESIDUI ed ERRORE

- ERRORE = teorico, non calcolabile.

Una volta che ho ottenuto i miei coefficienti β_1 e β_2, ho il mio modello.

$\underbrace{\hat{\beta}_1, \hat{\beta}_2}_{\text{COEFF. che ho stimato}} \xrightarrow{\text{APPLICO MODELLO}}$ $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ ≡ VALORE FITTATO ≡ valore che ottengo quando al mio modello sostituisco i valori trovati. Stimati $\hat{\beta}_1$ e $\hat{\beta}_2$.

$y_i - \hat{y}_i$: Esiste una differenza fra il mio valore reale e il mio valore fittato (ottenuto se sostituendo nel modello i valori stimati)

$y_i - \hat{y}_i = e_i$ (RESIDUO) (è come se fosse la STIMA dell'errore) → Differenza tra vero valore e valore FITTATO.

$$\Rightarrow y_i = \underbrace{\hat{y}_i}_{\text{Dipende dalla COMPONENTE SPIEGATA}} + e_i$$

e_i → Dipende dalla COMPONENTE che non sono in GRADO di SPIEGARE. └ COMPONENTE RESIDUALE

MODELLO STIMATO

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \underbrace{e_i}_{\text{RESIDUI}}$$

$\neq$

MODELLO VERO

$$y = \beta_1 + \beta_2 x + \underbrace{\varepsilon_i}_{\text{ERRORI}}$$

la mia STIMA (MODELLO TEORICO) non sarà mai uguale al mio modello vero.

$\Rightarrow \beta_1 \neq \hat{\beta}_1$

$\beta_2 \neq \hat{\beta}_2$

$\varepsilon \neq e_i$

⚠ NON SONO A PRIORI certo che $\beta_1 = \hat{\beta}_1$, $\beta_2 = \hat{\beta}_2$ ....

STIMA → PROBABILITÀ di OTTENERE un CERTO RISULTATO

(Non è detto che il valore che trovo sia quello effettivo)

Fino ad ora abbiamo visto un modello BIVARIATO → solo una variabile dipendente.
Passiamo ad un modello MULTIVARIATO!

$y = \beta_1 + \beta_2 X_i + \varepsilon_i$ → NON è REALISTICO!

MODELLO MULTIVARIATO!

$Y = \beta \quad X \quad \varepsilon_1$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \begin{pmatrix} 1 & x_{21} & x_{31} & \cdots & x_{k1} \\ 1 & x_{22} & x_{32} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{2n} & x_{3n} & \cdots & x_{kn} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

COSTANTE — K-1 VARIABILI

VETTORE = VETTORE MATRICE + VETTORE

$n \times 1 = \underbrace{K}_{} \times 1 \quad n \times K \quad n \times 1$

Se ho K variabili → devo avere K coefficienti.