METODI E MODELLI A SUPPORTO DELLE DECISIONI (II) ora 1

Titolo nota 29/09/2009

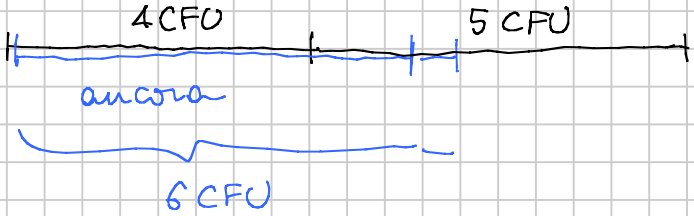
2 moduli : I Nicolodi → programmazione lineare 6 CFU
II Morandin → statistica multivariata 6 CFU

5 + 9 = 14 ⇨ 9 + (3) + 6 = 15 (18)

II modulo :

lun 14.30 – 16.30 LAB
mar 13.30 – 16.30 teoria
} ~50 ore = 24 + 36

rispetto all'anno scorso 24 lab + 24 teoria in comune
+12 teoria : ripasso + approfondimento di statistica + altro



PROGRAMMA :

Ross : 1 — 11 (1–8 già fatti quasi del tutto)
completare 8 : curva OC / potenza dei test / err I specie
9 : regressione
10 : ANOVA
11 : test di adattamento

approfondimenti :
→ uso di Excel
→ regressione avanzata : Afifi "Computer – aided mvariate analysis"
→ D.O.E. : Sleeper "Design for 6 sigma statistics"

BIBLIOGRAFIA :

Ross, Afifi, Sleeper
www.statsoft.com → stat. textbook
+ Middleton per Excel
Apogeo

dispense di Soliani → 2800 pagine

STUDENTI
- top → capire perfettamente → innovazione
- middle → impara a passare l'esame → fatica per niente
- bottom → cercare di non bloccarsi

- Ricevimento: martedì 16.30 – 18.30 dal 5 in poi
  (questa settimana mer 9.00 – 10.00)

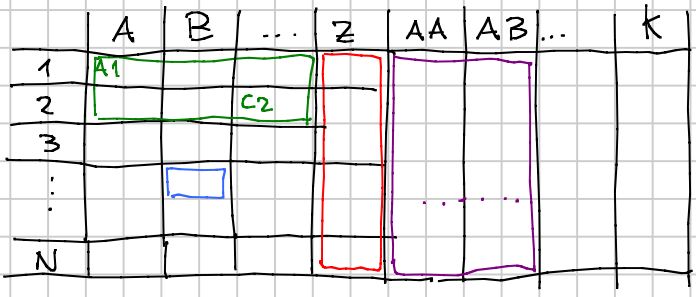
ESAME
- ★ test a crocette → idoneità per i due scritti (entrambi, ma soprattutto Morandin)
- ★ prova pratica (Excel) (Morandin)
- ★ prova scritta (Nicolodi)
- ★ orale (entrambi, ma soprattutto Nicolodi)

COSE PRATICHE
corsi.unipr.it → iscriversi a Metodi Statistici a Supporto delle Decisioni

- → foto
- → avvisi
- → forum
- → documenti
  - → pdf (anche vecchi)
  - → avi (scaricarli subito)
  - → .xlsx (office 2007)
  - → scritti vecchi

ora 2

COME FUNZIONA EXCEL



$N = 65536 = 2^{16}$ Office 2003 e precedenti

$= 1048576 = 2^{20}$ Office 2007

$k = 256 = 2^8$ Office 2003 e precedenti

$= 16384 = 2^{14}$ Office 2007

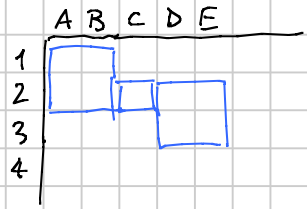
→ riferimenti: a una cella B4 = B4^2

a un rettangolo A1:C2 = SOMMA(A1:C2)

rettangoli "impropri" : AA:AB Z:Z

- Una cella Excel può contenere:
  a) Numeri
  b) Stringhe
  c) Formule (che possono restituire numeri o stringhe)
  d) (un pezzo di matrice)

- Formule: iniziano con = oppure −
  si possono usare + − * / ^ ( )
  e le relazioni < > <= >= = <> (↑ diverso)
  poi ci sono le funzioni interne
  =SOMMA(A1:B2 ; C2 ; D2:E3)
  il ; separa gli argomenti
  =SUM(...)



in generale i comandi sono tradotti automaticamente se si carica il file con una versione di lingua diversa, ma:

CASUALE.TRA	RANDBETWEEN
FRAZ.ANNO	YEARFRAC

invece danno err

→ elenco comandi più usati:
SOMMA, MEDIA, VAR, DEV.ST, INV.NORM.ST, DISTRIB.NORM.ST, INV.NORM, DISTRIB.NORM, DISTRIB.T, DISTRIB.CHI, INV.T, INV.CHI, SE CASUALE

★ CONSIGLIO: Fate molti esercizi dei Capitoli 7 e 8 del Ross usando Excel.

- Numeri: contenuto delle celle e visualizzazione sono due concetti distinti e complementari

MMSD

Titolo nota — 06/10/2009 — ora 3

Cambio orario : lun laboratorio 2 e 3 dalle 16:30 alle 18:30

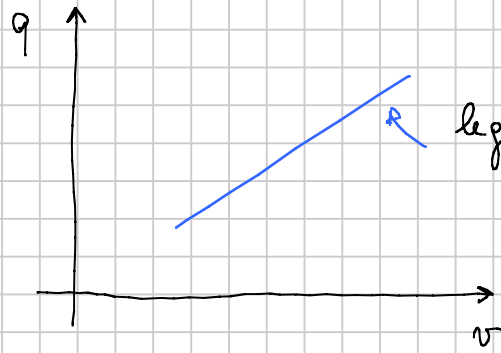
REGRESSIONE

- r. lineare semplice
- r. " multipla
- r. nonlineare

...

- La regressione studia il rapporto tra due variabili numeriche

es 1 : dosatore automatico di polvere → vel di rotazione
quantità di polvere

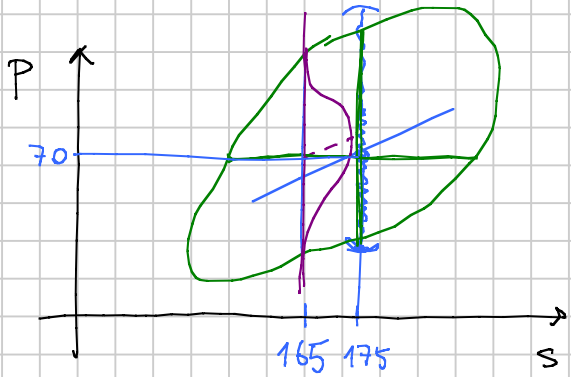


legame tra due variabili = funzione

$q = q(v)$

$v = v(q)$

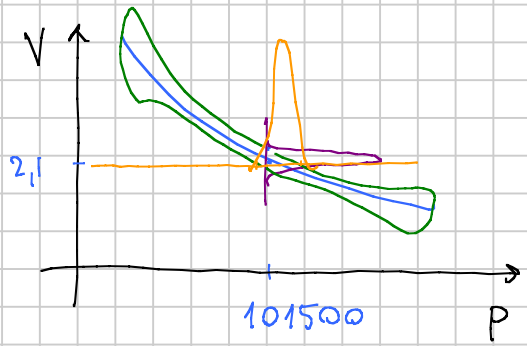
es 2 : misuro esseri umani → peso
statura



il legame tra queste variabili è "casuale" : sapere s permette di conoscere la distribuzione del peso ; ad esempio

$s = 165$ $\quad p \sim N(60, 12^2)$

es 3 : barometro elettronico : il sensore trasforma → pressione
diff di potenziale

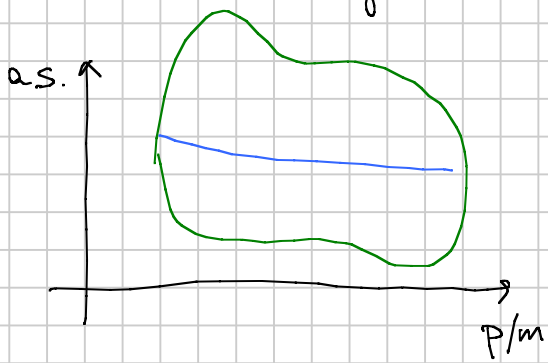


(situazione comune a tutti gli strumenti di misurazione)

- La regressione generalizza lo studio sul rapporto tra due variabili ai casi in cui c'è un margine di incertezza

- La prima risposta della regressione è se ci sia una qualche relazione tra le variabili, o se invece siano indipendenti

es 4: c'è un legame tra → pulsazioni a riposo e livello culturale (= anni di studi) ?



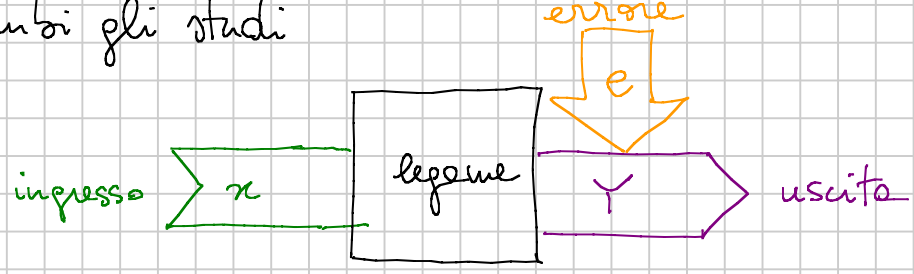
NB "correlazione non significa causalità"

- La regressione richiede di stabilire:
  - ★ una variabile "causa" che si chiama variabile di ingresso o indipendente (vel; stat; p; ?) 'x'
  - ★ una variabile "effetto" che si chiama variabile di risposta o dipendente (qtà; peso; V; ?) 'Y'

→ nell'es. 3 si potrebbero anche scambiare le due var tenendo conto che l'elettronica "vede" V e deve indovinare p

→ nell'es. 4 non è chiaro.

- Non è una scelta innocente; nonostante questo a volte deve essere fatta con un certo arbitrio. Si può comunque sempre scegliere di condurre entrambi gli studi

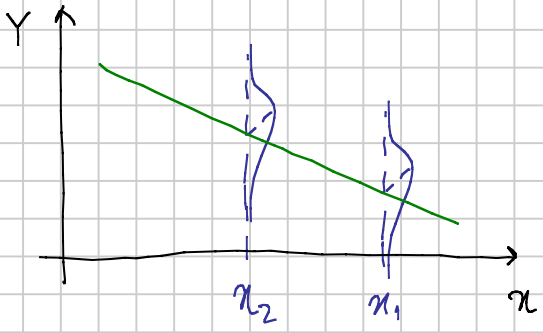


✱ Si sceglie quindi quale variabile è "deterministica" e quale è "aleatoria" perchè contiene l'incertezza o l'errore

→ Spesso questo punto di vista è irrealistico, ma si tratta di una semplificazione matematica di compromesso

ora 4

REGRESSIONE LINEARE SEMPLICE



legame lineare : "$Y = \alpha + \beta x$"

considerando anche l'errore :

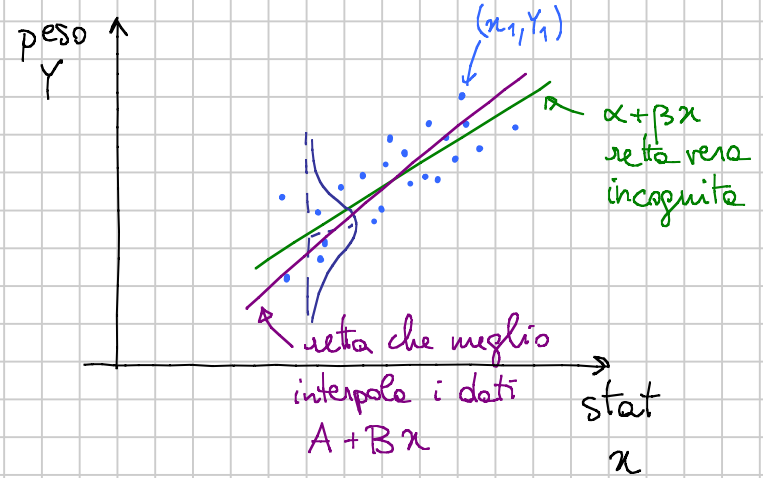
$$Y = \alpha + \beta x + e$$

dove $e \sim N(0; \sigma^2)$

✱ σ la dev. standard di e si suppone incognito ma non dipendente da x (= modello omoschedastico)

● Lo scopo della regr in questo caso è trovare α, β e σ!
Questo avviene a partire da un po' di dati sperimentali
(2 var per n esperimenti)

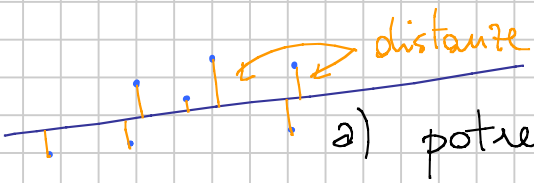
id	stat	peso
1	170	82
2	173	64
3	168	71
⋮	181	75
⋮	⋮	⋮
⋮	⋮	⋮
n	154	55



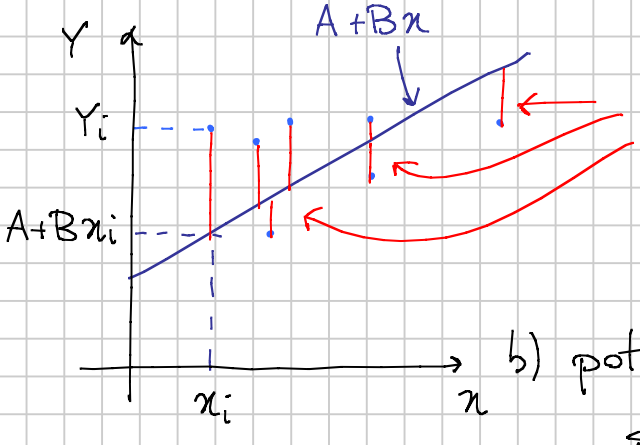
Dati i punti $(x_1, Y_1) \dots (x_n, Y_n)$ cerchiamo la retta $A + Bx$ che meglio li interpola, ovvero più "vicina" ai punti

- Data una retta si misura il suo "scarto" dai punti e poi si cerca tra tutte le rette quella con "scarto" minimo.

→ Come misuro lo "scarto"?



a) potrei sommare tutte le distanze — troppo complicato



$$R_i = Y_i - (A + B x_i)$$

b) potrei provare a sommare i residui

$$\text{scarto(retta)} = \sum_i R_i$$ — sbagliato

c) metto il valore assoluto

$$\text{scarto(retta)} = \sum_i |R_i|$$ — troppo complicato

d) faccio i quadrati

$$SS_R = SS = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} (Y_i - A - B x_i)^2$$

- Cerchiamo i coefficienti A, B a cui corrisponde il minimo di $SS = SS(A,B)$. Come si trovano A e B che rendono minima SS?
  - ↳ con il "risolutore" di Excel (solver)
  - ↳ facendo le derivate

$$\begin{cases} 0 = \frac{\partial SS}{A} = \frac{\partial}{\partial A} \sum R_i^2 = \sum \frac{\partial}{\partial A} R_i^2 = 2 \sum R_i \frac{\partial}{\partial A} R_i = -2 \sum R_i \\ 0 = \frac{\partial SS}{B} = \frac{\partial}{\partial B} \sum R_i^2 = \sum \frac{\partial}{\partial B} R_i^2 = 2 \sum R_i \frac{\partial}{\partial B} R_i = -2 \sum R_i x_i \end{cases}$$

equazioni normali $\begin{cases} \boxed{\sum R_i = 0} \\ \sum x_i R_i = 0 \end{cases}$ $\begin{cases} \sum_{i=1}^n (Y_i - A - Bx_i) = 0 \\ \sum_i x_i (Y_i - A - Bx_i) = 0 \end{cases}$ sist. lineare in A, B

$$\begin{cases} \sum Y_i = nA + B\sum x_i \\ \sum x_i Y_i = A\sum x_i + B\sum x_i^2 \end{cases}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i \qquad \overline{xY} = \frac{1}{n}\sum_{i=1}^n x_i Y_i \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$$

$$\begin{cases} \boxed{\bar{Y} = A + B\bar{x}} \\ \overline{xY} = A\bar{x} + B\overline{x^2} \end{cases} \qquad \begin{cases} A = \bar{Y} - B\bar{x} \\ \overline{xY} = (\bar{Y} - B\bar{x})\bar{x} + B\overline{x^2} \end{cases}$$

$$\boxed{\begin{cases} A = \bar{Y} - B\bar{x} \\ B = \dfrac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} \end{cases}}$$

coeff della retta che interpola meglio

★ $\bar{Y} = A + B\bar{x}$ significa che il punto $(\bar{x}, \bar{Y})$, che è il baricentro di tutti i punti, sta sulla retta che interpola meglio. Ovvero $A + Bx$ passa per il baricentro

→ Inoltre $\sum_i R_i = 0$ quindi residui positivi e negativi si bilanciano =

HW: Riscrivere la formula di B in funzione della varianza campionaria delle x_i, S_x^2 e della covarianza campionaria di x e Y (Capitolo 2 del Ross)

ora 5

- Si trova che A, B sono stimatori corretti e consistenti di α e β, con distribuzione normale di parametri "accessibili"
- Hanno distribuzione normale perché sono combinazioni lineari di vaa normali indipendenti:

$$A = \sum_{i=1}^n d_i Y_i \qquad B = \sum_{i=1}^n c_i Y_i \qquad c_i, d_i \in \mathbb{R}$$

- Nota bene: $Y_i = \alpha + \beta x_i + e_i$ con $e_i \sim \mathcal{N}(0, \sigma^2)$

  allora: $\boxed{Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)}$

  Siccome gli errori e_i si possono supporre indipendenti, anche le v.a. Y_i si possono supporre indipendenti

- Torniamo a $B = \sum_{i=1}^n c_i Y_i$ e giustifichiamolo

$$B = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} = \frac{\sum x_i Y_i - \bar{x}\sum Y_i}{n(\overline{x^2} - \bar{x}^2)} = \sum_{i=1}^n \underbrace{\frac{x_i - \bar{x}}{n(\overline{x^2} - \bar{x}^2)}}_{c_i} Y_i =: \sum_{i=1}^n c_i Y_i$$

Vale anche per A:

$$A = \bar{Y} - B\bar{x} = \frac{1}{n}\sum_i Y_i - \bar{x}\sum_i c_i Y_i = \sum_{i=1}^n \underbrace{\left(\frac{1}{n} - c_i \bar{x}\right)}_{d_i} Y_i =: \sum_{i=1}^n d_i Y_i$$

★ I coefficienti c_i e d_i sono numeri deterministici che dipendono solo dalle x_i

- Sono stimatori corretti: $E(A) = \alpha \quad E(B) = \beta$

$$E(B) = E\left(\sum_i c_i Y_i\right) \overset{\text{linearità di } E}{=} \sum c_i E(Y_i) = \sum_i c_i(\alpha + \beta x_i) = \alpha \underbrace{\sum_i c_i}_{0} + \beta \underbrace{\sum_i c_i x_i}_{1}$$

($c_i = \frac{x_i - \bar{x}}{n(\overline{x^2} - \bar{x}^2)}$)

$$= \alpha \cdot 0 + \beta \cdot 1 = \beta$$

infatti $$\sum_i c_i = \frac{\sum (x_i - \bar{x})}{n(\overline{x^2} - \bar{x}^2)} = \frac{n\bar{x} - n\bar{x}}{n(\overline{x^2} - \bar{x}^2)} = 0$$

$$\sum_i c_i x_i = \frac{\sum (x_i - \bar{x}) x_i}{n(\overline{x^2} - \bar{x}^2)} = \frac{\sum x_i^2 - \bar{x}\sum x_i}{n(\overline{x^2} - \bar{x}^2)} = \frac{\overline{x^2} - \bar{x}^2}{\overline{x^2} - \bar{x}^2} = 1$$

$$E(A) = E(\bar{Y} - B\bar{x}) = E(\bar{Y}) - \bar{x}E(B) = \frac{1}{n}\sum E(Y_i) - \beta\bar{x}$$

$$= \frac{1}{n}\sum (\alpha + \beta x_i) - \beta\bar{x} = \frac{1}{n}\sum_i \alpha + \beta \frac{1}{n}\sum_i x_i - \beta\bar{x} = \alpha$$

★ Punto della situazione : $A \sim \mathcal{N}(\alpha\,;\,?)$ $B \sim \mathcal{N}(\beta\,;\,?)$

- Troviamo le varianze

$$\mathrm{Var}(B) = \mathrm{Var}\left(\sum_i c_i Y_i\right) = \mathrm{Cov}\left(\sum_i c_i Y_i\,;\,\sum_j c_j Y_j\right) = \sum_i \sum_j c_i c_j \mathrm{Cov}(Y_i\,;Y_j)$$

$$= \sum_{i=1}^{n} c_i^2 \mathrm{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2$$

↑ $\mathrm{Cov}(Y_i\,;Y_j) = 0$ se $i \neq j$

$$\sum_i c_i^2 = \sum_i \frac{(x_i - \bar{x})^2}{n^2(\overline{x^2} - \bar{x}^2)^2} = \sum_i \frac{x_i^2 - 2\bar{x}x_i + \bar{x}^2}{n^2(\overline{x^2} - \bar{x}^2)^2} = \frac{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2}{n^2(\overline{x^2} - \bar{x}^2)^2}$$

$$= \frac{\overline{x^2} - 2\bar{x}^2 + \bar{x}^2}{n(\overline{x^2} - \bar{x}^2)^2} = \frac{\overline{x^2} - \bar{x}^2}{n(\overline{x^2} - \bar{x}^2)^2} = \frac{1}{n(\overline{x^2} - \bar{x}^2)} =: k_B$$

$$\mathrm{Var}(B) = \sigma^2 k_B \qquad k_B = \sum_i c_i^2 = \frac{1}{n(\overline{x^2} - \bar{x}^2)}$$

analogamente

$$\mathrm{Var}(A) = \sigma^2 k_A \qquad k_A = \sum_i d_i^2 = \frac{\overline{x^2}}{n(\overline{x^2} - \bar{x}^2)}$$

- Ho la distribuzione di A e B :

$$\boxed{A \sim \mathcal{N}(\alpha, \sigma^2 k_A)} \qquad \boxed{B \sim \mathcal{N}(\beta, \sigma^2 k_B)}$$

★ NON sono indipendenti : $\mathrm{Cov}(A\,;B)$ si può calcolare

- Manca solo uno stimatore per σ

$$SS = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

... se invece ci fossero α e β : $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$

$$\sum_{i=1}^{n} \left(\frac{Y_i - (\alpha + \beta x_i)}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\mathcal{N}(0,1)\right)^2 \sim \chi^2(n)$$ per definizione

$$\frac{SS}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{Y_i - A - Bx_i}{\sigma} \right)^2 \sim \chi^2(n-2)$$

↑ thm che faccio la prox volta

↑ ho sostituito 2 parametri con i relativi stimatori

* inoltre SS indip da A e SS indip da B

$$\boxed{\frac{SS}{\sigma^2} \sim \chi^2(n-2)}$$

- $\frac{SS}{n-2}$ è uno stimatore corretto di σ^2, indipendente da A e B.

Infatti $E\left(\frac{SS}{n-2}\right) = \frac{1}{n-2} E(SS) = \frac{\sigma^2}{n-2} E\left(\frac{SS}{\sigma^2}\right) = \frac{\sigma^2}{n-2} E\left(\chi^2(n-2)\right)$

$= \frac{\sigma^2}{n-2} \cdot (n-2) = \sigma^2$

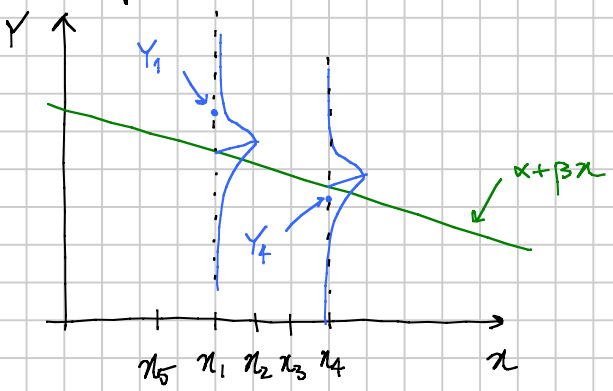
MMSD

ora 6

Titolo nota

13/10/2009

Regressione lineare semplice



$$Y_i \sim \mathcal{N}(\alpha+\beta x_i, \sigma^2)$$

Scopo: inferenza su α, β, σ

1) A, B stimatori per α, β

$$A \sim \mathcal{N}(\alpha, \sigma^2 k_A) \qquad B \sim \mathcal{N}(\beta, \sigma^2 k_B)$$

2) $\frac{SS}{n-2}$ stimatore per σ^2

$$\frac{SS}{\sigma^2} \sim \chi^2(n-2)$$

Vediamo ora le condizioni generali per dedurre che $\frac{SS}{\sigma^2} \sim \chi^2(n-2)$ e enunciati analoghi

→ Prima altri 2 esempi (e ripasso)

a) $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ μ e σ incognite

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mu \qquad \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$$

per definizione

$\mathcal{N}(0,1)$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\frac{S^2}{\sigma^2}(n-1) = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

b) due campioni X_i e Y_j $\quad X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ $\quad Y_j \sim \mathcal{N}(\mu_Y, \sigma^2)$

$$\bar{X}-\bar{Y} \sim \mathcal{N}\left(\mu_X-\mu_Y ; \frac{\sigma^2}{m}+\frac{\sigma^2}{n}\right)$$

m, n numerosità campioni

$$S_p^2 = \frac{m+1}{m+n-2}S_X^2 + \frac{n+1}{m+n-2}S_Y^2$$

$$\sum_{i=1}^{m}\left(\frac{X_i-\mu_X}{\sigma}\right)^2+\sum_{j=1}^{n}\left(\frac{Y_j-\mu_Y}{\sigma}\right)^2 \sim \chi^2(m+n) \qquad \text{per def}$$

$$\sum_{i=1}^{m}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2+\sum_{j=1}^{n}\left(\frac{Y_j-\bar{Y}}{\sigma}\right)^2 \sim \chi^2(m+n-2) \qquad \text{per thm}$$

$$\frac{S_p^2}{\sigma^2}(m+n-2) \quad \text{HW: verificare}$$

c) $x_1, \dots, x_n$ numeri $\quad Y_i \sim \mathcal{N}(\alpha+\beta x_i ; \sigma^2)$ indipendenti

$$\sum_{i=1}^{n}\left(\frac{Y_i-(\alpha+\beta x_i)}{\sigma}\right)^2 \sim \chi^2(n) \qquad \text{per def}$$

$$SS := \sum_{i=1}^{n}\left(Y_i - A - B x_i\right)^2$$

$$\frac{SS}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i-(A+Bx_i)}{\sigma}\right)^2 \sim \chi^2(n-2) \qquad \text{per thm}$$

IN GENERALE

hp: $Y_1, \dots, Y_n \quad Y_i \sim \mathcal{N}(\mu_i ; \sigma^2)$ indipendenti

$B_1, B_2, \dots, B_k$ ($k=2$ in b) e c)) combinazioni lineari a coeff noti delle Y_i:

$$B_1 = c_{11}Y_1 + c_{12}Y_2 + c_{13}Y_3 + \dots + c_{1n}Y_n$$
$$B_2 = c_{21}Y_1 + \dots$$
$$B_j = \sum_{i=1}^{n} c_{ji} Y_i$$

$v^1, v^2, \dots, v^k$ vettori di $\mathbb{R}^n$ linearmente indipendenti

$$v^j = (v_1^j, v_2^j, \dots, v_n^j) \in \mathbb{R}^n$$

$$E\left[Y_i - \left(v_i^1 B_1 + v_i^2 B_2 + \dots + v_i^k B_k\right)\right] = 0$$

ts: $$W := \sum_{i=1}^{n} \left(\frac{Y_i - (v_i^1 B_1 + v_i^2 B_2 + \dots + v_i^k B_k)}{\sigma} \right)^2 \sim \chi^2(n-k)$$

inoltre W è indipendente da $(B_1, B_2, \dots, B_k)$

- c) usa il thm:

$$W = \sum_{i=1}^{n} \left(\frac{Y_i - (A + Bx_i)}{\sigma} \right)^2$$

$k = 2 \quad B_1 = A, \quad B_2 = B$ ok

$v_i^1 = 1 \quad i = 1, 2, \dots, n \qquad v^1 = (1, 1, \dots, 1)$

$v_i^2 = x_i \quad i = 1, 2, \dots, n \qquad v^2 = (x_1, x_2, \dots, x_n)$

v^1 e v^2 sono lin. dipend. sse $v^2 = \lambda v^1$ ovvero se $v^2 = (\lambda, \lambda, \dots, \lambda)$ ovvero se $x_i \equiv \lambda \ \forall i$

→ Basta che gli x_i non siano tutti uguali ok

$B_1 = A = \sum_i d_i Y_i \qquad B_2 = B = \sum_i c_i Y_i$ sono comb. lin.

$$E(Y_i - A - Bx_i) = E(Y_i) - E(A) - x_i E(B)$$
$$= \alpha + \beta x_i - \alpha - x_i \beta = 0 \quad \text{ok}$$

Tesi: $\frac{SS}{\sigma^2} \sim \chi^2(n-2)$ inoltre SS è indipendente da (A, B)

HW: a) e b) usano il thm

ora 7

ERRORE STANDARD

$$\frac{SS}{\sigma^2} \sim \chi^2(n-2) \Rightarrow E\left(\frac{SS}{\sigma^2}\right) = n-2 \qquad \text{Var}\left(\frac{SS}{\sigma^2}\right) = 2n-4$$

$$\Rightarrow E(SS) = \sigma^2(n-2) \Rightarrow \boxed{E\left(\frac{SS}{n-2}\right) = \sigma^2}$$

$$\text{Var}\left(\frac{SS}{n-2}\right) = \text{Var}\left(\frac{SS}{\sigma^2} \cdot \frac{\sigma^2}{n-2}\right) = \left(\frac{\sigma^2}{n-2}\right)^2 \text{Var}\left(\frac{SS}{\sigma^2}\right) = \frac{2\sigma^4}{n-2}$$

tende a 0 per $n \to \infty$

- $\frac{SS}{n-2}$ è uno stimatore corretto e consistente di σ^2

HW: verificare se A e B sono stimatori consistenti di α, β.

- $S_e := \sqrt{\frac{SS}{n-2}}$ è uno stimatore (non corretto) di σ.
  Prende il nome di errore standard

HW: capire se $E(S_e) \gtrless \sigma$

☆ S_e è indipendente da (A, B)

FUNZIONI ANCILLARI

→ funzione che dipende dai dati, dai parametri noti e da un solo parametro incognito (quello su cui si vuole fare inferenza) di cui sia nota la distribuzione

→ esempi: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ se σ è nota

$\frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$ $\qquad$ $2n\lambda\bar{X} \sim \chi^2(2n)$

$$\boxed{\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)}$$

grazie al fatto che $\bar{X}$ e S sono indipendenti

- Veniamo ad A, B:

$A \sim \mathcal{N}(\alpha\,;\, \sigma^2 k_A)$ $\qquad$ $k_A = \frac{1}{n} \cdot \frac{\overline{x^2}}{\overline{x^2} - \bar{x}^2}$

$\frac{A-\alpha}{\sigma\sqrt{k_A}} \sim \mathcal{N}(0,1)$ $\qquad$ non è ancillare

Tuttavia, visto che S_e è stimatore di σ indipendente da A e visto che $\frac{S_e^2}{\sigma^2}(n-2) = \frac{SS}{\sigma^2} \sim \chi^2(n-2)$, si ha che:

$$\boxed{\frac{A-\alpha}{S_e\sqrt{k_A}} \sim t(n-2)}$$

funz ancillare

Analogamente : $B \sim \mathcal{N}(\beta, \sigma^2 k_B)$ $\quad k_B = \frac{1}{n} \cdot \frac{1}{\overline{x^2} - \overline{x}^2}$

$$\frac{B-\beta}{\sigma\sqrt{k_B}} \sim \mathcal{N}(0,1)$$

$$\boxed{\frac{B-\beta}{Se\sqrt{k_B}} \sim t(n-2)}$$ funz ancillare

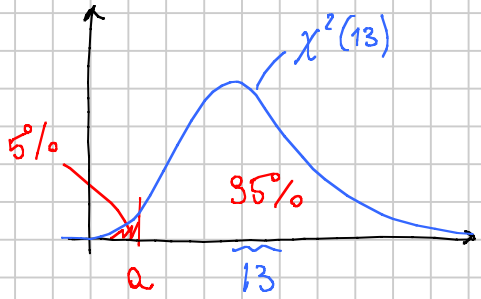
Infine $$\boxed{\frac{SS}{\sigma^2} = \frac{Se^2}{\sigma^2}(n-2) \sim \chi^2(n-2)}$$ funz ancillare

INFERENZA SULLA REGR. LIN. SEMPLICE

Siamo in grado di fare:

a) Intervalli di confidenza su α, β, σ

→ ad es.: int. conf. unil. sx su σ con lvl di conf 95%



$n = 15$, $Se = 0{,}7$

$a = INV.CHI(95\%; 13) \cong 5{,}89$

$$95\% = P(\chi^2(13) > a) = P\left(\frac{SS}{\sigma^2} > a\right)$$

$$= P(SS > a\sigma^2) = P\left(\frac{SS}{a} > \sigma^2\right) = P\left(\sigma < \sqrt{\frac{SS}{a}}\right)$$

Con il 95% di conf $\sigma < Se\sqrt{\frac{n-2}{a}} \cong 1{,}04$

b) Test statistici su α, β, σ

→ ad es.: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$ (p dei dati) $n=15$

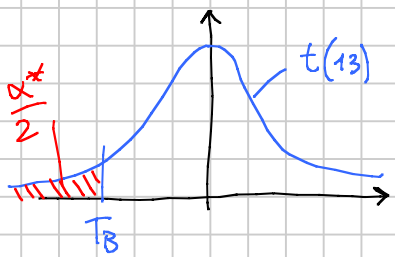
la statistica del test è

$$\frac{B-\beta}{Se\sqrt{k_B}} \sim t(n-2) \Rightarrow \frac{B}{Se\sqrt{k_B}} =: T_B \overset{H_0}{\sim} t(n-2) \quad (\beta_0 = 0)$$

Calcolo $T_B = -2{,}314$

$$\alpha^* = 2P(t(13) \leq -2{,}314)$$

$$\alpha^* = 2 - 2P(t(n-2) \leq |T_B|)$$



$$\alpha^* = 2 * \text{DISTRIB.T}(-2{,}314\ ;\ 13) = \text{NO}$$
$$= \text{DISTRIB.T}(2{,}314\ ;\ 13\ ;\ 2) \cong 3{,}8\%$$

ora 8

- Il test più importante: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

  Q: La var Y dipende dalla var x? (H_1 = sì ; H_0 = no)

  $$\alpha^* = \text{DISTRIB.T}(\text{ABS}(T_B)\ ;\ n-2\ ;\ 2)$$

  ★ Quando facciamo la regressione con gli strumenti automatici di Excel, questo valore è calcolato automaticamente

  ★ H_1 → tutto ok, la regressione è utile

  H_0 → vuol dire che $Y_i \sim N(\mu^{(\alpha)}, \sigma^2)$ è un campione omogeneo che si può trattare con tecniche elementari

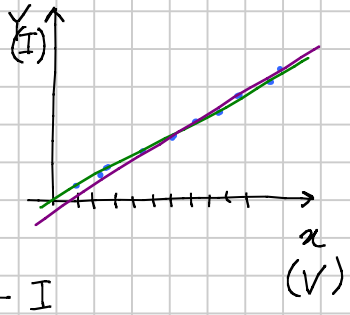
- Analogo su α $\quad H_0: \alpha = 0$ vs $H_1: \alpha \neq 0$

  Q: c'è termine noto nella relazione (H_1 = sì ; H_0 = forse no)

a) Ci sono grandezze che hanno una rel lineare senza termine noto

→ testo un resistore $\quad V = IR$

$$I = \frac{1}{R} V$$



Posso stimare $\frac{1}{R}$ (e quindi R) usando la regressione $\quad x \leftarrow V \quad Y \leftarrow I$

$$Y = \alpha + \beta x$$

$$I = \alpha + \beta V = 0 + \frac{1}{R} V \quad \text{perciò} \quad \alpha = 0 \quad \beta = \frac{1}{R}$$

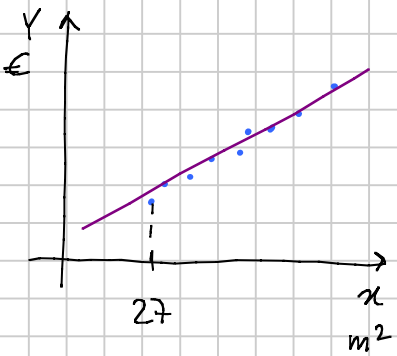
B stima $\frac{1}{R}$ $\qquad$ B^{-1} stima R

★ In questo caso sarebbe più sensato fare la regressione con un modello ancora più semplice: $\boxed{Y = \beta x}$

b) Ci sono casi dubbi: non so se $\alpha = 0$ o $\alpha \neq 0$

ad esempio: appartamenti → superficie $[m^2]$; prezzo $[€]$

x Y



$Y = \alpha + \beta x$

in questo caso: può avere senso che $\alpha = 0$?

se no, non faccio il test e tengo il termine noto anche se pdd è molto grande

se sì, faccio il test:

se H_1 → tutto ok, continuo così

se H_0 → devo rifare la regressione con $Y = \beta x$

HW: Costruire da zero fino all'inferenza la regressione senza α:

$Y = \beta x + e$ $\quad e \sim N(0, \sigma^2)$

- Inferenza sulla risposta media e sulle risposte future

→ esempio: regressione peso (Y) vs statura (x)

Q: Quanto dovrebbe pesare una persona alta 180 cm?

poniamo $x_0 = 180$; facciamo la regressione $Y = \alpha + \beta x$ sui dati a disposizione; troviamo A, B, S_e

i. $\alpha + \beta x_0$ (incognito) è il peso medio che hanno le persone alte x_0

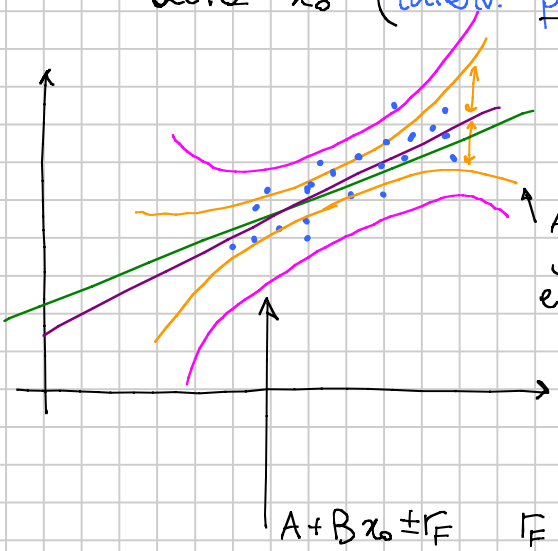
ii. $N(\alpha + \beta x_0 ; \sigma^2)$ (incognita) è la distribuzione di tutti i pesi delle persone alte x_0

iii. $A + B x_0$ stima puntuale del peso medio

→ iv. $A + B x_0 \pm r_M$ (r_M piccolo se n grande) intervallo di confidenza per $\alpha + \beta x_0$ (interv. conf. risposta media)

→ v. $A + B x_0 \pm r_F$ (r_F sempre maggiore di S_e) intervallo che con elevata probabilità conterrà il peso di una persona

alta x_0 (interv. predizione risposta futura singola)



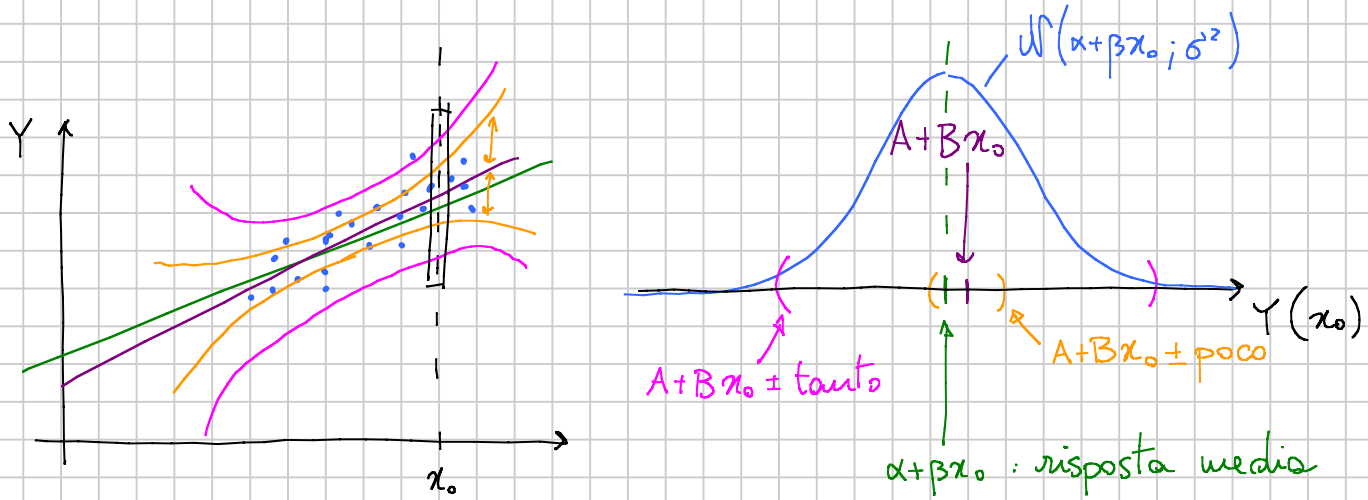
$A+Bx_0 \pm r_M$ $\quad r_M = r_M(x_0)$

è stretto attorno a $A+Bx$ (viola) e contiene $\alpha+\beta x$ (verde) con elevata confidenza

$A+Bx_0 \pm r_F$ $\quad r_F = r_F(x_0)$

è centrato attorno a $A+Bx$ (viola) e contiene i singoli punti (vecchi e futuri) con elevata probabilità

INFERENZA PER RISPOSTA MEDIA E FUTURA



$$Y \sim \mathcal{N}(\alpha+\beta x_0 ; \sigma^2)$$

$A+Bx_0 \pm a\, S_e \sqrt{k_m}$: (arancione) intervallo di confidenza per la risposta media — può essere molto stretto, soprattutto se n è grande

$A+Bx_0 \pm a\, S_e \sqrt{1+k_m}$: (fucsia) intervallo di predizione : una singola risposta futura Y relativa ad un livello di ingresso x_0 apparterrà a questo intervallo con probabilità assegnata

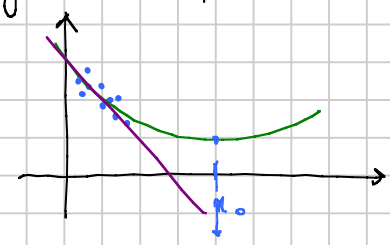
- $k_m = k_m(x_0) = \frac{1}{n}\left[1 + \frac{(x_0-\bar{x})^2}{\overline{x^2}-\bar{x}^2}\right]$

  → è minima quando $x_0 = \bar{x}$ e si allarga verso i bordi

  → non è mai minore di $\frac{1}{n}$

  → tende a zero per $n \to \infty$

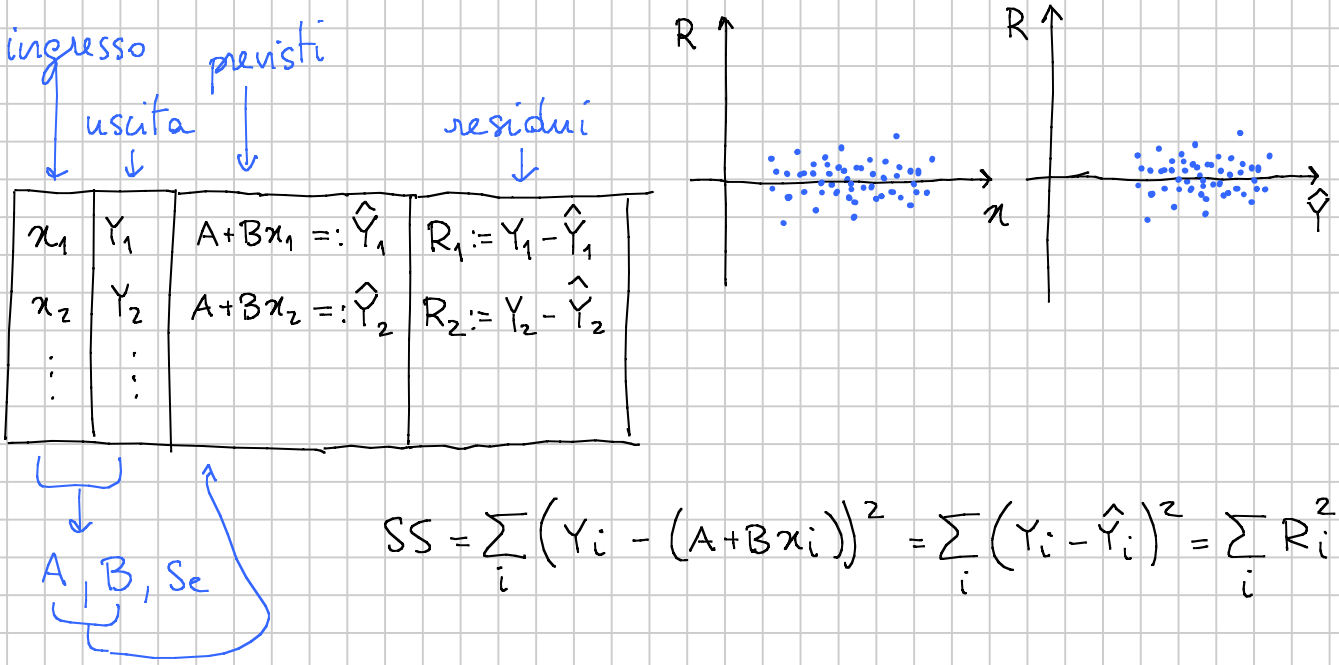
☆ In generale l'affidabilità dell'inferenza a ingresso x_0 non è garantita per valori esterni a quelli dei dati o vicini ai bordi.



- a deve essere un quantile t di Student con $n-2$ g.d.l.
(Si può prendere bilaterale o unilaterale...)

ANALISI DEI RESIDUI

Il modo migliore per capire se la regressione è affidabile è analizzare qualitativamente il diagramma di dispersione dei residui vs la var di ingresso o il valore previsto



$$SS = \sum_i \left(Y_i - (A+Bx_i)\right)^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2 = \sum_i R_i^2$$

- Intendiamo che la regressione è affidabile se davvero:

$$Y \sim \mathcal{N}\left(\alpha + \beta x ; \sigma^2\right)$$

la varianza non dipende da x (modello omoschedastico)

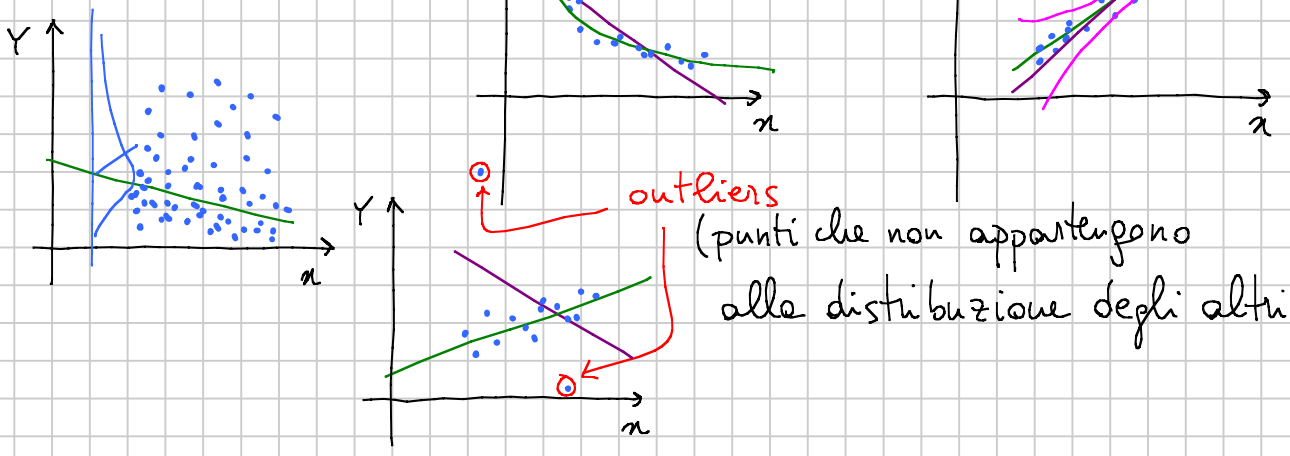
e se invece $\sigma^2 = \sigma^2(x)$?

funzione lineare di x

e se non è lineare ?

legge normale

e se non lo è ?



(punti che non appartengono alla distribuzione degli altri)

Informazioni di servizio:

- gio 29 ~~Nicolodi~~ → Morandin
- gio 22 Nicolodi normalmente
- file .xlsx → è un file .zip ... click col DX → apri con... Excel
  Office XP - 2003 → google → open file office 2007

Bizzarrie da evitare

$R_1, R_2, \dots, R_n$ sono un campione "quasi" gaussiano

$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ media campionaria — è sempre 0

(x_i, R_i) → faccio la regressione — viene sempre $A_R = 0 = B_R$

$$S_R^2 := \frac{1}{n-1}\sum_i \left(R_i - \underbrace{\overline{R}}_{0}\right)^2 = \frac{1}{n-1}\sum_i R_i^2 = \frac{SS}{n-1}$$

è uno stimatore di σ^2 che sottostima

$$\sigma^2 \simeq \frac{SS}{n-2}$$

Coefficiente di Determinazione

$$R^2 = 1 - \frac{SS}{S_{YY}}$$

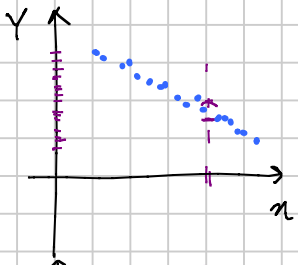
dove $SS = \sum_{i=1}^{n} \left(Y_i - (A + B x_i)\right)^2$ imprecisione sulla conoscenza di Y_i, conoscendo le x_i oppure no

$$S_{YY} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

$\frac{SS}{S_{YY}}$ è la frazione di imprecisione che non viene rimossa anche se si conoscono le x_i

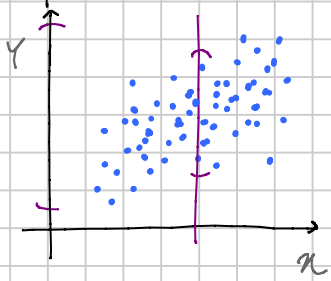
$R^2 = 1 - \frac{SS}{S_{YY}}$ è la fraz. di imprecisione rimossa dalla conoscenza delle x_i, o anche "la frazione di varianza spiegata dalla regressione"

★ Un elevato valore di R^2 significa che la regressione è buona e utile ovvero che le nostre previsioni e stime su Y sono molto migliori quando conosciamo x



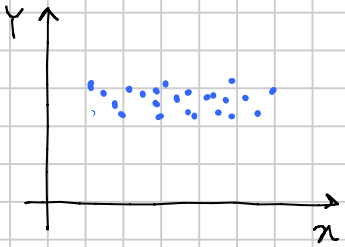
$R^2 \approx 97\%$ grande

p dei dati $H_0: \beta = 0$ 10^{-40} piccolo



$R^2 \approx 6\%$ piccolo

p dei dati $H_0: \beta = 0$ 10^{-20} piccolo



$R^2 \approx 3\%$ piccolo

p dei dati $H_0: \beta = 0$ 0,72 grande

★ $R^2 \in [0,1]$

★ Nel caso della regressione lineare semplice $\boxed{R^2 = r^2}$ dove r è il coefficiente di correlazione lineare campionaria (Cap 2)

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} \Rightarrow r(x_i, Y_i) = \dots$$

→ Se due variabili hanno una correlazione di 0,4 ... non è gran che perché $0,4^2 = 0,16 = 16\%$

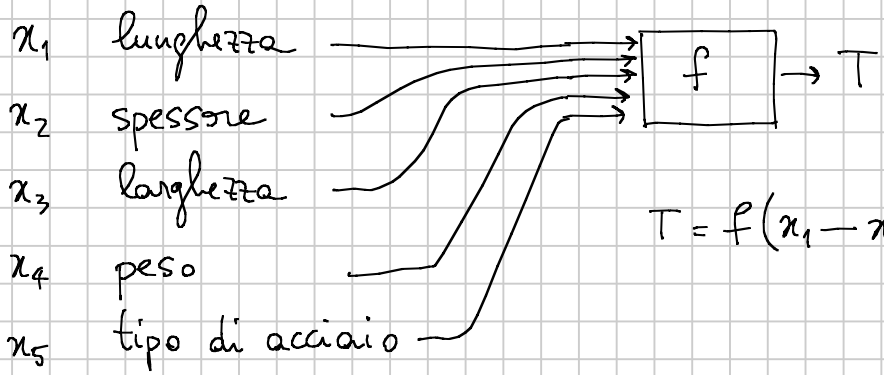
REGRESSIONE LINEARE MULTIPLA

→ Esempio: acciaieria → linea di lavorazione → programma

coil di acciaio

↓

l'ingresso viene occupato per tempo T non prevedibile facilmente



$T = f(x_1 - x_5)$??

★ La regr. lineare multipla permette di studiare le relazioni tra $p \geq 1$ variabili di ingresso e una variabile di uscita

ora 11

Modello: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + e$

Dati:

id	Y	x_1	x_2	....	x_p
1	$Y(1)$	$x_1(1)$	$x_2(1)$		$x_p(1)$
2	$Y(2)$	$x_1(2)$			
⋮	⋮	⋮	⋮		⋮
n	$Y(n)$	$x_1(n)$	....		$x_p(n)$

$Y(i)$ risposta del punto i

$x_j(i)$ var di ingresso j-esima del punto i

$$Y(i) \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(i) \; ; \; \sigma^2\right)$$

modello lineare, gaussiano, omoschedastico

alta notazione:

- $Y(i) = Y_i$
- $x_j(i) = x_{ij}$ (come matrice)

★ Molto spesso si inventa ed inserisce $x_0 \equiv 1$ in modo che sia:

$$Y \sim \mathcal{N}\left(\sum_{j=0}^{p} \beta_j x_j \; ; \; \sigma^2\right)$$

Sia X la matrice x_{ij} $i = 1, 2, \dots, n$ $j = 0, 1, \dots, n$

$$\begin{pmatrix} 1 & x_1(1) & \dots & x_p(1) \\ 1 & x_1(2) & & \vdots \\ 1 & & \ddots & \vdots \\ \vdots & & & \vdots \\ 1 & \dots & & x_p(n) \end{pmatrix} =: X \in M_{n,p+1} \qquad \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} =: Y \in \mathbb{R}^n$$

$\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{pmatrix} =: \beta \in \mathbb{R}^{p+1}$ ‖ $\begin{pmatrix}B_0\\ B_1\\ \vdots\\ B_p\end{pmatrix} =: B \in \mathbb{R}^{p+1}$ ‖ $B_i \simeq \beta_i$ stimatori

- Previsti : $\hat{Y}(i) = \sum_{j=0}^{p} B_j x_j(i) = \sum B_j x_{ij} = \sum_j x_{ij} B_j = [XB]_i$
- Residui : $R(i) = Y(i) - \hat{Y}(i) = Y_i - [XB]_i = [Y - XB]_i$
- SS : $SS = \sum_i R(i)^2 = \|Y - XB\|^2$
- Cerchiamo i coefficienti B_i che minimizzano SS

$$0 = \frac{\partial SS}{\partial B_j} = \frac{\partial}{\partial B_j} \sum_i \left(Y(i) - \sum_{k=0}^{p} B_k x_k(i) \right)^2 = 2 \sum_i \left(Y(i) - \sum_0^p B_k x_k(i) \right)\left(-x_j(i)\right)$$

$$0 = \sum_i \left(Y(i) x_j(i) - \sum_{k=0}^{p} B_k x_k(i) x_j(i) \right) = \sum_i Y(i) x_j(i) - \sum_i x_j(i) [XB]_i$$

$$= \sum_i Y_i x_{ij} - \sum_i [XB]_i x_{ij} = [Y^T X]_j - [(XB)^T X]_j = [Y^T X - B^T X^T X]_j \quad \forall j$$

quindi $Y^T X - B^T X^T X = 0$ trasposto : $X^T Y - X^T X B = 0$

- Equazioni normali : $\boxed{X^T Y = X^T X B}$
- Ricavo B : $(X^T X)^{-1} X^T Y = B$

$$\boxed{B = (X^T X)^{-1} X^T Y}$$

- Funzioni da usare su Excel

=MATR.PRODOTTO (matrice 1 ; matrice 2) =MMULT(..

=MATR.TRASPOSTA (matrice) =TRANSPOSE(..

=MATR.INVERSA (matrice) =MINVERSE(...

=MATR.SOMMAPRODOTTO (vett 1 ; vett 2) prodotto scalare

=SUMPRODUCT(...

- Distribuzione coefficienti

1) $B_j = \left[(X^TX)^{-1}X^TY\right]_j = \sum_i \left[(X^TX)^{-1}X^T\right]_{ji} Y_i$ è comb. lineare delle Y_i

($(X^TX)^{-1}X^T$: $p+1 \times n$)

Quindi B_j ha distribuzione normale

2) $\boxed{E(B_j) = \ldots = \beta_j}$ stimatore corretto

3) $Var(B_j) = \ldots$

è più facile e più utile : $Cov(B_j ; B_k) =: \Sigma_{jk}$

Σ 'Sigma' matrice di covarianza $\in M_{p+1,p+1}$

$$\boxed{\Sigma = \ldots = \sigma^2 (X^TX)^{-1}}$$

$$B = (X^TX)^{-1}X^TY$$

$$E(B) = (X^TX)^{-1}X^T E(Y) = (X^TX)^{-1}X^TX\beta = \beta$$

$$Cov(B ; B) =$$

Regr. multilineare → matrici

var di ingresso

$$X = \begin{pmatrix} 1 & x_1(1) & x_2(1) & \cdots & x_p(1) \\ | & | & | & & | \\ 1 & x_1(n) & & & x_p(n) \end{pmatrix} \in M_{n,p+1}$$

$x_0(i) = 1 \quad \forall i = 1, 2, \dots, n$

var di risposta

$$Y = \begin{pmatrix} Y(1) \\ | \\ Y(n) \end{pmatrix} \in \mathbb{R}^n$$

$$\beta = \begin{pmatrix} \beta_0 \\ | \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1} \qquad B = \begin{pmatrix} B_0 \\ | \\ B_p \end{pmatrix} \in \mathbb{R}^{p+1} \qquad (X^T X)^{-1} \in M_{p+1,p+1}$$

$$\boxed{B = (X^T X)^{-1} X^T Y}$$

- Distribuzione di $B = (B_0, B_1, \dots, B_p)$ è un vettore → distrib. congiunta

Per fortuna non occorre scrivere $f_B(t_0, t_1, \dots, t_p) = ???$

Infatti B ha distribuzione gaussiana multidimensionale ...

... basta conoscere quali sono i parametri che la identificano:

→ la media di ciascuna componente $E(B_i)$, $i = 0, 1, \dots, p$

→ la varianza di ciascuna componente $\mathrm{Var}(B_i)$, $i = 0, \dots, p$

→ la covarianza di ogni coppia di comp. $\mathrm{Cov}(B_i, B_j)$ $i, j = 0, \dots, p$

★ Ovviamente $\mathrm{Var}(B_i) = \mathrm{Cov}(B_i; B_i)$ quindi la II è inclusa nella III

a) $B = NY \qquad N := (X^T X)^{-1} X^T \in M_{p+1,n}$

B è una trasf lineare di Y, quindi è normale.

b) $E(B) = E(NY) = N E(Y) = N X \beta = (X^T X)^{-1} X^T X \beta = \beta$ stimatore corretto

infatti:

$$Y(i) \sim \mathcal{N}\left(\beta_0+\beta_1 x_1(i)+\beta_2 x_2(i)+\dots+\beta_p x_p(i)\ ;\ \sigma^2\right) \sim \mathcal{N}\left([X\beta]_i\ ;\ \sigma^2\right)$$

$$E(Y(i)) = [X\beta]_i \quad \Rightarrow \quad E(Y) = X\beta$$

c) $\Sigma := \mathrm{Cov}(B) \in M_{p+1,p+1}$ matrice di covariante di B

$\Sigma_{ij} := \mathrm{Cov}(B_i\ ;\ B_j)$

$\mathrm{Cov}(B\ ;\ B) = \mathrm{Cov}(NY\ ;\ NY) = ?$

Più in generale: X, Y vettori aleatori con $S = \mathrm{Cov}(X\ ;\ Y)$ matrice di covariante; M, N matrici:

$$\left[\mathrm{Cov}(MX\ ;\ NY)\right]_{ij} := \mathrm{Cov}\left([MX]_i\ ;\ [NY]_j\right) = \mathrm{Cov}\left(\sum_k M_{ik}X_k\ ;\ \sum_\ell N_{j\ell}Y_\ell\right)$$

$$= \sum_k \sum_\ell M_{ik}N_{j\ell}\,\mathrm{Cov}(X_k\ ;\ Y_\ell) =: \sum_k \sum_\ell M_{ik}N_{j\ell}\,S_{k\ell} = \sum_{k,\ell} M_{ik}\cdot S_{k\ell}\cdot N^T_{\ell j} =$$

$$=: \left[MSN^T\right]_{ij}$$

$$\boxed{\mathrm{Cov}(MX\ ;\ NY) = M\,\mathrm{Cov}(X\ ;\ Y)\,N^T}$$

$$\mathrm{Cov}(B\ ;\ B) = \mathrm{Cov}(NY\ ;\ NY) = N\,\mathrm{Cov}(Y\ ;\ Y)\,N^T = N\sigma^2 I\,N^T = \sigma^2 NN^T$$

$$\mathrm{Cov}(Y_i\ ;\ Y_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

$$NN^T = (X^TX)^{-1}X^T\,X\,(X^TX)^{-1} = (X^TX)^{-1}$$

$$\boxed{\mathrm{Cov}(B\ ;\ B) = \sigma^2 (X^TX)^{-1}}$$

- Serve uno stimatore di σ^2

$$\boxed{S_e^2 := \frac{SS}{n-(p+1)} \approx \sigma^2}$$ stimatore consistente e corretto di σ^2

$$\rightarrow \quad \boxed{\frac{SS}{\sigma^2} \sim \chi^2(n-p-1)} \quad \text{indipendente da } B$$

HW: Verificate l'applicabilità del thm del chi-quadro

INFERENZA SUI β_k

$$B_k \sim \mathcal{N}\left(\beta_k ; \sigma^2 \cdot (X^TX)^{-1}_{k,k}\right)$$

$$\Rightarrow \frac{B_k - \beta_k}{\sigma\sqrt{(X^TX)^{-1}_{k,k}}} \sim \mathcal{N}(0,1)$$ se σ incognita...

$$\Rightarrow \boxed{\frac{B_k - \beta_k}{S_e\sqrt{(X^TX)^{-1}_{k,k}}} \sim t(n-p-1)}$$ funzione ancillare

$\rightarrow$ $H_0: \beta_k = 0$ $\quad H_1: \beta_k \neq 0$ test per vedere se $\beta_k = 0$

statistica $T_k := \dfrac{B_k}{S_e\sqrt{(X^TX)^{-1}_{k,k}}} \overset{H_0}{\sim} t(n-p-1)$

- Dopo la regressione preliminare, andrebbero eliminate le variabili per cui risulta plausibile $H_0: \beta_k = 0$.
  In effetti tali variabili non vanno mai rimosse tutte assieme: se ne toglie una e si rifà la regressione: può essere che a questo punto una o più delle altre risultino in $H_1: \beta_k \neq 0$.

ord 13

Inferenza risposta media e futura

$p = 1$ $\quad x_0 \rightarrow Y(x_0) \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$

$\alpha + \beta x_0$??

$p \geq 1$ $\quad \tilde{x} \in \mathbb{R}^{p+1}$, $\tilde{x} = (1, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_p) \rightarrow Y(\tilde{x}) \sim \mathcal{N}(\tilde{x}\cdot\beta ; \sigma^2)$

$\tilde{x}\cdot\beta = \beta_0 + \beta_1\tilde{x}_1 + \dots + \beta_p\tilde{x}_p$??

$\rightarrow$ Stimatore puntuale $\tilde{x}\cdot B$

$\rightarrow$ Che distribuzione ha? Certamente ha legge normale (scalare)

$\rightarrow$ Parametri: $E(\tilde{x}\cdot B) = \tilde{x}\cdot E(B) = \tilde{x}\cdot\beta$

$\tilde{x}\cdot B$ è uno stimatore corretto

$$\mathrm{Var}(\vec{x}\cdot B) = \mathrm{Cov}(\vec{x}\cdot B ; \vec{x}\cdot B) = \mathrm{Cov}(MB ; MB) = \circledast$$

dove $M = \vec{x}^T = (1, \vec{x}_1, \dots, \vec{x}_p) \in M_{1,p+1}$

$$\circledast = M\,\mathrm{Cov}(B;B)\,M^T = \vec{x}^T \Sigma \vec{x} = \vec{x}\cdot\Sigma\vec{x} = \langle \vec{x}, \Sigma\vec{x}\rangle$$

prod matrici — notazione

dove $\Sigma = \sigma^2 (X^T X)^{-1}$

$$\boxed{\mathrm{Var}(\vec{x}\cdot B) = \sigma^2 \vec{x}^T (X^T X)^{-1} \vec{x}}$$

→ funz. ancillare:

$$\frac{\vec{x}\cdot B - \vec{x}\cdot\beta}{\sigma\sqrt{\vec{x}^T(X^TX)^{-1}\vec{x}}} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \boxed{\frac{\vec{x}\cdot B - \vec{x}\cdot\beta}{S_e\sqrt{\vec{x}^T(X^TX)^{-1}\vec{x}}} \sim t(n-p-1)}$$

→ int. di conf.

quantile $t(n-p-1)$ con lvl di conf prescelto

$$\vec{x}\cdot\beta \in \vec{x}\cdot B \pm q\, S_e \sqrt{\vec{x}^T(X^TX)^{-1}\vec{x}}$$

- Risposta futura:

$\vec{x}$ livello di ingresso $\quad \vec{x} = (1, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_p)$

$Y(1), Y(2), \dots, Y(n)$ con ingressi $x(i)$ diversi

$\vec{x} \longrightarrow Y(n+1)$ risposta per $\vec{x}$

$Y(n+1) \sim \mathcal{N}(\vec{x}\cdot\beta ; \sigma^2)$ indipendente da Y e quindi da B

$$Y(n+1) - \vec{x}\cdot B \sim \mathcal{N}\left(0 ; \sigma^2 + \sigma^2 \vec{x}^T(X^TX)^{-1}\vec{x}\right)$$

$$\frac{Y(n+1) - \vec{x}\cdot B}{\sigma\sqrt{1+\vec{x}^T(X^TX)^{-1}\vec{x}}} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \boxed{\frac{Y(n+1) - \vec{x}\cdot B}{S_e\sqrt{1+\vec{x}^T(X^TX)^{-1}\vec{x}}} \sim t(n-p-1)}$$

(Non si tratta di una funz. ancillare)

$1-\alpha$ prob. fissata dell'intervallo di predizione

q quantile $t(n-p-1)$ bilaterale corrispondente a lvl conf $1-\alpha$

$$q = F^{-1}_{t(n-p-1)}\left(1-\frac{\alpha}{2}\right) = \text{INV.T}\left(\alpha;\, n-p-1;\, 2\right)$$

$$1-\alpha = P\left(-q \le t(n-p-1) \le q\right) =$$

$$= P\left(-q \le \frac{Y(n+1) - \vec{x}\cdot B}{S_e\sqrt{1+\vec{x}^T(X^TX)^{-1}\vec{x}}} \le q\right) = \ldots = P\left(Y(n+1) \in \vec{x}\cdot B \pm q\, S_e\sqrt{1+\vec{x}^T(X^TX)^{-1}\vec{x}}\right)$$

$$Y(n+1) \in \vec{x}\cdot B \pm q\, S_e\sqrt{1+\vec{x}^T(X^TX)^{-1}\vec{x}}$$ intervallo di predizione

$Y(n+1)$: v.a. futura; $+1$: differenza rispetto int. di conf. $\vec{x}\cdot\beta$

HW: trovare intervallo di predizione per la somma e per la media delle prossime m risposte future, tutte con lvl di impr. $\vec{x}$

$$\vec{x} \longrightarrow \begin{matrix} Y(n+1) \\ Y(n+2) \\ \vdots \\ Y(n+m) \end{matrix} \quad \begin{matrix} V := Y(n+1) + \ldots + Y(n+m) \\ \\ U := \frac{V}{m} \end{matrix}$$

COEFFICIENTE DI DETERMINAZIONE

$$R^2 := 1 - \frac{SS}{S_{YY}}$$ frazione di devianza spiegata

dove $S_{YY} = \sum_i Y_i^2 - n\bar{Y}^2 = (n-1)S_Y^2$

S_Y^2: varianza campionaria $\quad S_Y^2 = \frac{1}{n-1}\left(\sum_i Y_i^2 - n\bar{Y}^2\right) = \frac{1}{n-1}\sum_i\left(Y_i-\bar{Y}\right)^2$

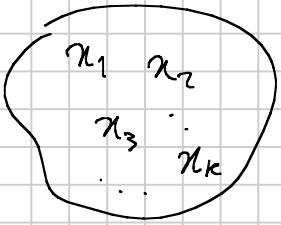
S_{YY}: devianza campionaria $\quad S_{YY} = \sum_i \left(Y_i-\bar{Y}\right)^2$

SS: devianza dei residui $\quad SS = \sum_i \left(Y_i - \hat{Y}_i\right)^2 \quad \hat{Y}_i = B\cdot x(i)$

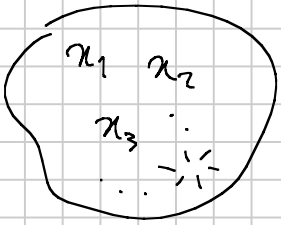
$$R^2 = 1 - \frac{SS}{S_{YY}} = 1 - \frac{(n-p-1)S_e^2}{(n-1)S_Y^2} = 1 - \underbrace{\frac{n-p-1}{n-1}}\ \frac{S_e^2}{S_Y^2}$$

circa 1 se $p \ll n$ come auspicabile

- Se si effettuano due regressioni, con o senza una var di ingresso x_k il valore di R^2 è necessariamente migliore (più alto) in quella con la variabile in più



R^2 maggiore



R^2 minore

→ Si può definire un indice R^2_c coeff di determinazione corretto che tiene conto anche di p:

$$R^2_c = f(R^2, p) \qquad R^2_c \leq R^2$$

È fatto in modo che se si toglie una var x_k non significativa, R^2_c non diminuisce, anzi, magari aumenta

$$SS = \|Y\|^2 - B^T X^T X B \quad = \|Y\|^2 - Y^T X B$$

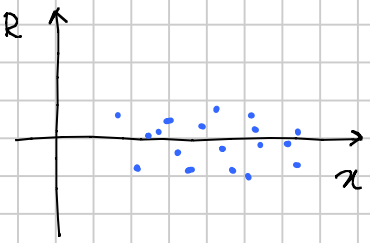
ora 14

$$\|Y\|^2 = \sum_i Y(i)^2$$

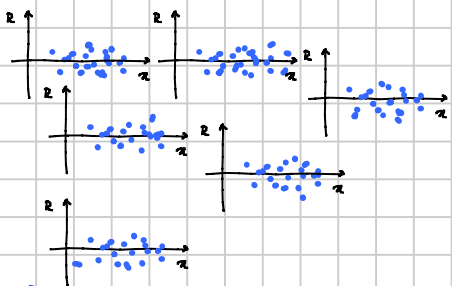
QUESTIONI PRATICHE SULLA REGRESSIONE

- Analisi dei residui

$p = 1$

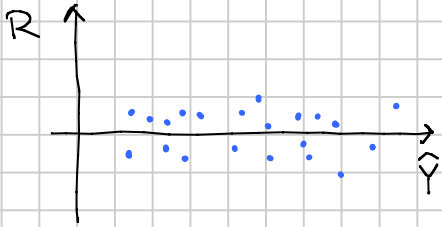


$p > 1$



un grafico per ogni x_i
→ p grafici (tutti diversi)

Possiamo anche farci dare il grafico previsti vs residui

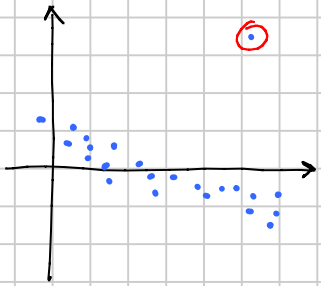


$p=1$ è lo stesso grafico perché $\hat{Y} = \alpha + \beta x$ è una trasf lineare di x

$p>1$ non è nessuno di quelli precedenti, ma ne sintetizza alcuni aspetti

★ MINITAB ad esempio mostra di default questo grafico.

- Outliers : sono punti → errati
  → provenienti da un'altra distribuzione



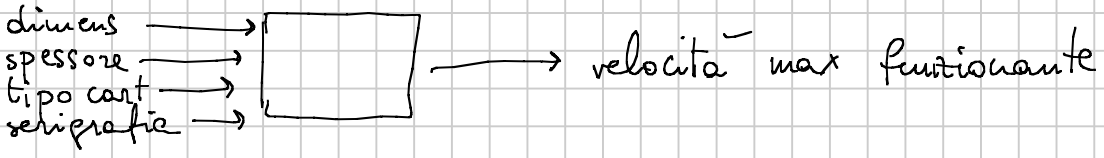
→ vanno capiti, studiati a parte e eliminati dal campione, poi si rifà
→ esempi : - stelle / seq principale
- case di Philadelphia

- Nonlinearità : cerchiamo un andamento a parabola o simile nei residui

→ se coinvolgono una variabile sola (o poche variabili) si corregge quella variabile, aggiungendo nel modello termini di grado più alto (vedi oltre)
→ in alternativa, soprattutto se le nonlinearità sono su tutte le var e si somigliano, si può tentare qualche trasformazione nonlineare della var di risposta (Y)

- Trasformazioni nonlineari di Y

es : macchina che crea confezioni / scatole di cartone



V = velocità $= f(x_1, x_2, \dots, x_p)$

$V \overset{?}{=} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ lineare

invece:

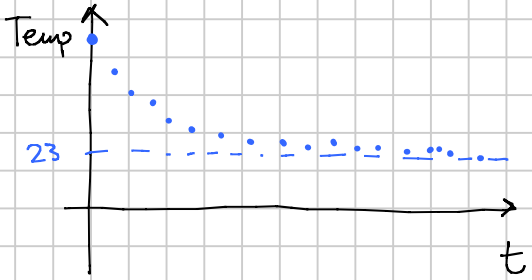
T = tempo dedicato a ogni cartone $\overset{?}{=} \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

$T = \beta_0 + \dots + \beta_p x_p + e$

$V = \frac{1}{T} = \frac{1}{\beta_0 + \dots + \beta_p x_p + e}$ nonline

In pratica si fa il reciproco di Y $\quad Y \longmapsto \frac{1}{Y} = T$
(una funzione nonlineare)

→ esempio: tempo di raffreddamento di metallo



andamento esponenziale

$Temp = 23 + e^{-\lambda t + \nu}$

$\log(Temp - 23) = \nu - \lambda t$

$\log(Temp - 23)$ è una misura della Temp che potrebbe dipendere linearmente da t

$Temp \longmapsto \log(Temp - 23)$ è una trasformazione nonlineare

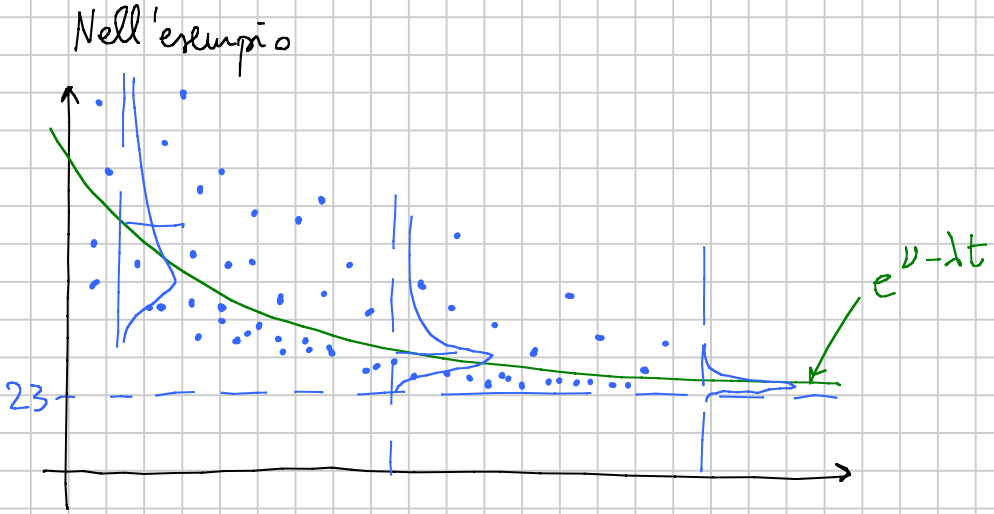
★ Occhio all'effetto sui residui:

supponiamo che $\log(Temp - 23) = \nu - \lambda t$ sia giusto

$\log(Temp - 23) = \nu - \lambda t + e$ $\quad e \sim \mathcal{N}(0, \sigma^2)$

Allora il modello originale è:

$Temp = 23 + e^{\nu - \lambda t + e} \neq 23 + e^{\nu - \lambda t} + e$

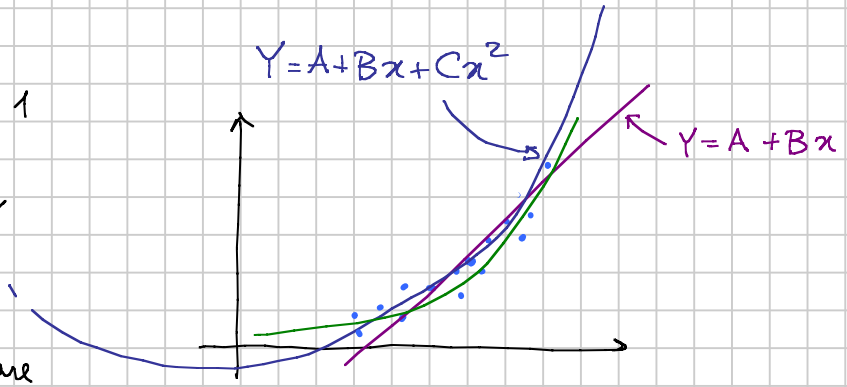

Nell'esempio
23
$e^{\nu-\lambda t}$

NONLINEARITÀ

1) Trasf. nonlineare di Y → può correggere
2) Aggiustare la nonlinearità di singole var di ingresso aggiungendo termini nonlineari

- Caso base: $p = 1$

$x \longrightarrow$ [nonlin] $\longrightarrow Y$



Se il modello lineare $Y = \alpha + \beta x$ non funziona, provo con polinomi di grado via via più alto, ogni volta controllando i residui

→ in pratica:

Y	x	x^2	x^3	...
$Y(1)$	$x(1)$	$x^2(1)$		
⋮	⋮	⋮		
$Y(n)$	$x(n)$	$x^2(n)$		

var ingresso di regr. lineare multipla

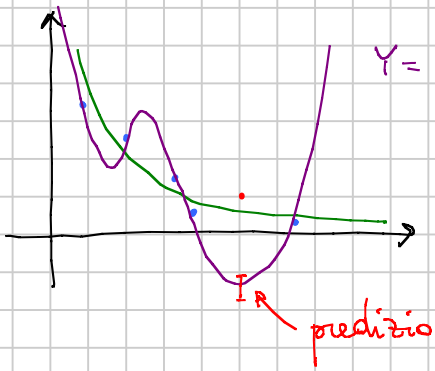
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

(x, x^2)

cioè si aggiungono variabili fittizie (dummy)

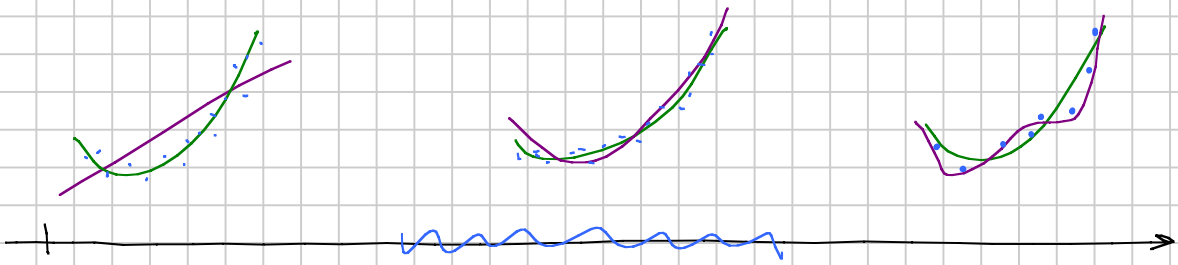
→ attenzione a non esagerare con il grado!

per n punti passa esattamente un solo polinomio di grado $n-1$



$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

in realtà il fenomeno è progressivo: se il grado p



★ La situazione in cui p è troppo alto si chiama overfitting. In generale, per essere abbastanza sicuri che non vi sia overfitting ci vole $n \gg p$.
In letteratura si trovano diverse convenzioni, come $\frac{n}{p} \geq 5$, $\frac{n}{p^2} \geq 2$

★ Ci si può accorgere che p è eccessivo se le variabili risultano non significative al test $H_0: \beta_k = 0$
→ in particolare i p dei dati vengono tutti alti e circa tutti uguali (spesso)

★ Inoltre maggiore è $\frac{n}{p}$, $\frac{n}{p^2}$, maggiore è la potenza dei test della regressione.

◎ Nonlinearità nella regressione multipla $p > 1$
tutto simile, ma più laborioso:

→ il lavoro va fatto su ogni variabile di ingresso

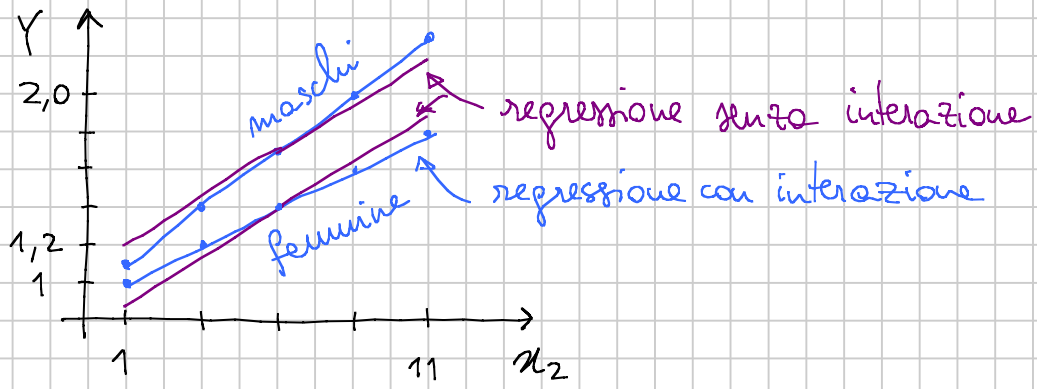
→ $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$
inoltre ci sono i termini di interazione!

x_2 →	1	3	5	7	11	— anni dalla laurea
maschi 0	1,1	1,4	1,7	2,0	2,3	
femmine 1	1	1,2	1,4	1,6	1,8	

↑ x_1

interazione ↓

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$



→ i termini di interazione sono necessari quando il modo in cui Y dipende da x_2 è legato a x_1

★

2	3	4	5	6	....	20
1	3	6	10	15		190

il numero di termini di interazione cresce rapidamente con p

$$\binom{p}{2} = \frac{p(p-1)}{2} \approx \frac{p^2}{2}$$

in questo modo, il "p effettivo", cioè il numero di coeff da stimare sul modello nonlineare sia troppo grosso (per n normali)

→ per questo motivo i termini di interazione da inserire nel modello vanno scelti con cura e parsimonia.

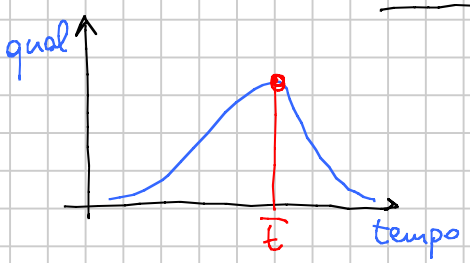
→ in ogni caso un termine inserito va testato ($H_0: \beta_k = 0$) per vedere se è significativo.

---

ora 16

● Regressione e ottimizzazione

Se lo scopo della regressione è massimizzare o minimizzare qualcosa...



dipendenza nonlineare

... i modelli lineari si adattano male, perché trovano max e min sempre sul bordo.

→ NB anche i termini di interazione mantengono questo problema

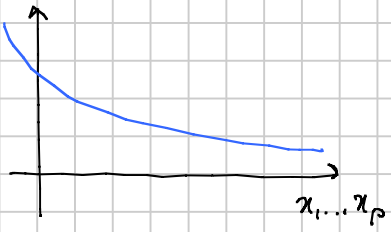
- Nel caso di metodo 1) : trasf nonlineare di Y

quali sono le trasf più comuni ?

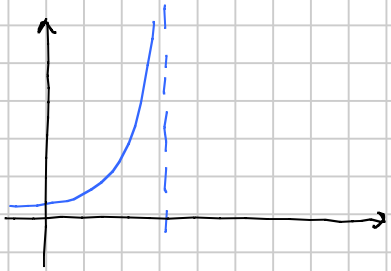
i. $\log(Y)$

$Y = e^{\beta_0 + \beta_1 x_1 + \dots}$

→ $\sqrt{Y}$, $\sqrt[3]{Y}$ sono simili

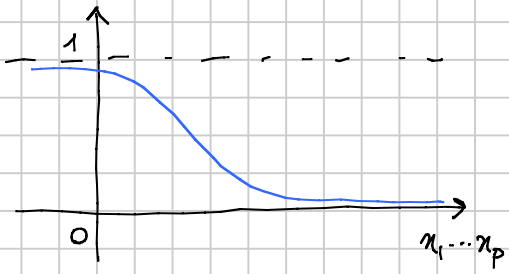


ii. Y^{-1}, Y^{-2}, ...



iii. $\log\left(\frac{Y}{1-Y}\right)$

"odds ratio"



forma sigmoide

"trasformazione logistica" = logit transform

$$Y = \frac{e^{\beta_0 + \beta_1 x_1 + \dots}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots}}$$

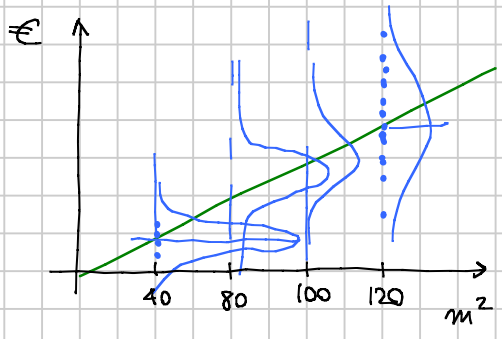
Se la sigmoide non va da 0 a 1, prima della trasf logistica, fare una trasf lineare in modo che Y vari circa tra 0 e 1.

MANCANZA DI OMOSCHEDASTICITÀ (problema 3)

Laborioso e spesso dagli effetti dubbi ...

= la varianza di $e \sim N(0, \sigma^2(x))$ dipende dalle var di ingresso (quindi anche le var dei residui)

- Diagnosi



prezzi case

$\sigma \propto x$ oppure $\sigma^2 \propto x$

oppure $\sigma \propto Y$ oppure $\sigma^2 \propto Y$

o altre dipendenze ancora ...

→ Se la variabilità è ±20% del prezzo ($\Rightarrow \sigma \propto Y$)

$Y = \beta x + e$ $\quad$ dev.st$(e) \approx 20\% \cdot \beta x$

$\Rightarrow \boxed{\sigma \approx 0{,}2 \cdot \beta \cdot x \propto x}$, ma anche $\boxed{\sigma \propto Y}$

(sia a x sia a Y perché $Y \propto x$)

$Y = \alpha + \beta x + e$ $\quad$ dev.st$(e) \approx 20\% (\alpha + \beta x)$

$\Rightarrow \boxed{\sigma \propto Y}$ ma $\sigma \not\propto x$

→ x: # incroci Y: tempo di percorrenza

$Y = \beta x + e$

$Y = T_1 + T_2 + \ldots + T_x$ modello mentale: il tempo complessivo è somma di tanti contributi indipendenti, uno per incrocio

se $\text{Var}(T_i) = a^2$ circa uguali $\Rightarrow \text{Var}(Y) = x a^2$

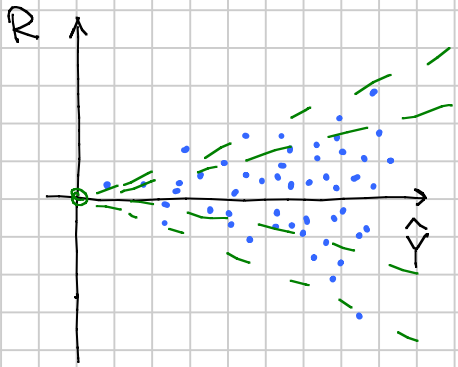
(si sommano per l'indipendenza)

$\underbrace{T_1 + T_1 + T_1 + \ldots + T_1}_{x} = x T_1 \Rightarrow \text{Var}(x T_1) = x^2 a^2$

$\sigma = \sqrt{\text{Var}(Y)} = \sqrt{x}\, a \propto \sqrt{x} \quad \Rightarrow \boxed{\sigma^2 \propto x} \quad \Rightarrow \boxed{\sigma^2 \propto Y}$

$Y = \alpha + \beta x + e \quad \Rightarrow \quad \boxed{\sigma^2 \propto x} \qquad \boxed{\sigma^2 \not\propto Y}$

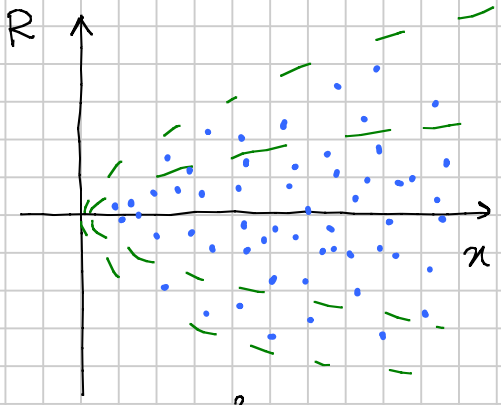
- Più frequentemente non abbiamo idea di motivi teorici per supporre una certa dipendenza ... comunque è bene guardare i residui



$\sigma \propto Y$

$\sigma \propto x$

stesso grafico x in ascissa



$\sigma^2 \propto x \quad \Rightarrow \quad \sigma \propto \sqrt{x}$

$\sigma^2 \propto Y$

stesso grafico Y in ascissa

laboratorio dalla sett prox spostato al mar 9.30 – 11.30

$\vec{x} \longrightarrow Y(n+1)$, $\dots$, $Y(n+m)$ } m risposte future indipendenti

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$$
$$e \sim \mathcal{N}(0, \sigma^2)$$
modello

$$Y(n+i) \sim \mathcal{N}(\vec{x}\cdot\beta ; \sigma^2) \qquad \forall i = 1, 2, \dots, m$$

$$V = \sum_{i=1}^{m} Y(n+i) \quad \text{la somma} \qquad V \sim \mathcal{N}(m\,\vec{x}\cdot\beta ; m\sigma^2)$$

$$V - m\vec{x}\cdot B \sim \mathcal{N}\left(0 ; m\sigma^2 + m^2 \underbrace{\sigma^2 \vec{x}^T (X^TX)^{-1}\vec{x}}_{\uparrow \text{Var}(\vec{x}\cdot B)}\right)$$

$$\vec{x}\cdot B \sim \mathcal{N}\left(\vec{x}\cdot\beta ; \vec{x}^T \underbrace{\sigma^2 (X^TX)^{-1}}_{\text{matrice di covarianza di } B} \vec{x}\right)$$

$$\frac{V - m\vec{x}\cdot B}{\sigma\sqrt{m + m^2 \vec{x}^T (X^TX)^{-1}\vec{x}}} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \frac{V - m\vec{x}\cdot B}{S_e\sqrt{m + m^2 \vec{x}^T (X^TX)^{-1}\vec{x}}} \sim t(n-p-1)$$

$$\frac{S_e^2}{\sigma^2}(n-p-1) \sim \chi^2(n-p-1) \quad \text{indip da } B$$

Concludendo, se q è un quantile $t(n-p-1)$ con prob $1-\alpha$ (bilaterale):

$$V \in m\vec{x}\cdot B \pm q\, S_e \sqrt{m + m^2 \vec{x}^T (X^TX)^{-1}\vec{x}} \quad \text{con prob } 1-\alpha$$

Analogamente per $U = \frac{V}{m}$ (media campionaria delle risposte future):

$$U \in \vec{x}\cdot B \pm q\, S_e \sqrt{\frac{1}{m} + \vec{x}^T (X^TX)^{-1}\vec{x}} \quad \text{con prob } 1-\alpha$$

- Riguardo al lab 05

$Y = f(x_1, x_2) + e$ — f nonlineare (di II grado)

x_1 = t = temperatura; x_2 = c = 1,2 canale

Solo canale 1 : $Y = \beta_0 + \beta_1 t + \beta_2 t^2 + e$

B_0, B_1, B_2 stimatori

Solo canale 2 : ... C_0, C_1, C_2

Multipla $Y = \alpha_0 + \alpha_1 t + \alpha_2 c + \alpha_3 t^2 + \alpha_4 ct + \alpha_5 ct^2$

$A_0, A_1, \ldots$ A_5 stimatori

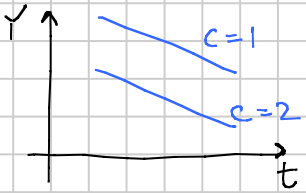
Provo a considerare il modello multiplo per $c = 1$

$$Y = \alpha_0 + \alpha_1 t + \alpha_2 + \alpha_3 t^2 + \alpha_4 t + \alpha_5 t^2$$
$$= (\alpha_0 + \alpha_2) + (\alpha_1 + \alpha_4) t + (\alpha_3 + \alpha_5) t^2$$

Si può verificare che $A_0 + A_2 = B_0$, $A_1 + A_4 = B_1$, $A_3 + A_5 = B_2$

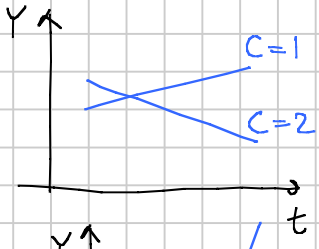
$A_0 + 2A_2 = C_0$, $A_1 + 2A_4 = C_1$, $A_3 + 2A_5 = C_2$

$Y = \alpha_0 + \alpha_1 t + \alpha_2 c$



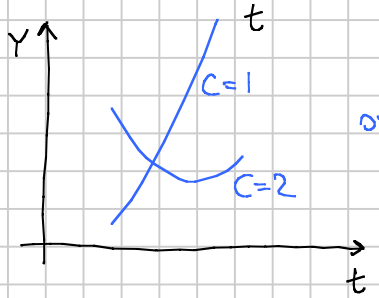
coeff ang α_1 per entrambe

$Y = \alpha_0 + \alpha_1 t + \alpha_2 c + \alpha_3 ct$



libere

$Y = \alpha_0 + \alpha_1 t + \alpha_2 c + \alpha_3 t^2 + \alpha_4 ct$



stesso coeff di II grado

Se si aggiunge $\alpha_5 ct^2$

si possono ottenere tutte le coppie di parabole possibili

★ L'unica differenza operativa tra il metodo in 2 casi e quello con regr multipla è che nel primo stimo due distinti valori di σ, nel secondo uno solo: se i due gruppi (canali nell'esempio) mostrano varianze diverse, devo separare, se no posso tenere assieme.
(Tenerle assieme è leggermente più potente.)

REGR NON OMOSCHEDASTICA: CURA

$Y = \beta \cdot x + e \qquad e \sim \mathcal{N}(0, \sigma^2(x)) \qquad$ ad es $\sigma \propto x$

$w(i) \quad i = 1, 2, \dots, n$ dei pesi che vanno presi inversamente proporzionali a σ^2

$w(i) \propto \frac{1}{\sigma^2(x(i))} \qquad$ nell'es $\qquad w(i) \propto \frac{1}{x(i)^2}$

- Si rifà daccapo la regressione, minimizzando $SS^{(w)}$

$$SS^{(w)} := \sum_{i=1}^{n} \left(Y(i) - x(i) \cdot B\right)^2 \cdot w(i)$$

infatti, se prima minimizzavo $SS := \sum_{i=1}^{n} \left(Y(i) - x(i) \cdot B\right)^2$

e $Y(i) - x(i) \cdot \beta \sim \mathcal{N}(0, \sigma^2)$, ora

$Y(i) - x(i) \cdot \beta \sim \mathcal{N}(0, \sigma(i)^2)$ ma

$\left(Y(i) - x(i) \cdot \beta\right)\sqrt{w(i)} \sim \mathcal{N}(0, \underbrace{w(i)\sigma(i)^2}_{\text{costante}})$

- In pratica, si "distribuisce" $\sqrt{w}$ nell'equazione

$$SS^{(w)} := \sum_{i=1}^{n} \left(\sqrt{w(i)}Y(i) - \sqrt{w(i)}\, x(i) \cdot B\right)^2$$

ovvero si modificano X e Y moltiplicando ogni riga per il corrispondente $w(i)$ e si fa una normale regressione.

Y	x_0	x_1	$\dots$	x_p
$Y(1)$	1	$x_1(1)$		$x_p(1)$
$Y(2)$	1	$x_1(2)$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		

var di ingresso

termine noto

Y'	x_0'	x_1'	$\dots$	x_p'
$Y(1)\sqrt{w(1)}$	$\sqrt{w(1)}$	$x_1(1)\sqrt{w(1)}$		$x_p(1)\sqrt{w(1)}$
$Y(2)\sqrt{w(2)}$	$\sqrt{w(2)}$	$x_1(2)\sqrt{w(2)}$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		

niente termine noto!

★ Lanciando la regressione su Y', X', si trovano i coefficienti $B^{(w)}$ che minimizzano $SS^{(w)}$

★ Se servono interv di conf o di predizione, si fanno su Y', X', poi si "torna" ai valori Y, X dividendo per $\sqrt{w(\vec{x})}$

★ Il coeff di determinazione R^2 del modello Y', X' non è confrontabile con quello di Y, X.

★ L'unica possibile verifica della bontà del procedimento è vedere se i residui di Y', X' ora sembrano omoschedastici.

FINE REGRESSIONE

ANALISI DELLA VARIANZA (ANOVA) (Cap 10)

- Un breve excursus su due test del capitolo 8

1. Test per il confronto delle medie di due popolazioni normali:

$$X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad \bar{X} \approx \mu_1 \qquad \bar{X} \sim \mathcal{N}\left(\mu_1; \frac{\sigma_1^2}{m}\right)$$

$$Y_1, \dots, Y_n \sim \mathcal{N}(\mu_2, \sigma_2^2) \qquad \bar{Y} \approx \mu_2 \qquad \bar{Y} \sim \mathcal{N}\left(\mu_2; \frac{\sigma_2^2}{n}\right)$$

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2 \quad \Rightarrow \quad H_0: \mu_1 - \mu_2 = 0 \quad H_1: \mu_1 - \mu_2 \neq 0$$

$$\bar{X} - \bar{Y} \approx \mu_1 - \mu_2 \qquad \bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0,1)$$

se σ_1 e σ_2 note, questa è la funz ancillare

In alternativa, se $\sigma_1 = \sigma = \sigma_2$ sono uguali (ipotesi di omoschedasticità)

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$$

dove S_p^2 è lo stimatore pooled della varianza

$$S_p^2 = \frac{m-1}{m+n-2}S_X^2 + \frac{n-1}{m+n-2}S_Y^2$$

media pesata di S_X^2 e S_Y^2

coeff positivi a somma 1

$$\frac{S_p^2}{\sigma^2}(m+n-2) \sim \chi^2(m+n-2)$$

2 Test per il confronto delle varianze di due popolazioni normali:

$X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu_1, \sigma_1^2)$ $\quad S_X^2 \approx \sigma_1^2$ $\quad \frac{S_X^2}{\sigma_1^2}(m-1) \sim \chi^2(m-1)$

$Y_1, \dots, Y_n \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\quad S_Y^2 \approx \sigma_2^2$ $\quad \frac{S_Y^2}{\sigma_2^2}(n-1) \sim \chi^2(n-1)$

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad H_1: \sigma_1^2 \neq \sigma_2^2$ $\Rightarrow$ $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $\quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

$$\frac{S_X^2}{S_Y^2} \approx \frac{\sigma_1^2}{\sigma_2^2}$$

- **Def**: Se W_1 e W_2 sono due v.a. chi-quadro, con a e b g.d.l. rispettivamente, e tra loro indipendenti, il loro rapporto ha (a meno di costanti) distribuzione F di Fisher:

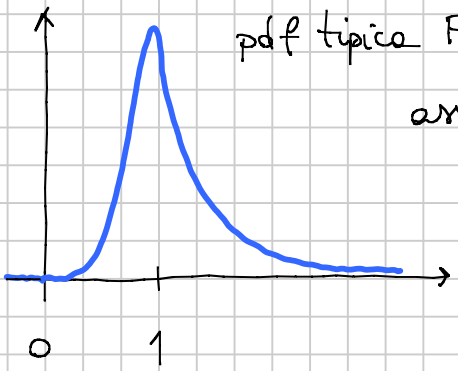
$W_1 \sim \chi^2(a)$

$W_2 \sim \chi^2(b)$

$$\frac{W_1/a}{W_2/b} \sim F(a;b)$$

"g.d.l. del denom."

"g.d.l. del numeratore"



pdf tipica F di Fisher

assume valori circa attorno a 1

$$\frac{S_X^2}{S_Y^2} \approx \frac{\sigma_1^2}{\sigma_2^2} \qquad W_1 = \frac{S_X^2}{\sigma_1^2}(m-1) \sim \chi^2(\overset{a}{m-1}) \qquad W_2 = \frac{S_Y^2}{\sigma_2^2}(n-1) \sim \chi^2(\overset{b}{n-1})$$

$$\frac{W_1/a}{W_2/b} = \frac{W_1/(m-1)}{W_2/(n-1)} = \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} = \frac{S_X^2}{S_Y^2} \Big/ \frac{\sigma_1^2}{\sigma_2^2} \sim F(\overset{a}{m-1}; \overset{b}{n-1})$$

$$\frac{\frac{S_X^2}{S_Y^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F(m-1; n-1)$$

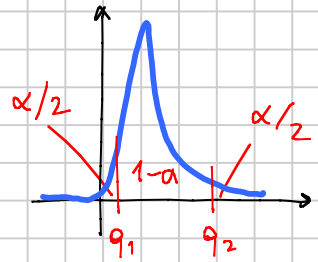
è una f. ancillare per $\frac{\sigma_1^2}{\sigma_2^2}$

→ La statistica del test si ottiene sostituendo 1 a $\frac{\sigma_1^2}{\sigma_2^2}$

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$V = \frac{S_X^2}{S_Y^2} \overset{H_0}{\sim} F(m-1; n-1)$$ statistica del test

→ livello di sign α (test bilaterale)



$$q_1 = \text{INV.F}(1-\alpha/2; m-1; n-1)$$

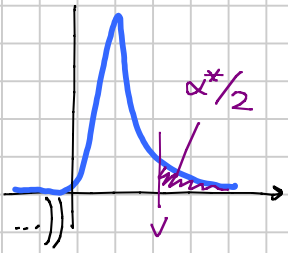
HW: verificare sintassi confrontando con le tabelle o gli esempi sul libro

$$q_2 = \text{INV.F}(\alpha/2; m-1; n-1)$$

... si guarda dove cade V

→ p dei dati

$$\alpha^* = 2 \min(\text{DISTRIB.F}(v; \ldots); 1 - \text{DISTRIB.F}(v; \ldots))$$



★ Questo test è prezioso come preliminare a [1] per testare l'ipotesi di omoschedasticità

★ Tutta l'ANOVA si basa su test simili a questo

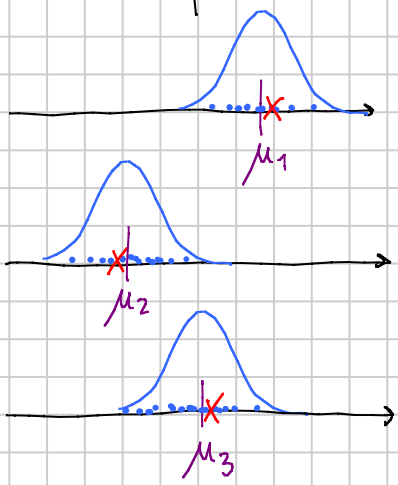
ANOVA A 1 VIA

Generalizza il test [1] (nel caso omoschedastico) quando i campioni da confrontare sono un numero qualsiasi.

→ Nel caso i campioni siano 2, [1] e ANOVA danno sempre lo stesso risultato (α^* coincidente)

ord 19

m campioni normali omoschedastici di medie $\mu_1, \mu_2, \dots, \mu_m$



$X_{1,1}, X_{1,2}, \dots, X_{1,n_1} \quad \mathcal{N}(\mu_1, \sigma^2)$

$X_{2,1}, X_{2,2}, \dots, X_{2,n_2} \quad \mathcal{N}(\mu_2, \sigma^2)$

$X_{3,1}, X_{3,2}, \dots, X_{3,n_3} \quad \mathcal{N}(\mu_3, \sigma^2)$

$\dots\dots$

$X_{m,1}, \dots \quad X_{m,n_m} \quad \mathcal{N}(\mu_m, \sigma^2)$

★ In questo contesto posso stimare σ^2 come facevo in [1]

$S_1^2, S_2^2, \dots, S_m^2$ varianze campionarie

Sono tutti stimatori indipendenti di σ^2.

Per ciascuno di essi vale: $\frac{S_k^2}{\sigma^2}(n_k - 1) \sim \chi^2(n_k - 1)$

● Def Si chiama devianza campionaria il prodotto di una qualunque varianza campionaria per i propri g.d.l.

Di solito si denota con SS (sum of squares)

→ Ad es. nella regressione $\frac{SS}{n-2} = s_e^2 \iff SS = s_e^2(n-2)$

$SS_1, SS_2, \dots, SS_m \qquad SS_k := s_k^2 (n_k - 1)$

perciò $\frac{SS_k}{\sigma^2} \sim \chi^2(n_k - 1)$

Quindi: $\frac{SS_1 + SS_2 + \dots + SS_m}{\sigma^2} \sim \chi^2\left(\sum_{k=1}^{m}(n_k - 1)\right)$

Ad esempio nel caso $m = 2$ $\qquad \frac{SS_1 + SS_2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$

$SS_1 = (n_1 - 1) s_1^2 \qquad SS_2 = (n_2 - 1) s_2^2$

$$\frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2 = s_p^2$$

$SS_p = SS_1 + SS_2 \qquad$ dove $\qquad SS_p = (n_1 + n_2 - 2) s_p^2$

★ In generale per $m \geq 2$ il termine pooled viene sostituito con within
(= entro i campioni ... poi ci sarà quella tra i campioni)

$SS_W = SS_1 + SS_2 + \dots + SS_m \qquad$ devianza within

$\frac{SS_W}{\sigma^2} \sim \chi^2\left(\sum_k (n_k - 1)\right) \qquad$ perciò $\qquad \frac{SS_W}{\sum_k (n_k - 1)} \approx \sigma^2$ stimatore corretto

$$s_W^2 = \frac{SS_W}{\sum_k (n_k - 1)} \qquad \text{varianza within} \qquad s_W^2 \approx \sigma^2$$

$$s_W^2 = \sum_j \frac{SS_j}{\sum_k (n_k - 1)} = \sum_j \frac{n_j - 1}{\sum_k (n_k - 1)} s_j^2$$

media pesata delle s_j^2

coeff positivi a somma 1

★ Per semplicità di notazione supponiamo che tutti i campioni abbiano la stessa numerosità n : $\qquad n_k = n \quad \forall k$

$$S_W^2 = \frac{SSw}{mn-m}$$

★ Sia $X_{k,*} = \bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{k,i}$ la media del k-esimo campione

Se i campioni hanno tutti la stessa media (vera) ....

$$\mu_1 = \mu_2 = \dots = \mu_m = \mu$$

$$X_{k*} \sim \mathcal{N}\left(\mu_k, \frac{\sigma^2}{n}\right) \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$X_{1,*}, X_{2,*} \dots, X_{m,*}$ sono un campione iid di v.a. $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Sia S_M^2 la var campionaria di questo campione

$$S_M^2 = \frac{1}{m-1}\sum_{k=1}^{m}\left(X_{k,*} - X_{**}\right)^2 \qquad S_M^2 \approx \sigma^2/n$$

↑ media campionaria degli X_{k*}

$$\frac{S_M^2}{\sigma^2/n}(m-1) \sim \chi^2(m-1)$$

$$S_B^2 := n S_M^2 \qquad \text{varianza between}$$

$$\frac{S_B^2}{\sigma^2}(m-1) \sim \chi^2(m-1) \qquad S_B^2 \approx \sigma^2 \quad \text{stimatore corretto}$$

$$SS_B = (m-1) S_B^2 \qquad \text{devianza between}$$

- Test dell'anova a 1 via (n_k tutti uguali)

$H_0: \mu_1 = \mu_2 = \dots = \mu_m$ $\qquad$ H_1: non tutte uguali

→ Se è vera H_0, S_W^2 e S_B^2 sono entrambi stimatori corretti di σ^2

Quindi: $\frac{S_B^2}{S_W^2} \sim F(m-1; mn-m)$

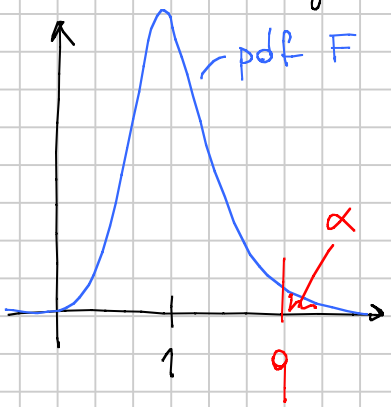
infatti $\frac{S_B^2}{\sigma^2}(m-1) \sim \chi^2(m-1)$ $\qquad$ $\frac{S_W^2}{\sigma^2}(mn-m) \sim \chi^2(mn-m)$

$$\frac{S_B^2}{\sigma^2} \Big/ \frac{S_W^2}{\sigma^2} = \frac{S_B^2}{S_W^2} \sim F(\dots)$$

→ Se è vera H_1, S_W^2 è comunque uno stimatore corretto di σ^2 mentre S_B^2 è sempre grande ($=$ sovrastima σ^2) perché le $X_{k,*}$ sono "sgranate".

Perciò sotto H_1 la statistica $\frac{S_B^2}{S_W^2}$ sarà tipicamente (molto) più grande di 1

☆



Si prende sempre un quantile unilaterale e sempre a destra

$$q = INV.F(\alpha;\ m-1;\ m*n-m)$$

DATE PROPOSTE PER GLI APPELLI

	Crocette & Excel	Scritto (Nicolodi)	Orale
1	19/01/2010	28/01/2010	08/02/2010
2	09/02/2010	17/02/2010	19/02/2010
3	08/06/2010	17/06/2010	28/06/2010
4	29/06/2010	08/07/2010	19/07/2010
5	01/09/2010	13/09/2010	16/09/2010
6	15/09/2010	27/09/2010	30/09/2010

ANOVA A 1 VIA

m campioni normali → varianza costante (omosched.)
↳ medie forse uguali o forse no (test)

$X_{1,1}, X_{1,2}, \dots, X_{1,n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$

$X_{i,1}, X_{i,2}, \dots, X_{i,n_i} \sim \mathcal{N}(\mu_i, \sigma^2) \qquad i = 1, 2, \dots, m$

$n_1, n_2, \dots, n_m$ le numerosità $\qquad N = n_1 + n_2 + \dots + n_m$

- Due stimatori della varianza

1) $$S_W^2 = \sum_{i=1}^{m} \frac{n_i - 1}{N - m} \cdot \underbrace{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{i,j} - X_{i,*})^2}_{S_i^2 \text{ var. camp. del campione } i}$$

$$= \frac{1}{N-m} \sum_{i,j} (X_{i,j} - X_{i,*})^2$$

within (generalizzazione dello stimatore pooled)

dove $X_{i,*} = \underbrace{\frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}}_{\overline{X_i} \text{ media camp del camp. } i}$

$$\frac{S_W^2}{\sigma^2}(N-m) \sim \chi^2(N-m)$$

2) $$S_B^2 = \frac{1}{m-1} \sum_{i=1}^{m} n_i (X_{i*} - X_{**})^2$$

between

dove $X_{**} = \underbrace{\frac{1}{N} \sum_{i,j} X_{i,j}}_{\text{media camp. totale}} = \sum_{i=1}^{m} \frac{n_i}{N} X_{i,*}$

$\neq \frac{1}{m} \sum_{i=1}^{m} X_{i*}$

$\frac{S_B^2}{\sigma^2}(m-1) \overset{H_0}{\sim} \chi^2(m-1)$ vera solo se è vera $H_0: \mu_1 = \mu_2 = \dots = \mu_m$

se H_0 è falsa, S_B^2 in genere sovrastima σ^2

★ $x_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ se H_0 è vera $x_{ij} \sim \mathcal{N}(\mu, \sigma^2)$

$x_{i*} = \bar{x}_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n_i}\right)$

$\frac{x_{i*} - \mu}{\sigma/\sqrt{n_i}} \sim \mathcal{N}(0,1)$

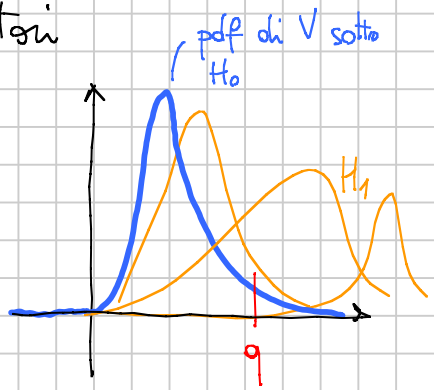
$\frac{1}{\sigma^2}\sum_{i=1}^{m} n_i (x_{i*} - \mu)^2 = \sum_{i=1}^{m} \left(\frac{x_{i*} - \mu}{\sigma/\sqrt{n_i}}\right)^2 \sim \chi^2(m)$ definizione di χ^2

Se al posto di μ metto $x_{**} \approx \mu$ per il solito thm

(verificare le ipotesi) $m \longrightarrow m-1$

$\frac{1}{\sigma^2}\sum_{i=1}^{m} n_i (x_{i*} - x_{**})^2 \sim \chi^2(m-1)$ CVD

- L'ANOVA funziona confrontando i due stimatori



$V = \frac{S_B^2}{S_W^2}$ statistica del test

→ sotto H_0, $V \sim F(m-1; N-m)$

→ sotto H_1, V è in genere più grande

(quindi il test viene sempre fatto con quantili unilaterali)

- Identità algebrica delle deviazioni

$S_W^2(N-m) = SS_W$

$S_B^2(m-1) = SS_B$

$S^2(N-1) = SS_x$

$\boxed{SS_x = SS_W + SS_B}$

$S^2 = \frac{1}{N-1}\sum_{i,j}(x_{ij} - x_{**})^2$ varianza campionaria globale

→ stima σ^2 sotto H_0, sovrastima sotto H_1

→ Si usa spesso per ricavare $S_B^2 = \frac{1}{m-1}(SS_x - SS_W)$

$$\rightarrow \quad \frac{SS_X}{\sigma^2} = \frac{SS_W}{\sigma^2} + \frac{SS_B}{\sigma^2}$$

(dal punto di vista statistico)

$\frac{SS_X}{\sigma^2} \sim_{H_0} \chi^2(N-1)$ $\quad$ $\frac{SS_W}{\sigma^2} \sim \chi^2(N-m)$ $\quad$ $\frac{SS_B}{\sigma^2} \sim_{H_0} \chi^2(m-1)$

sono indipendenti (H_0)

DUE COLLEGAMENTI CON LA REGRESSIONE

Test globale di regressione

$$Y = \sum_{i=0}^{p} \beta_i x_i + e \qquad x_0 = 1 \qquad e \sim \mathcal{N}(0, \sigma^2)$$

C'è regressione (globalmente)?

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ $\qquad$ H_1 : non tutti nulli

S_e^2 è l'analogo di S_W^2

$$SS_W = \sum_{ij} (X_{ij} - X_{i\cdot})^2$$

$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ $\quad$ $X_{i\cdot} \approx \mu_i$

$$SS_e = SS = \sum_{i=1}^{n} \left[Y(i) - (B_0 + B_1 x_1(i) + B_2 x_2(i) + \dots + B_p x_p(i)) \right]^2$$

$Y(i) \sim \mathcal{N}(\beta \cdot x(i), \sigma^2)$ $\quad$ $B \cdot x(i) \approx \beta \cdot x(i)$

S_Y^2 ha lo stesso ruolo di S^2

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y(i) - \bar{Y})^2 \qquad SS_Y = S_Y^2 (n-1)$$

dev. totale

$$\frac{SS_Y}{\sigma^2} = \frac{SS_e}{\sigma^2} + \frac{?}{\sigma^2}$$

dev. residui $\quad$ analogo di SS_B (dev. spiegata)

$$SS_D := SS_Y - SS_e$$

devianza spiegata

dev. spiegata

$$R^2 = 1 - \frac{SS_e}{SS_Y} = \frac{SS_Y - SS_e}{SS_Y}$$

$$\frac{SS_Y}{\sigma^2} = \frac{SS_e}{\sigma^2} + \frac{SS_D}{\sigma^2}$$

$\frac{SS_Y}{\sigma^2} \sim_{H_0} \chi^2(n-1)$ $\quad$ $\frac{SS_e}{\sigma^2} \sim \chi(n-p-1)$ $\quad$ $\frac{SS_D}{\sigma^2} \sim_{H_0} \chi^2(p)$

indipendenti (sotto H_0)

→ Posso effettuare il test per la regressione globale così:

$V = \frac{S_D^2}{S_e^2}$ statistica:

sotto H_0 ha legge $F(p; n-p-1)$

sotto H_1 in genere assume valori maggiori

(quindi per il test si usano quantili unilaterali)

ora 21

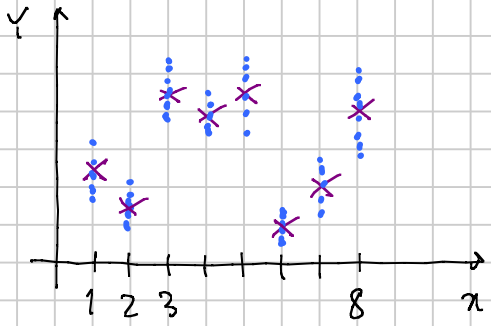
ANOVA come regressione con var di ingresso categorica.

Esempio → verbali di multe

id	data	persona	# di verbali
1	~	1	17
2	~	6	9
3	~	2	15
...	...	1	24
		8	32
		2	6
		⋮	⋮

Q: il # verbali (Y) dipende dalla persona (x)?

Si tratta di due var numeriche, quindi faccio la regressione .... o no?

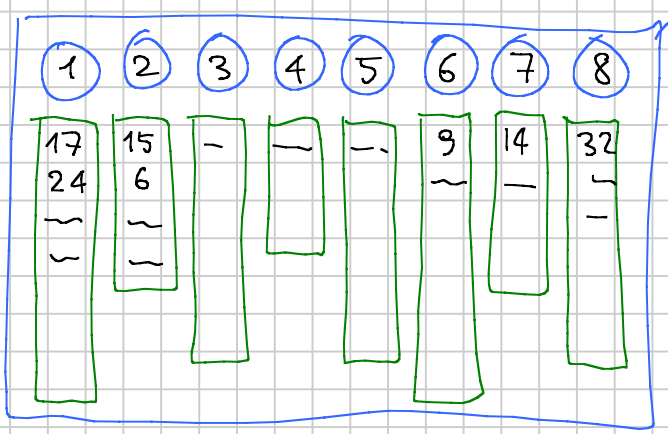


siccome 1, 2, ..., 8 (in questo ordine) è una codifica arbitraria la regressione non viene

✱ la var "persona" infatti è una variabile categorica

1) Modo migliore: ANOVA a UNA VIA

si riorganizzano i dati:



H_1: le medie non sono tutte uguali → il numero di verbali dipende dalla persona

H_0: tutti gli operatori sono uguali

Proseguendo l'analogia, il modello "di regressione" sarebbe

$Y = \mu_x + e$ $\quad Y = \alpha + \beta x + e$ $\quad Y = e^{\alpha+\beta x} + e$ $\quad Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$

$x = 1, 2, \dots, m$

2) C'è anche modo di recuperare la regressione
È meno potente e ha un sacco di difetti, però permette ad esempio di aggiungere altre variabili (in questo esempio la data, per vedere se c'è un trend)

Si "esplode" la variabile categorica con k scelte in $k-1$ variabili dicotomiche

persona	d1	d2	d3	d4	d5	d6	d7
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	0	1	0	0
6	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1
(8)	0	0	0	0	0	0	0

la scelgo come categoria "di default"

id	data	persona	#di verbali	d1	d2	d3	...	d6	...
1	~	1	17	1	0	0		0	0
2	~	6	9	0	0	0	...	1	0
...									

$Y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \dots + \beta_7 d_7 + e$

$\beta_0 = \mu_8$ risposta della persona 8

$\beta_0 + \beta_3 = \mu_3$ risposta della persona 3

$\beta_3 = \mu_3 - \mu_8$ differenza tra la pers 3 e la 8

Ora posso inserire la var data (x)

$Y = \beta_0 + \sum_1^7 \beta_i d_i + \beta_8 x + e$

→ esempio di scelta del default : stato civile

	default	coniug.	divor.	vedovo
– celibe/nubile		0	0	0
– coniugato		1	0	0
– divorziato		0	1	0
– vedovo		0	0	1

ipotizzo dati sull'uso di internet

x categorica → [ANOVA 1 via] → Y numerica omosch

x_1 numerica, x_2 numerica, ... → [REGR] → Y numerica omosch

x_1 categorica, x_2 categorica → [ANOVA 2 vie] → Y numerica omosch.

ANOVA a DUE VIE

esempio : Y : vendite di sigarette

x_1 : giorno della settimana (6 categ : lun ... sab)

x_2 : marca (4 categ : A, B, C, D)

id	x_1	x_2	Y
1	sa	B	12
2			
...			

	lu	ma	me	gi	ve	sa
A		×	×	×	×	×
B	×	×	×	×	×	12
C	×	××	××	×	×	×
D	×	×	×		×	×

n colonne

m righe

★ L'ANOVA a 2 VIE richiede esattamente un dato per ogni cella.

i. se ci sono dei buchi al limite posso pensare di togliere tutta la riga o la colonna

ii. se in ogni cella ci sono k dati, posso pensare di fare la media in ogni cella

iii. se le numerosità cambiano, posso pensare di buttare via qualche dato (sempre scelto a caso)

- Ipotesi di lavoro:
  - normalità + omosch. $X_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$ dato cella i,j
  - **effetto additivo** $\mu_{ij} = \mu + \alpha_i + \beta_j$ ← effetto colonna; α_i: effetto riga; μ: "media globale"
  - $\alpha_1, \alpha_2, \dots, \alpha_m$ e $\beta_1, \beta_2, \dots, \beta_n$ sono un po' positivi e un po' negativi, e $\sum_{i=1}^{m} \alpha_i = 0 = \sum_{j=1}^{n} \beta_j$

★ Sull'effetto additivo

vuol dire ad esempio che mediamente tutte le marche A, ..., D vendono 7 pacchetti in più il sabato, rispetto al mercoledì

	me	sa
A	6	13
B	4	11
C	1	8
D	2	9

(−4 tra B e C; μ_{ij} tra D e 9; +7 effetto additivo)

$\mu_{ij} = \mu + \alpha_i + \beta_j$ ok

Forse è più ragionevole:

	me	sa
	6	9
	4	6
	1	1,5
	2	3

(−66% ×0,33; +50% ×1,5 effetto moltiplicativo)

$\mu_{ij} = \mu \times \alpha_i \times \beta_j$ KO

★ Nel caso si sospetti che gli effetti siano moltiplicativi si può pensare di fare il logaritmo di tutti i dati

$$\underbrace{\log \mu_{ij}}_{\mu'_{ij}} = \log(\mu \cdot \alpha_i \cdot \beta_j) = \underbrace{\log \mu}_{\mu'} + \underbrace{\log \alpha_i}_{\alpha'_i} + \underbrace{\log \beta_j}_{\beta'_j}$$

Funziona bene se i dati erano lognormali ...

- Che ipotesi verifica l'ANOVA a 2 vie?

Vengono fatti due test distinti

a) Verifica se vi sia effetto riga

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$ $\quad H_1$: non tutti nulli

b) Verifica se vi sia effetto colonna

$H_0: \beta_1 = \beta_2 = \dots = \beta_n = 0$ $\quad H_1$: non tutti nulli

- Come si fa l'ANOVA a 2 vie "a mano"?

→ Come si stimano α_i, β_j e μ?

$$X_{i*} = \overline{X_i} = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad \text{medie per riga} \qquad i = 1, 2, \dots, m$$

$$X_{*j} = \overline{X^j} = \frac{1}{m}\sum_{i=1}^{m} X_{ij} \qquad \text{medie per colonna} \qquad j = 1, \dots, n$$

$$X_{**} = \overline{X} = \frac{1}{mn}\sum_{i,j} X_{ij} = \frac{1}{m}\sum_{i=1}^{m} X_{i*} = \frac{1}{n}\sum_{j=1}^{n} X_{*j} \qquad \text{media globale}$$

i. X_{i*}, X_{*j}, X_{**} tutte normali

ii. Calcolo i valori attesi

$$E(X_{i*}) = \frac{1}{n}\sum_j \underbrace{E(X_{ij})}_{\mu_{ij}} = \frac{1}{n}\sum_j (\mu + \alpha_i + \beta_j) = \mu + \alpha_i + 0 = \mu + \alpha_i$$

$$E(X_{*j}) = \mu + \beta_j \qquad E(X_{**}) = \mu$$

iii. Varianze

$$\mathrm{Var}(X_{i*}) = \frac{1}{n^2}\sum_{j=1}^{n} \underbrace{\mathrm{Var}(X_{ij})}_{\sigma^2} = \frac{\sigma^2}{n} \qquad \mathrm{Var}(X_{*j}) = \frac{\sigma^2}{m} \qquad \mathrm{Var}(X_{**}) = \frac{\sigma^2}{mn}$$

(indip)

iv. $\alpha_i = (\mu + \alpha_i) - \mu \approx X_{i*} - X_{**}$

$$\left.\begin{aligned} \alpha_i &\approx X_{i*} - X_{**} \\ \beta_j &\approx X_{*j} - X_{**} \\ \mu &\approx X_{**} \end{aligned}\right\} \quad \mu_{ij} = \mu + \alpha_i + \beta_j \approx X_{i*} + X_{*j} - X_{**}$$

★ Questi stimatori hanno legge normale con media ovvia (sono stimatori corretti) e varianza complicata (ad esempio X_{i*} e X_{**} non sono indipendenti)

→ Calcolo varianze e devianze campionarie delle medie per riga e per colonna

$X_{1*}\ X_{2*}\ \dots\ X_{m*}$ medie per riga X_{**} è la loro $\bar{X}$

$$S^2_{MR} = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_{i*} - X_{**}\right)^2 \quad \approx \frac{\sigma^2}{n} \qquad m-1 \text{ g.d.l.}$$

sotto H_0 2) ovvero se $\alpha_i \equiv 0$

ovvero se $X_{i*} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

ANOVA A DUE VIE

$X_{i,j}$ $\quad \begin{matrix} x_{1,1} & \dots & x_{1,n} \\ x_{2,1} & \dots & x_{2,n} \\ \vdots & & \\ x_{m,1} & \dots & x_{m,n} \end{matrix}$ matrice completa $\quad X_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2)$ indipendenti

$\longrightarrow$ modello additivo $\quad \mu_{i,j} = \mu + \alpha_i + \beta_j$

α_i: effetti riga $\quad \alpha_1, \dots, \alpha_m$

β_j: effetti colonna $\quad \beta_1, \dots, \beta_n$

$\longrightarrow \sum_{i=1}^{m} \alpha_i = 0 = \sum_{j=1}^{n} \beta_j$

$\begin{matrix} x_{1,1} & \dots & x_{1,n} & \longrightarrow & x_{1,*} \\ x_{2,1} & \dots & x_{2,n} & \longrightarrow & x_{2,*} \\ \vdots & & & & \\ x_{m,1} & \dots & x_{m,n} & \longrightarrow & x_{m,*} \\ \downarrow & & \downarrow & \searrow & \\ x_{*,1} & \dots & x_{*,n} & & x_{*,*} \end{matrix}$

medie per riga $\quad X_{i,*} := \frac{1}{n}\sum_{j=1}^{n} X_{i,j} \sim \mathcal{N}\left(\mu + \alpha_i, \frac{\sigma^2}{n}\right)$

media totale $\quad X_{*,*} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{mn}\right)$

medie per colonna $\quad X_{j,*} \sim \mathcal{N}\left(\mu + \beta_j, \frac{\sigma^2}{m}\right)$

$\longrightarrow \begin{cases} \alpha_i \approx X_{i,*} - X_{*,*} \\ \beta_j \approx X_{*,j} - X_{*,*} \\ \mu \approx X_{*,*} \end{cases}$ stimatori corretti (e consistenti)

STIMATORE (BUONO SEMPRE) DI σ^2

È l'analogo degli stimatori pooled e within

$$\mu_{i,j} = \mu + \alpha_i + \beta_j \approx \cancel{X_{*,*}} + X_{i,*} - \cancel{X_{*,*}} + X_{*,j} - X_{*,*}$$

$$\mu_{i,j} \approx X_{i,*} + X_{*,j} - X_{*,*} \sim \mathcal{N}(\mu_{i,j}, ?)$$

$$X_{i,j} - (X_{i,*} + X_{*,j} - X_{*,*}) \sim \mathcal{N}(0, ?)$$

$X_{i,j}$: $\mu_{i,j}$ $\quad$ $(X_{i,*} + X_{*,j} - X_{*,*})$: $\mu_{i,j}$

HW: $\mathrm{Var}\left(X_{i,j} - X_{i,*} - X_{*,j} + X_{*,*}\right) = \sigma^2 \frac{(m-1)(n-1)}{mn}$

$$\sum_{i,j} \frac{(x_{i,j} - \mu_{i,j})^2}{\sigma^2} \sim \chi^2(mn)$$ fatto ovvio

$$\sum_{i,j} \frac{(x_{i,j} - \overbrace{x_{i,*}}^{\text{stimatore}} - x_{*,j} + x_{**})^2}{\sigma^2} \sim \chi^2(mn - m - n + 1) \sim \chi^2((m-1)(n-1))$$

gdl persi

HW: Verificare ipotesi e applicazione del "teorema chi-quadro"

$$SS_e = \sum_{i,j} (x_{i,j} - x_{i,*} - x_{*,j} + x_{*,*})^2$$ devianza errori

$$S_e^2 = \frac{SS_e}{(m-1)(n-1)}$$ varianza errori

$$\frac{SS_e}{\sigma^2} \sim \chi^2((m-1)(n-1))$$

HW: Calcolare $\mathbb{E}$ e Var di S_e^2

$\longrightarrow \sigma^2 \approx S_e^2$ stimatore corretto e consistente

STIMATORI DI σ^2 BUONI SOTTO H_0

a) $H_0: \alpha_i \equiv 0 \ \forall i$ H_1: non tutti nulli test sull'effetto riga

b) $H_0: \beta_j \equiv 0 \ \forall j$ H_1: non tutti nulli test sull'effetto colonna

$$x_{i,*} \sim \mathcal{N}\left(\mu + \alpha_i, \frac{\sigma^2}{n}\right)$$

sotto H_0 a)

x_{**} al posto di μ e colonna di 1 i gdl

$$x_{i,*} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \sum_{i=1}^{m} n \frac{(x_{i,*} - \mu)^2}{\sigma^2} \sim \chi^2(m)$$

$$SS_R = \sum_{i=1}^{m} n (x_{i,*} - x_{**})^2$$ devianza righe

$$\frac{SS_R}{\sigma^2} \overset{H_0 a)}{\sim} \chi^2(m-1) \qquad S_R^2 = \frac{SS_R}{m-1}$$ varianza righe

★ Sempre sotto H_0 a) S_e^2 e S_R^2 sono indipendenti, quindi:

$$V := \frac{S_R^2}{S_e^2} \overset{H_0 a)}{\sim} F(m-1; (m-1)(n-1))$$ statistica test a)

★ Sotto H_1 a) V tende ad assumere valori più grandi

Analogamente ...

sotto H_0 b)

$$SS_C = \sum_{j=1}^{n} m\left(x_{*,j} - x_{**}\right)^2 \qquad \text{devianza colonne} \qquad \frac{SS_C}{\sigma^2} \overset{H_0^{b)}}{\sim} \chi^2(n-1)$$

$$S_C^2 = \frac{SS_C}{n-1} \qquad \text{varianza colonne}$$

$$\bigstar \quad W := \frac{S_C^2}{S_e^2} \overset{H_0^{b)}}{\sim} F\left(n-1\,;(m-1)(n-1)\right) \qquad \text{statistica del test b)}$$

$\bigstar$ Sotto $H_1^{b)}$ W tende ad assumere valori più grandi

ANALISI DEI RESIDUI

→ esempio: dieta delle banane – 10 persone per il test

pesi prima:	pesi dopo:
82 $= x_{1,1}$	85 $= x_{1,2}$
97 $= x_{2,1}$	94 $= x_{2,2}$
75	74
63	68
68	67
79	80
...	...
111 $= x_{10,1}$	109 $= x_{10,2}$

$$x_{1,1} \sim \mathcal{N}\left(\mu_{1,1}; \sigma^2\right)$$
$$x_{2,1} \sim \mathcal{N}\left(\mu_{2,1}; \sigma^2\right)$$

effetto dieta

$$x_{1,2} \sim \mathcal{N}\left(\mu_{1,1} + \gamma\,; \sigma^2\right)$$
$$x_{2,2} \sim \mathcal{N}\left(\mu_{2,1} + \gamma\,; \sigma^2\right)$$
$$x_{i,2} \sim \mathcal{N}\left(\mu_{i,1} + \gamma, \sigma^2\right)$$

sto supponendo che l'effetto della dieta sia lo stesso per tutti

$$\mu := \frac{1}{10}\sum_{i=1}^{10} \mu_{i,1} + \frac{\gamma}{2} \qquad \alpha_i := \mu_{i,1} + \frac{\gamma}{2} - \mu \qquad \beta_1 := -\frac{\gamma}{2} \quad \beta_2 := \frac{\gamma}{2}$$

$\bigstar$ ho definito tutto in modo che:

→ $\boxed{\mu_{i,j} = \mu + \alpha_i + \beta_j \quad \forall i,j}$

$$\mu + \alpha_i + \beta_1 = \mu_{i,1} + \frac{\gamma}{2} - \frac{\gamma}{2} = \mu_{i,1}$$
$$\mu + \alpha_i + \beta_2 = \mu_{i,1} + \frac{\gamma}{2} + \frac{\gamma}{2} = \mu_{i,1} + \gamma = \mu_{i,2}$$

ok

→ $\sum_{j=1}^{2} \beta_j = 0$

→ $\sum_{i=1}^{10} \alpha_i = \sum_{i=1}^{10} \mu_{i,1} + 10\frac{\gamma}{2} - 10\mu = \sum_i \mu_{i,1} + 10\frac{\gamma}{2} - \sum_i \mu_{i,1} - 10\frac{\gamma}{2} = 0$

ora 24

82	85	83,5
97	94	95,5
75	74	74,5
.....		
82,2	83,4	82,8

$\beta_1 \approx 82{,}2 - 82{,}8 = -0{,}6$
$\beta_2 \simeq 83{,}4 - 82{,}8 = 0{,}6$

$\alpha_1 \approx 83{,}5 - 82{,}8 = 0{,}7$
$\alpha_2 \approx 12{,}7$
$\alpha_3 \approx -8{,}3$

diff = effetto dieta $+ 1{,}2 \approx \gamma$

- Calcolo i previsti

$\mu_{1,1} \approx 82$ ma meglio $\mu_{1,1} \approx 82{,}8 + 0{,}7 + (-0{,}6) = 82{,}9$

è una stima che usa tutti i dati

previsti ↓

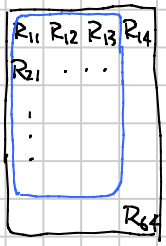
82,9	84,1	83,5
94,9	96,1	95,5
73,9	85,1	...
...	...	
82,2	83,4	

$82{,}8 - 8{,}3 - 0{,}6 = 72{,}8 + 1{,}1 = 73{,}9$

- Calcolo i residui (osservato − previsto)

−0,9	0,9	0
2,1	−2,1	0
1,1	−1,1	0
...	...	
0	0	0

i residui non possono per definizione avere effetti riga e colonna



i numeri liberi sono meno di $m \cdot n$

sono: $(m-1)(n-1)$

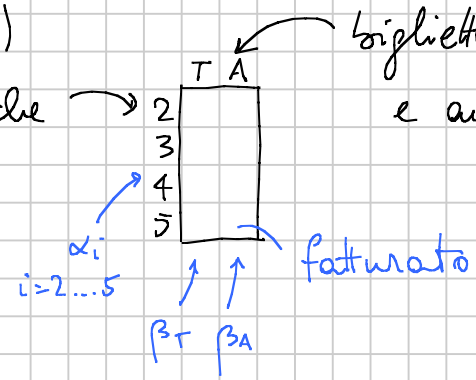
(ovvero i g d l dello stimatore S_e^2)

SS_e = somma quadrati tutti i residui

☆ A parte il "bordo" i residui dovrebbero essere normali di media 0 e var circa S_e^2, indipendenti e senza struttura

▨ esempio 2: (sbagliato)

biglietterie automatiche in funzione



biglietterie tradizionali e automatiche

$\mu_{ij} = \mu + \alpha_i + \beta_j$

$\beta_T = -\beta_A$ $\beta_A = \frac{\gamma}{2}$ dove γ è la differenza tra A e T

☆ Non è naturale che γ sia lo stesso per tutte le righe

In realtà osservo:

65	12	38,5
64	14	39
62	17	39,5
60	20	40
6275	15,75	78,5

ora 25

[Vedi commento al file xlsx]

RELAZIONE TRA ANOVA A 1 VIA E A 2 VIE

- Analogie:

→ ANOVA 1 via generalizza il test per il confronto delle medie di campioni normali omoschedastici a $n \geq 2$ campioni

→ ANOVA 2 vie generalizza il test t-paired (vedi Cap 8 Ross) a $n \geq 2$ campioni; inoltre lo fa sia sulle righe, sia sulle colonne

- Ambiti di applicazione:

→ Se i campioni hanno lunghezze diverse o se non c'è corrispondenza tra le righe di diverse colonne si può fare solo l'ANOVA a 1 via

→ In caso contrario, capita spesso che uno dei due effetti sia evidente a priori che sarà forte: può sembrare inutile verificarlo esplicitamente e si ha la tentazione di fare ANOVA a 1 via sull'altro effetto.
In realtà questo è **sbagliatissimo**: proprio in questi casi la ANOVA a 2 VIE risulta molto più potente dell'ANOVA a 1 via nel cercare l'altro effetto (quello non ovvio), perché, tenendo conto che le righe (ad esempio) sono diverse riesce a schermare l'effetto e "vedere" un rumore minore (2 vs 14) sull'effetto delle colonne

→ Se non è evidente se le due variabili hanno effetti significativi chiaramente uso l'ANOVA a 2 vie.

a) Se entrambi i test dicono H_1, fine
b) Se un solo test dice H_1, fine
c) Se entrambi i test dicono H_0: in questo caso può darsi (è frequente) che le due ANOVA a 1 via siano più potenti.

rifare con ANOVA 1 via

ANOVA A DUE VIE CON INTERAZIONI

Per ogni cella delle $m \times n$ combinazioni dei due fattori serve un campione di $\ell \geq 2$ dati

$X_{i,j,k} \sim \mathcal{N}(\mu_{i,j}, \sigma^2)$ $\quad i = 1, \dots, m$ (righe) $\quad j = 1, \dots, n$ (colonne) $\quad k = 1, \dots, \ell$ (campione)

- ipotesi di omoschedasticità
- le medie non dipendono da k

$\mu_{i,j} = \mu + \alpha_i + \beta_j \Rightarrow$ $\mu_{i,j}$ qualsiasi $\Leftrightarrow \mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$

ANOVA 2 VIE — ANOVA 2 VIE CON INT.

è sempre possibile scrivere $\mu_{i,j}$ come nell'espressione a destra e in modo tale che $\sum_i \alpha_i = 0 = \sum_j \beta_j = \sum_{i,j} \gamma_{i,j}$

- $\gamma_{i,j}$ si chiama termine di interazione
- Test che si fanno:

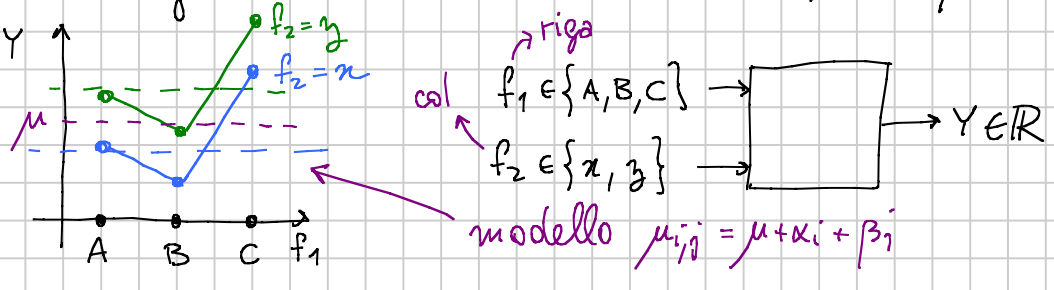
a) $H_0: \alpha_i \equiv 0 \ \forall i$ $\quad H_1$: non tutti nulli — test sull'effetto riga

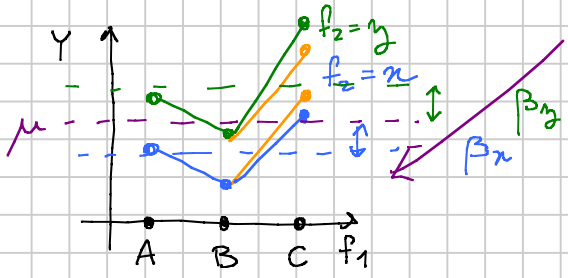
b) $H_0: \beta_j \equiv 0 \ \forall j$ $\quad H_1$: non tutti nulli — test sull'effetto colonna

c) $H_0: \gamma_{i,j} \equiv 0 \ \forall i,j$ $\quad H_1$: non tutti nulli — test sulle interazioni

→ Se $H_0^{c)}$ allora il modello additivo è adeguato e si può usare la ANOVA a 2 VIE (+ potente, + robusto)

→ Se $H_1^{c)}$ allora vi sono interazioni non nulle; il modello additivo non è adeguato; la presenza/assenza di effetti riga e colonna globali vanno lette su a) e b) di questa ANOVA





modello con interazioni $\mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$

i punti arancio sono i previsti di $\mu + \alpha_i + \beta_j$

i termini di interazione $\gamma_{c,x}$ e $\gamma_{c,y}$ dicono di quanto il modello osservato si discosta da quello semplificato additivo (arancione)

TEST DI ADATTAMENTO (Cap 11 Ross)

1) test del chi-quadro e derivati
2) test di Kolmogorov-Smirnov e parenti (leggere sul Ross)

TEST DEL CHI-QUADRO ELEMENTARE

esempio: v.a. discreta con una legge teorica nota

S: scolarizzazione $\quad S \in \{$ laurea, maturità, III media, meno $\}$

dati Istat (distribuzione nota) $\rightarrow$ 4% | 75% | 7% | 14%

Prendo un campione particolare: residenti a Treviso

$n = 100$ | 2 | 61 | 20 | 17

Q: questo campione può sensatamente (H_0) venire da quella distribuzione o c'è evidenza (H_1) che la sua distrib sia diversa?

→ Il test del χ^2 elementare funziona su una v.a. X che assume solo i valori $\{1, 2, .., n\}$ dove n è finito, piccolo e si suppone una certa legge φ_0

$H_0: P(X=1) = \varphi_0(1), P(X=2) = \varphi_0(2), \ldots, P(X=n) = \varphi_0(n)$

H_1: non tutte uguali

→ Si chiama test di adattamento a distribuzione discreta completamente specificata

- In pratica

id	sex	età	--	S
1				III m la
2				—
⋮				III m mat
100				⋮

n numero di dati

conto

k numero di categorie

1 : laurea
2 : mat.
3 : III m
4 : —

$O_1 = 2$
$O_2 = 61$
$O_3 = 20$
$O_4 = 17$

osservati

φ_0 nota

moltiplico per n

$A_1 = 4$
$A_2 = 75$
$A_3 = 7$
$A_4 = 14$

attesi (normalmente non sono interi)

$$W := \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i}$$ statistica del test

→ W è grande sotto H_1 e piccola sotto H_0

→ Sotto H_0 W ha una distribuzione calcolabile esattamente con metodi sofisticati (numerici o montecarlo) (in lab?)

→ Fortunatamente, di solito, sotto H_0 : $W \dot{\sim} \chi^2(k-1)$

ipotesi perché la approssimazione sia buona:

- Occorre che gli attesi siano grandi:
  ↳ "rule of thumb": tutti $A_i \geq 1$ e i $4/5$ degli $A_i \geq 5$

MMSD

ora 27

Titolo nota 15/12/2009

6 CFU → 8 h teoria
↳ 10 h pratica

1 CFU → 25 h

20/10 = 2 CFU
32/8 = 4 CFU } 6 CFU
52 + 98 h di lavoro a caso

lun 21 dic 16.30-18.30
Nicolodi + Morandin lab 2+3
(da confermare)

TEST DEL CHI QUADRO

$A_i = n \cdot p_i$ affinché l'approssimazione sia buona bisogna che i p_i non siano piccoli

→ in particolare se k è grande, questo non è possibile
↳ si può risolvere con metodi MC
↳ oppure raggruppando i valori:

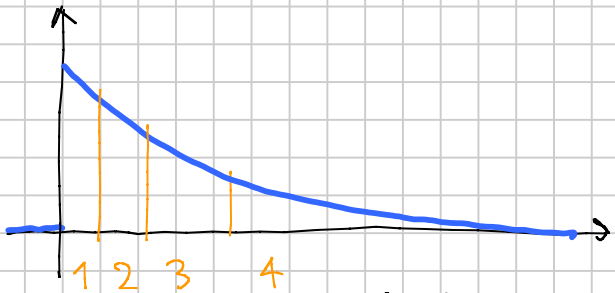
- Quando X ha tanti valori possibili o è continua

→ se X è discreta raggruppo valori simili in modo da ridurre il numero elevato di valori con prob piccole ad un numero basso di gruppi (bin) con prob non piccole

Vogliamo testare per X una distribuzione di Poiss

0	4,53999E-05	p0	
1	0,000453999	p1	
2	0,002269996	p2	
3	0,007566655		
4	0,018916637		50
5	0,037833275		
6	0,063055458	0,130141	6,507071
7	0,090079226	0,090079	4,503961
8	0,112599032	0,112599	5,629952
9	0,125110036	0,12511	6,255502
10	0,125110036	0,12511	6,255502
11	0,113736396	0,113736	5,68682
12	0,09478033	0,09478	4,739017
13	0,072907946	0,208444	10,42218
14	0,052077104	8 gruppi di valori poss	
15	0,03471807		
16	0,021698794		
17	0,012763996		

→ se X è continua, uguale legge esponenziale di media 10



4 bin di prob elevata

☆ Principio generale: se i bin sono più o meno equiprobabili l'approx è migliore e il test è più potente

→ Se X è continua è possibile (e raccomandato) ottenere bin esattamente equiprobabili

→ voglio k bin equip. della distrib continua con pdf f e CdF F

$c_1, c_2, \dots, c_{k-1}$ punti di taglio sono tali che:

$$\frac{1}{k} = \int_{-\infty}^{c_1} f(t)dt = \int_{c_1}^{c_2} f(t)dt = \dots = \int_{c_{k-1}}^{\infty} f(t)dt$$

$$= F(c_1) \quad = F(c_2) - F(c_1) \quad \dots \quad 1 - F(c_{k-1})$$

$$\Rightarrow F(c_1) = 1/k, \quad F(c_2) = 2/k, \quad \dots \quad F(c_{k-1}) = \frac{k-1}{k}$$

$$\Rightarrow c_i = F^{-1}\left(\frac{i}{k}\right)$$

Vogliamo testare per X una distribuzione esponenziale di media 10

	7 k
i	ci
1	1,541507
2	3,364722
3	5,596158
4	8,472979
5	12,52763
6	19,4591

o_i , frequenza (array; ...)

★ Gli attesi sono sempre n per la prob di ciascun bin
Gli osservati sono il numero di dati che cadono in ogni bin

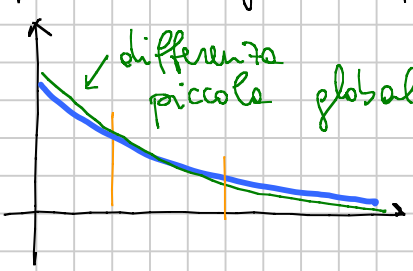
● Quanti bin bisogna fare:

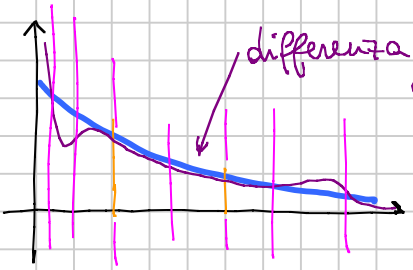
i. maggiore n, maggiore k

ii. k mai troppo grande (gli attesi devono rispettare la "rule of thumb")

iii. più k è piccolo più il test è potente (differenze piccole)

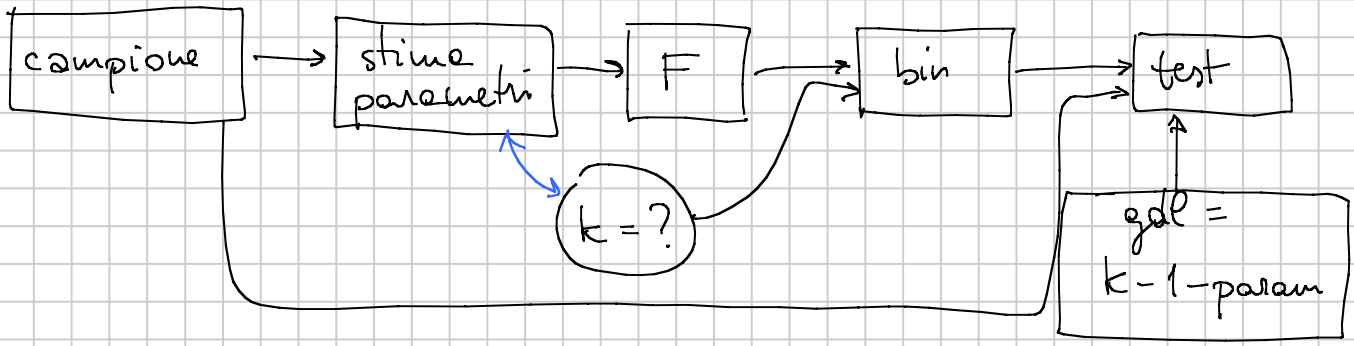
iv. più k è grande più il test è sensibile (differenze locali)





Test di adattamento a distribuzione con parametri incogniti

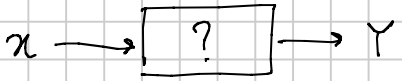
→ Si stimano puntualmente i parametri incogniti sui dati e poi si esegue il test usando quelli. In questo modo il campione si adatta però artificialmente bene alla distribuzione. Per correggere si devono calare i gdl della chi-quadro del numero di parametri stimati



ora 28

TABELLE DI CONTINGENZA

Primo scopo: vedere se due variabili sono indipendenti



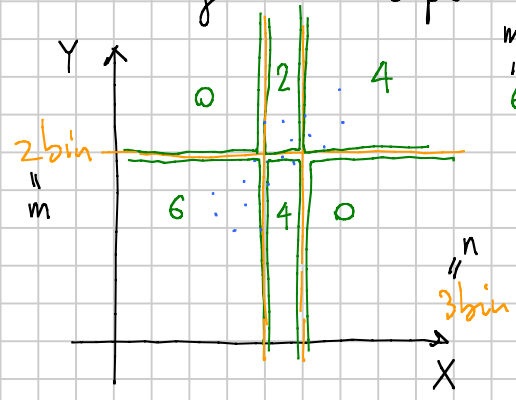
regressione: x, Y numeriche

anova: x categorica, Y numerica

t d cont: x, Y qualunque (ma è poco potente, quindi dice spesso H_0)

$X \longleftrightarrow Y$ il legame di dipendenza cercato è simmetrico

id	X	Y
1	~	~
2	~	~
⋮	~	~
N	~	~



caso entrambe numeriche

Y \ X	1	2	3	
1	0	2	4	6
2	6	4	0	10
	6	6	4	16

Y \ X	1	2	3	
1	0	2	4	6
2	6	4	0	10
	6	6	4	16

H_0: le due var sono indipendenti

H_1: no

sotto H_0, ipotizzando che i totali verdi siano dati, che valori mi aspetterei nelle 6 celle?

2,25	2,25	1,5	6
3,75	3,75	2,5	10
6	6	4	16

$\frac{6}{16} \cdot \frac{6}{16} = \frac{36}{16^2}$ del totale $\quad \frac{36}{16} = 2{,}25$ attesi

$\frac{6}{16}$ della riga da 10 $= \frac{6}{16} \cdot 10 = \frac{15}{4} = 3{,}75$

Questi sono gli attesi A_i che si avrebbero se ci fosse indipendenza

0	2	4
6	4	0

gli osservati

$$W = \sum_{i=1}^{6} \frac{(A_i - O_i)^2}{A_i} = \frac{(0-2{,}25)^2}{2{,}25} + \frac{(2-2{,}25)^2}{2{,}25} + \frac{(4-1{,}5)^2}{1{,}5} + \frac{(6-3{,}75)^2}{3{,}75} + \frac{(4-3{,}75)^2}{3{,}75} + \frac{(0-2{,}5)^2}{2{,}5}$$

la statistica richiede la solita "rule" sugli A_i, che qui è false

sotto H_0 : $W \dot{\sim} \chi^2(gdl)$

$gdl = (m-1)(n-1)$ dove m e n sono righe e colonne

★ Nella scelta dei numeri di bin m, n occorre ricordare che i bin totali $m \cdot n$ devono essere adeguati alla numerosità N del campione

→ Se una variabile è numerica, converrà "tagliare" i bin ai percentili campionari $= \text{PERCENTILE}(\text{array}_x; 1/3)$

$= \text{PERCENTILE}(\text{array}_x; 2/3)$

→ Se una var è categorica, converrà raggruppare le categorie fino ad ottenere il numero voluto di bin, che devono risultare ± equifrequenti

stato famiglia	#
celibe	64
convivente	112
convivente	45
divorziato	38
separato	35
vedovo	44
	338

3 bin ~ 113 l'uno

cel + conv	109
coniug	112
altro	117

ora 29

TUTTI GLI SCOPI DELLE TAB. DI CONT.

1) Vedere se H_0: X, Y sono indipendenti H_1: no

2) Vedere se: H_0: no H_1: Y dipende da X
(= X dipende da Y)

★ alternativa non-parametrica a regressione semplice ($p=1$) e ANOVA a 1 via
↳ non ha ipoteri né di omoschedasticità, né di gaussianità, né di linearità.
↳ è meno potente delle tecniche parametriche corrispondenti

3) Vedere se m campioni hanno la stessa distribuzione

$X_{1,1}\ X_{1,2} \dots X_{1,n_1} \sim D_1$

$X_{2,1} \dots X_{2,n_2} \sim D_2$

...

$X_{m,1} \dots X_{m,n_m} \sim D_m$

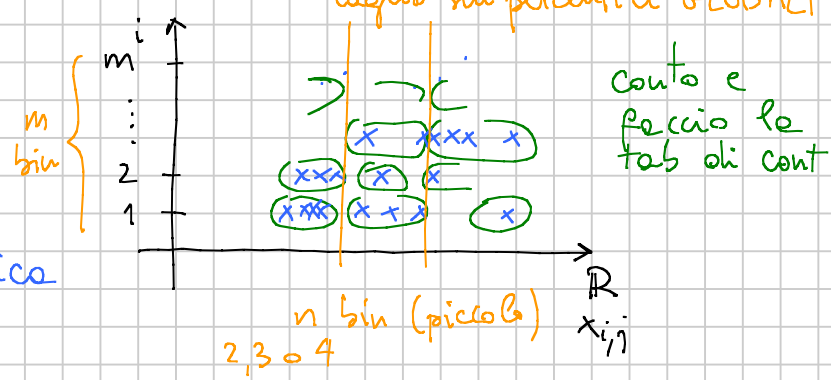
$H_0: D_1 = D_2 = \dots = D_m$

H_1: no

$X_{i,j}$ dipende da i?

$X_{i,j}$: numerica ma non necessariamente

i: categorica



★ Ricordiamo che il test del chi-quadro standard permette similmente di confrontare $D_1 \sim x_1 \dots x_n$ con una distribuzione assegnata

comp 1:	37	24	17	12	90
comp 2:	63	51	32	20	166

test, 3 gdl
hanno la stessa distrib?

distrib:	0,4	0,3	0,2	0,1	
comp 1:	37	24	17	12	90
atteso:	36	27	18	10	90

test, 3 gdl
hanno la stessa distrib?

4) Nel caso una delle variabili sia dicotomica, vedere se

H_0: la prob di successo non dipende dall'altra variabile

H_1: dipende

	ER	H	GR	...
non ces.	74	45	~	~
cesarei	22	18	~	~

osservati

	ok	ko	
farmaco	26	14	40
placebo	19	21	40
	45	35	80

SUI TEST STATISTICI

(2008 da ore 13.13)

→ parametro incognito all'interno di classe di distribuzioni nota (ad es μ per $\mathcal{N}(\mu, \sigma^2)$)

→ due ipotesi contrapposte (ad es: $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$; $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$) spesso con un valore target di confronto (ad es: μ_0) detto AQL acceptable quality level

→ stimatore parametro $(\bar{X} \approx \mu)$ a volte una statistica $\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right)$

- ESEMPIO : misuro 50 pezzi $\bar{X} = 2{,}758$ il target dovrei essere

$\mu_0 = 2{,}735$

la media con cui sto producendo questi pezzi (μ) è in target ?

H_0 : in target $\mu = \mu_0$

H_1 : fuori target $\mu \neq \mu_0$

→ test dell'idiota :

se $\bar{X}$ = target, ok se no regolo la macchina

dico sempre H_1 e mai H_0

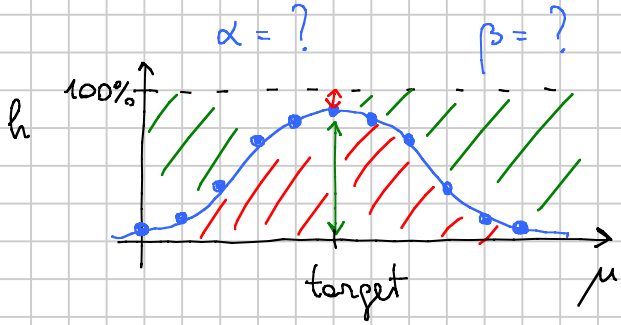
$\alpha = 100\%$ $\beta = 0\%$

→ test del dilettante : ci vuole un po' di tolleranza

R.A.

se $\bar{X} \in$ target $\pm 0{,}010$, ok se no regolo

↑ scelta a occhio



h misura la prob di dire H_0

$h = h(\mu)$

funzione o curva operativa caratteristica

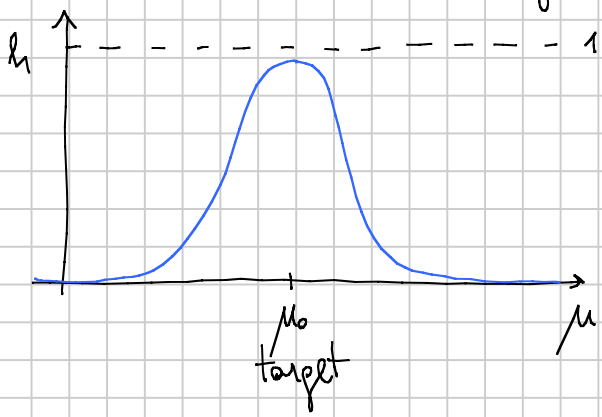
quando è vera H_0, sono in target, dico H_1 con prob $1 - h(\mu_0)$

$\alpha = 1 - h(\mu_0)$ ↕

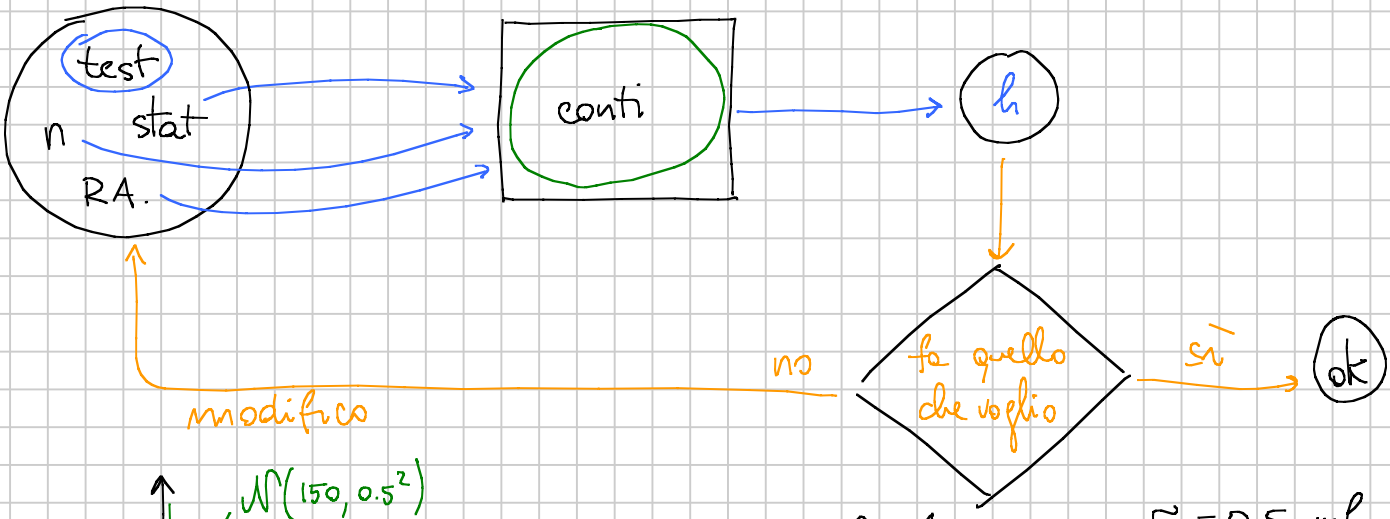
in tutti gli altri casi è vera H_1

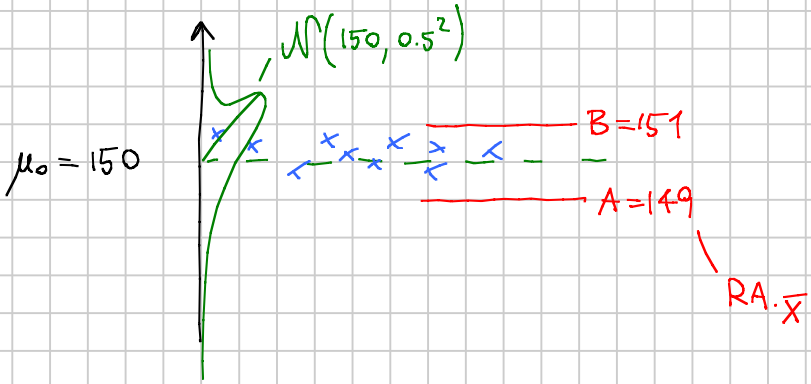
$\beta = h(\mu)$ cioè β dipende da μ

Il "professionista" è in grado di calcolare la COC.



$h(\mu)$: prob di dire H_0 quando il parametro vale μ





$n = 4$ $\quad \sigma = 0{,}5$ ml

$\bar{X} = \frac{1}{4}\sum_{i=1}^{4} x_i$ $\quad \mu_0 = 150$ ml

$RA_{\bar{X}} = [149; 151]$

$h = h(\mu) :=$ "prob di dire H_0" $= P(\bar{X} \in [A, B]) = P(A \le \bar{X} \le B)$

$\bar{X} \sim ?$ $\quad F_{\bar{X}} = ?$

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad F_{\bar{X}}(t) = \text{DISTRIB.NORM}\left(t; \mu; \frac{\sigma}{\sqrt{n}}\right)$

$h(\mu) = F_{\bar{X}}(B) - F_{\bar{X}}(A) = \text{DISTRIB.NORM}\left(B; \mu; \frac{\sigma}{\sqrt{n}}\right) - \text{DISTRIB.NORM}\left(A; \mu; \frac{\sigma}{\sqrt{n}}\right)$

$A = \mu_0 - r$
$B = \mu_0 + r$
} r diventa l'unica incognita

$r \longleftrightarrow \square \longleftrightarrow \alpha$

$$\alpha = 1 - h(\mu_0) = 1 - \left(F_{\bar{X}}(\mu_0 + r) - F_{\bar{X}}(\mu_0 - r)\right) = 1 - \Phi\left(\frac{\mu_0 + r - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{\mu_0 - r - \mu_0}{\sigma/\sqrt{n}}\right)$$

$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ $\quad F_{\bar{X}}(t) = \Phi\left(\frac{t - \mu_0}{\sigma/\sqrt{n}}\right)$

$$= 1 - \Phi\left(\frac{r\sqrt{n}}{\sigma}\right) + \Phi\left(-\frac{r\sqrt{n}}{\sigma}\right) = 2\left(1 - \Phi\left(\frac{r\sqrt{n}}{\sigma}\right)\right)$$

$$\frac{\alpha}{2} = 1 - \Phi\left(\frac{r\sqrt{n}}{\sigma}\right)$$

$$1 - \frac{\alpha}{2} = \Phi\left(\frac{r\sqrt{n}}{\sigma}\right) \iff \frac{r\sqrt{n}}{\sigma} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \iff r = \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

- Esempio 2 (unilaterale): mtst0605.1

p : percentuale difettosi nelle forniture
↳ campione di n pezzi

X : numero di difettosi sul campione

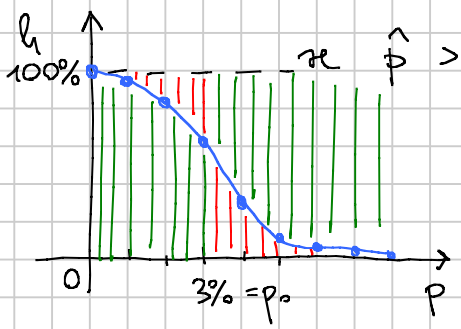
$\hat{p} = \frac{X}{n} \approx p$ frazione difettosi sul campione

$p < 3\%$? $\quad p > 3\%$?

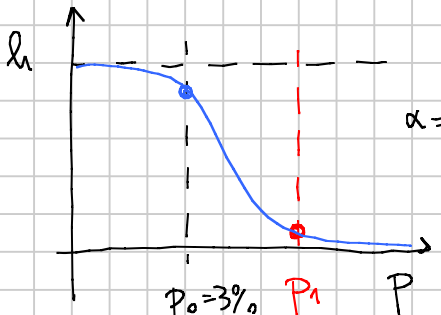
→ test dell'idiota
se $\hat{p} < 3\% \rightarrow p < 3\%$ tutto ok
se $\hat{p} > 3\% \rightarrow p > 3\%$ protesto col fornitore
(o rifiuto i pezzi)

→ test del dilettante: un po' di tolleranza
se $\hat{p} \leq 3\% + 1{,}5\% \rightarrow p \leq 3\%$ tutto ok (H_0)
↑ margine arbitrario

se $\hat{p} > 3\% + 1{,}5\% \rightarrow p > 3\%$ protesto (H_1)



$h = h(p)$: "prob di dire H_0"



α: "prob err di I specie" $= 1 - h(p)$ $\quad p \le 0{,}03$

$\alpha = \alpha(p) \le \bar{\alpha} := 1 - h(p_0) = 1 - h(3\%)$

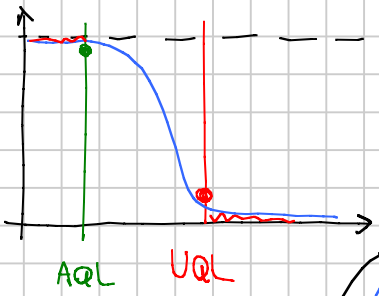
livello di significatività : max err I sp.

β: "prob err di II specie" $= h(p)$ $\quad p > 0{,}03$

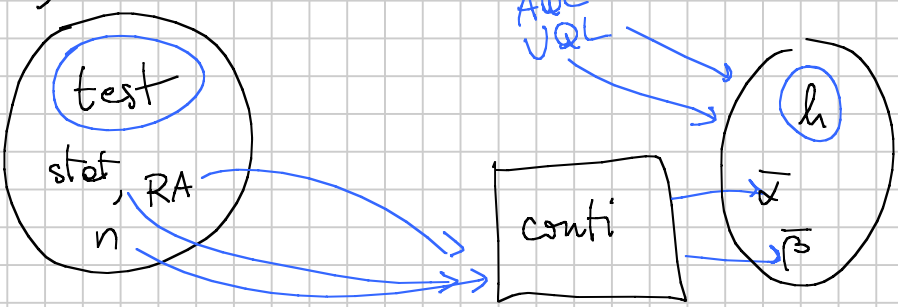
$\beta = \beta(p) \le h(p_0) = 1 - \bar{\alpha}$

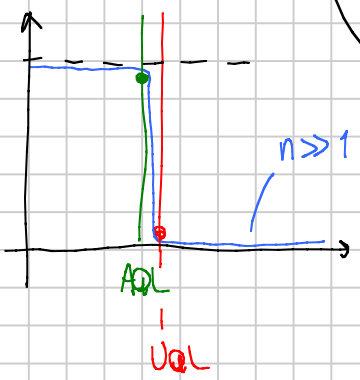
se $p > p_1$ allora $\beta = \beta(p) \le \bar{\beta} = h(p_1)$

★ La definizione dell'UQL simmetrizza il test : scambiando le ipotesi e il ruolo di $\bar{\alpha}$ e $\bar{\beta}$ si ottiene il medesimo risultato



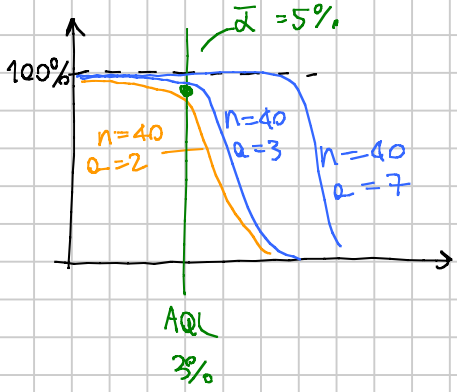
Si fissano UQL e AQL, $\bar{\alpha}$ e $\bar{\beta}$ e si ottiene un test con quelle caratteristiche





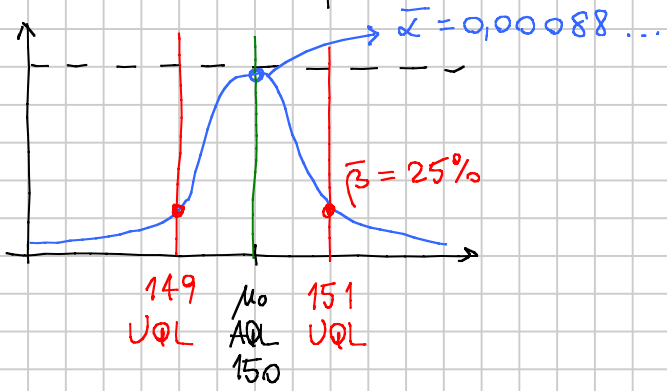
$h(p) = P(X \le a) = F_X(a) = \text{DISTRIB.BINOM}(a; n; p; 1)$

$\text{bin}(n,p)$



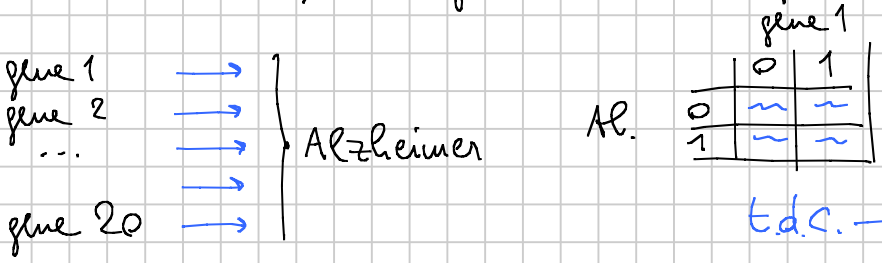
★ Può essere utile definire UQL anche nei casi bilaterali

Nel primo esempio:



CORREZIONE DI BONFERRONI

Se in un ambito singolo sono costretto a fare più test, le probabilità dei falsi positivi (α, err I specie) si sommano e quindi la frequenza globale di falsi positivi non è più piccola



t.d.c. → gene 1 è legato a Al.

$\alpha = 5\%$ per ciascun test ; ipotizziamo vere tutte H_0

ciascun test dice H_1 con prob 5% → la prob che dicano tutti H_0 è $0{,}95^{20} \approx 36\%$ → 64% di prob che almeno un gene risulti significativo pur non essendolo

★ Questo test : (H_0 : nessuno di questi geni è legato alla malattia vs H_1 : qualcuno sì) fatto così, ha livello di significatività del 64% INACCETTABILE

★ Correzione di Bonferroni : se devo fare n test e voglio globalmente lvl di sign. $\bar{\alpha}$, ciascuno di questi test deve essere eseguito con lvl di sign. $\bar{\alpha}/n$.

→ Ne escono test poco potenti

TEST DI FISHER-IRWIN (Cap 8)

tabelle di contingenza 2×2 quando i numeri sono piccoli per il χ^2

Al.

gene 1	0	1	
0	6	7	13
1	1	14	15
	7	21	28

3,25	9,75
3,75	11,25

gli attesi hanno 2 numeri < 5
non uso il chi-quadro

Mettiamoci in H_0: le due var sono indipendenti

Y		13
		15
7	21	28

è equivalente immaginare un'urna con 28 palline, di cui 15 nere (malattia) e 13 bianche che vengono pescate a caso da 28 persone: 7 con gene 0 e 21 con gene 1; alla fine l'associazione tra i due caratteri è indipendente.

Sia Y il numero di gene 0 - sani: che distribuzione ha Y?

$Y \sim \text{hypgeom}(7; 13; 28)$ ipergeometrica

1	12	13
6	9	15
7	21	28

Sia y il valore osservato per Y

$$\alpha^* = 2 \cdot \min\left(F_Y(y); 1 - F_Y(y-1)\right)$$

SULLA GENERAZIONE DI V.V.A.A.

= CASUALE() genera unif(0,1)

= F^{-1}(CASUALE()) genera una v.a con CdF F

↳ =INV.NORM(CASUALE(); media; devst)

↳ =INV.GAMMA(CASUALE(); α o 1; β)

...

↳ = CRIT.BINOM(CASUALE(); n; p)

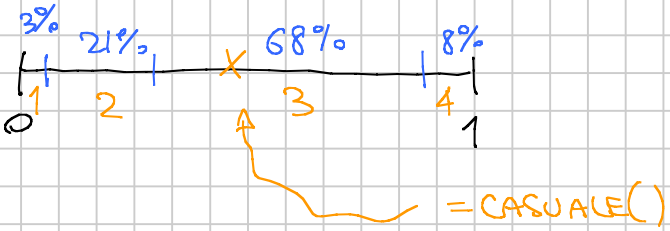
★ Poisson: non c'è INV.POIS Pois(ν) ν la media

= CRIT.BINOM(CASUALE(); 1000; ν/1000)

★ Categorica generica $n = 50$

1	no	3%	2
2	medie	21%	11
3	maturità	68%	30
4	laurea	8%	7

$X_1, X_2, \dots, X_{50} \rightarrow \{1, 2, 3, 4\}$



=CASUALE() [A1]

=SE(A1 < 0,03 ; 1 ; SE(A1 < 0,03+0,21 ; 2 ; SE(A1 < 0,03+0,21+0,68 ; 3 ; 4)))

★ Si ripete su 50 celle e si ottiene il campione X_i

★ Si usa =CONTA.SE() per contare quante di ogni tipo nel campione