METODI E MODELLI A SUPPORTO DELLE DECISIONI (B) 01

Note Title 02/03/2011

Nicolodi + Morandin
↳ statistica 6 CFU

A Design of experiment
(Sleeper - Design for Six Sigma Statistics - 59 tools for ...)
Capitolo 10

★ Middleton (Apogeo)
"Analisi statistiche con Excel"

★ Altre fonti:
- www.statsoft.com "statistics textbooks"
- dispense di Soliani ~3000 pagine

Modalità di esame da definire

Materiale → lea.unipr.it
scritti, lezioni (avi/pdf), forum, avvisi, ...
↳ anche vecchie, soprattutto lab

Ricevimento: gio 13-14.30

10 laboratori di 2 ore
32 ore di lezione

★ Green belt, black belt, master black belt

Ross:
3 — 8 già fatti
7 intervalli di confidenza
8 test statistici ★
9 regressione
10 analisi della varianza
11 i test del chi-quadro

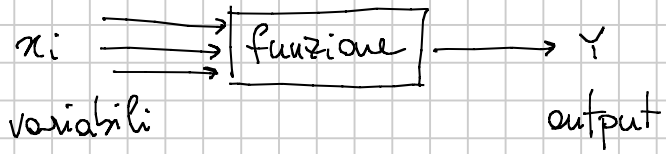
★ Più curve OC e simulazione MC

REGRESSIONE

- R. lineare semplice
- R. lineare multipla
- R. nonlineare
- R. ... eteroschedastica

} omoschedastiche

+ DOE



dati ← esperimenti → dati

↓

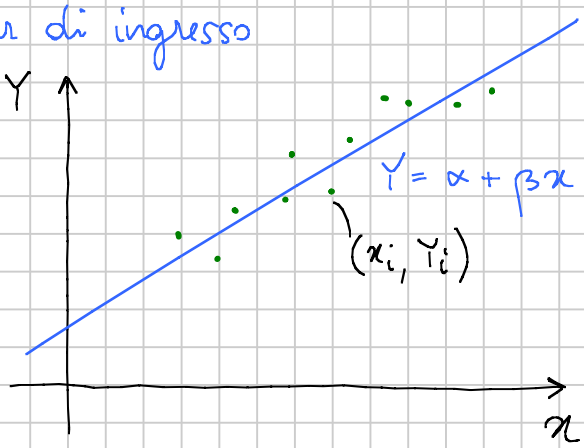
regressione → descrizione funzione

misura del rumore

↳ ne tiene conto

REGRESSIONE LINEARE SEMPLICE

↑ funzione (retta) ↑ 1 var di ingresso



(x_i, Y_i) $i = 1, 2, \dots, n$

i dati su cui lavorare

per trovare la retta, ovvero α, β

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

↳ errore

$Y_1 = \alpha + \beta x_1 + e_1$

$Y_2 = \alpha + \beta x_2 + e_2$

. . . .

→ supponiamo che $e_1, \dots, e_n$ siano

- $\mathcal{N}(0, \sigma^2)$ ← σ non dipende da $i / x_i / Y_i$
- indipendenti

OMOSCHEDASTICITÀ

★ $Y_i = \alpha + \beta x_i + e_i \sim \mathcal{N}(\alpha + \beta x_i ; \sigma^2)$

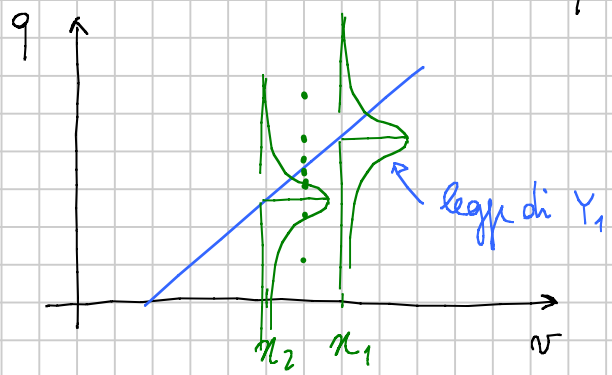
$Y_1, Y_2, \dots, Y_n$ sono indipendenti ma non con la stessa legge

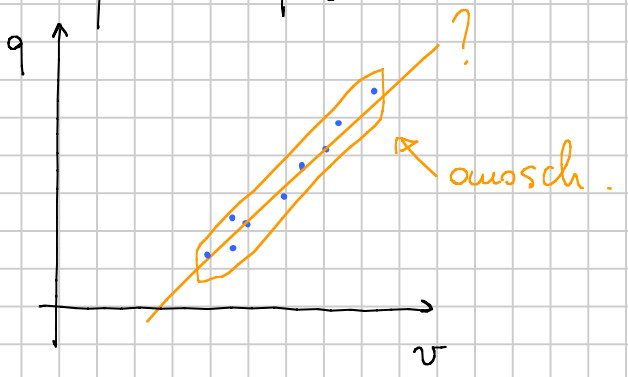
(tranne quando $\beta = 0$ → si dice che non c'è regressione)

↳ quindi Y_i iid e si usano tecniche elementari

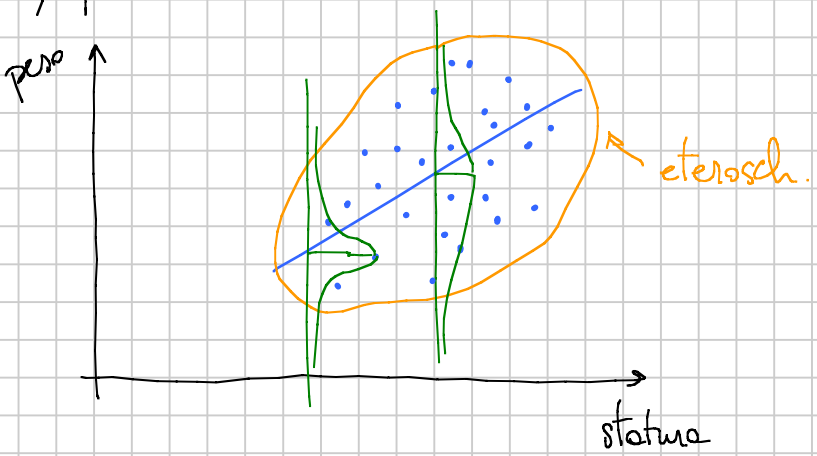
Esempi

1) Ripartitore polveri - velocità v della cocka
- quantità q della polvere

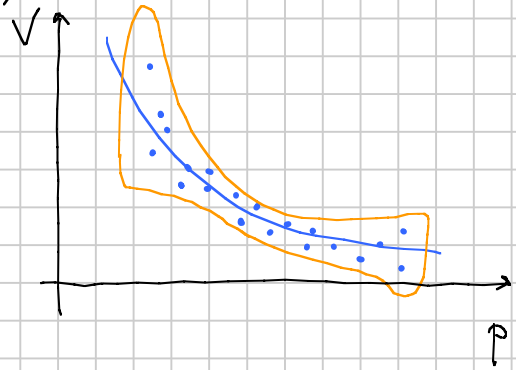




2) peso vs statura adulti italiani maschi



3) barometro elettronico - tensione V - pressione p



due variabili legate → c'è regressione → le var sono correlate

→ rapporto di causa / effetto

★ È possibile scambiare x e Y $\quad Y = \alpha + \beta x \iff x = -\frac{\alpha}{\beta} + \frac{1}{\beta} Y$

NB. Se si fanno le due regressioni si ottengono risultati (un po') diversi

★ Come scegliere?

- x causa, Y effetto
- x deterministica/precisa, Y casuale/rumorosa
- Y deterministica + rumore canale

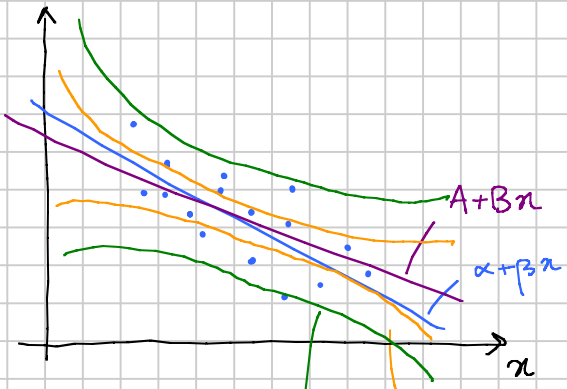
● Pratica:

$Y \sim \mathcal{N}(\alpha+\beta x ; \sigma^2)$ $\quad \alpha, \beta, \sigma$ incognite da stimare

$\downarrow \quad \downarrow \quad \downarrow$

$A \quad B \quad Se$ stimatori

$(x_i, Y_i) \quad i=1,\dots,n$



intervallo di confidenza per la risposta media

$\alpha+\beta x \in A+Bx \pm \text{raggio}(x)$

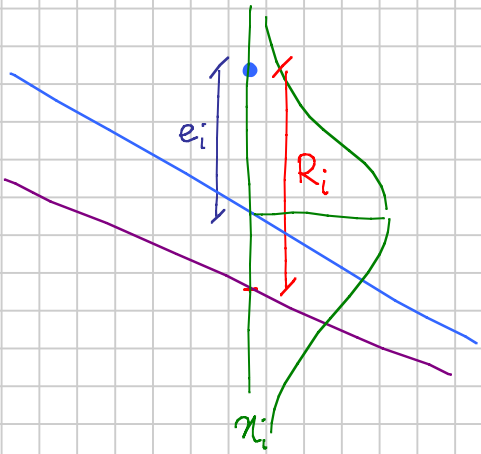
intervallo di predizione per una risposta futura

$\alpha+\beta x+e \in A+Bx \pm \text{raggio\_maggiore}(x)$

● Come si trova la retta viola?

Si minimizza la somma dei quadrati dei residui

★ Residui o scarti sono l'errore misurato relativamente a $A+Bx$



$R_i := Y_i - (A+Bx_i)$

può essere positivo o negativo

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

★ $SS_R = SS_R(A;B)$ funzione di A, B

HW: Trovare A, B che minimizzano SS_R (in funzione di x_i, Y_i)

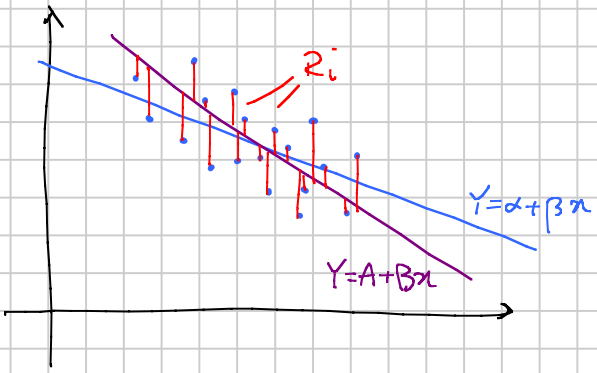
MMSD (B) — ora 3

Note Title 03/03/2011

REGR. LIN. SEMPLICE

$$Y_i = \alpha + \beta x_i + e_i \qquad (x_i, Y_i),\ i = 1, 2, \dots, n \qquad e_i \sim \mathcal{N}(0, \sigma^2)$$

α, β, σ parametri incogniti



stimatori (random) → $A \approx \alpha$, $B \approx \beta$, $S_e \approx \sigma$

formule — distribuzioni

Troviamo A, B che rendono minima la somma dei quadrati

$$SS_R(A,B) = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

$$0 = \frac{\partial SS_R}{\partial A} = \sum_{i=1}^{n} 2(Y_i - A - Bx_i)\cdot(-1)$$

$$0 = \frac{\partial SS_R}{\partial B} = \sum_{i=1}^{n} 2(Y_i - A - Bx_i)\cdot(-x_i)$$

$$\begin{cases} \sum Y_i = \sum (A + Bx_i) \\ \sum x_i Y_i = \sum (A + Bx_i) x_i \end{cases}$$

$$\begin{cases} \sum Y_i = nA + B\sum x_i \\ \dots \end{cases} \qquad \begin{cases} \bar{Y} = A + B\bar{x} \\ \overline{xY} = A\bar{x} + B\overline{x^2} \end{cases}$$

sistema delle equazioni normali

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

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

- Stimatore di σ

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

$\frac{1}{n}\sum (Y_i - \alpha - \beta x_i)^2 \approx \sigma^2$ ma non ho α, β

$$S_e^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - A - Bx_i)^2 = \frac{SS_R}{n-2}$$

errore standard

★ $\bar{Y} = A + B\bar{x}$: il punto $(\bar{x}, \bar{Y})$ sta sulla retta viola

↑ baricentro

★ $B = \dfrac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} = \dfrac{S_{xY}}{S_x^2}$ ← covarianza campionaria

$$S_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - n\bar{x}^2\right]$$

$$S_{xY} = \frac{1}{n-1}\left[\sum x_i Y_i - n\bar{x}\bar{Y}\right]$$

HW: cosa succede (e perché) se il den. è zero?

★ $Y_i - \alpha - \beta x_i \sim \mathcal{N}(0, \sigma^2)$ indipendenti

$R_i = Y_i - A - Bx_i$ NO (non hanno Var $= \sigma^2$ e non sono indip.)

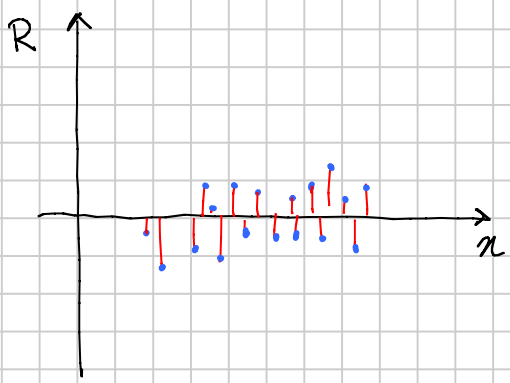
$$"S_R^2" = \frac{1}{n-1}\left[\sum R_i^2 - n\bar{R}^2\right] = \frac{n-2}{n-1} S_e^2$$

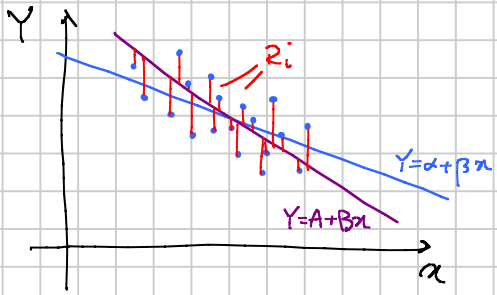
↑ HW

stimatore di σ^2 che sottostima sistematicamente

★ Per come sono definiti i residui:

- $\bar{R} = 0$
- se faccio la regressione sui punti (x_i, R_i) trovo l'asse delle ascisse





DISTRIBUZIONE DI A E B

$$B = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} = \frac{\frac{1}{n}\sum x_i Y_i - \bar{x}\frac{1}{n}\sum Y_i}{\overline{x^2} - \bar{x}^2} = \frac{1}{n(\overline{x^2} - \bar{x}^2)}\left[\sum x_i Y_i - \bar{x}\sum Y_i\right]$$

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

i. B comb lineare di $Y_i \Rightarrow B$ è $\mathcal{N}$

ii. $E(B) = \beta$ (stimatore corretto)

iii. $\mathrm{Var}(B) = \sigma^2 k_B$ (stimatore consistente)

$$E(X+Y) = E(X) + E(Y)$$
$$E(aX+bY) = aE(X) + bE(Y)$$

$$E(B) = E\left(\sum c_i Y_i\right) = \sum_{i=1}^{n} c_i E(Y_i)$$

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

$$= \sum c_i(\alpha + \beta x_i) = \frac{1}{n(\quad)}\left[\underbrace{\sum (x_i - \bar{x})\alpha}_{=0} + \sum (x_i - \bar{x})\beta x_i\right]$$

$\frac{1}{n}\sum x_i = \bar{x}$

$\sum x_i = n\bar{x}$

$\sum \bar{x} = n\bar{x}$

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

$$\mathrm{Var}(X) = \mathrm{Cov}(X ; X)$$
$$\mathrm{Cov}\left(\sum a_i x_i ; \sum b_j Y_j\right) = \sum_i \sum_j a_i b_j \mathrm{Cov}(x_i ; Y_j)$$
$$X, Y \text{ indipendenti} \Rightarrow \mathrm{Cov}(x ; Y) = 0$$

ord 4

$$\mathrm{Var}(B) = \mathrm{Cov}(B ; B) = \mathrm{Cov}\left(\sum_{i=1}^{n} c_i Y_i ; \sum_{j=1}^{n} c_j Y_j\right) = \sum_i \sum_j c_i c_j \mathrm{Cov}(Y_i ; Y_j)$$

$$= \sum_{i=1}^{n} c_i^2 \mathrm{Cov}(Y_i ; Y_i) = \sum c_i^2 \mathrm{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 =: \sigma^2 k_B$$

$$k_B := \sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{n(\bar{x^2} - \bar{x}^2)} \right]^2 = \frac{1}{n^2(\ )^2} \sum (x_i - \bar{x})(x_i - \bar{x})$$

$$\frac{1}{n^2(\ )^2} \left[\underbrace{\sum (x_i - \bar{x}) x_i}_{= n(\bar{x^2} - \bar{x}^2) \text{ come sopra}} - \underbrace{\sum (x_i - \bar{x}) \bar{x}}_{= 0 \text{ come sopra}} \right] = \frac{1}{n(\bar{x^2} - \bar{x}^2)}$$

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

★ per $n \to \infty$ $k_B \to 0$

• Rifaccio per A

$$A = \bar{Y} - B\bar{x} = \bar{Y} - \sum_{i=1}^{n} c_i Y_i \bar{x} = \frac{1}{n} \sum_i Y_i - \sum_i \frac{x_i - \bar{x}}{n(\bar{x^2} - \bar{x}^2)} \bar{x}\, Y_i$$

$$= \sum_{i=1}^{n} Y_i \left\{ \frac{1}{n} - \frac{x_i \bar{x} - \bar{x}^2}{n(\bar{x^2} - \bar{x}^2)} \right\} = \sum_{i=1}^{n} Y_i \frac{\bar{x^2} - \bar{x}^2 - x_i \bar{x} + \bar{x}^2}{n(\bar{x^2} - \bar{x}^2)} =: \sum_{i=1}^{n} d_i Y_i$$

$$d_i := \frac{\bar{x^2} - x_i \bar{x}}{n(\bar{x^2} - \bar{x}^2)}$$

HW: verificare che:

i. $A \sim \mathcal{N}$

ii. $E(A) = \alpha$ (stimatore corretto)

iii. $Var(A) = \sigma^2 k_A$ (stimatore consistente)

dove $k_A = \dfrac{\bar{x^2}}{n(\bar{x^2} - \bar{x}^2)}$

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

NB A, B non sono indipendenti

HW: $Cov(A; B) = -\sigma^2 \dfrac{\bar{x}}{n(\bar{x^2} - \bar{x}^2)}$

$Var(A + Bx) = ?$

DISTRIBUZIONE DI S_e

$$S_e^2 := \frac{SS_R}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - A - Bx_i)^2$$

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

perché? Vedremo... per ora qualche analogia

a) $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.

$$\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1) \quad \text{i.i.d.}$$

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

$$\frac{n}{\sigma^2}\cdot\underbrace{\frac{1}{n}\sum(X_i-\mu)^2}_{S_*^2} = \frac{S_*^2}{\sigma^2}n \sim \chi^2(n)$$

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

b) $X_1 \dots X_n$, $Y_1 \dots Y_m$

$X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ iid $\qquad Y_j \sim \mathcal{N}(\mu_Y, \sigma^2)$ iid.

$$S_X^2 = \frac{1}{n-1}\sum(X_i-\bar{X})^2 \qquad S_Y^2 = \frac{1}{m-1}\sum(Y_i-\bar{Y})^2$$

$$S_X^2 \approx \sigma^2 \qquad S_Y^2 \approx \sigma^2$$

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

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

$$= \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{j=1}^{m}(Y_i-\bar{Y})^2\right]$$

$$\frac{S_p^2}{\sigma^2}(n+m-2) = \sum_i\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + \sum_j\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2(n+m-2)$$

$$\sum_i \left(\frac{X_i - \mu_X}{\sigma}\right)^2 + \sum_j \left(\frac{Y_i - \mu_Y}{\sigma}\right)^2 \sim \chi^2(n+m) \qquad \text{(HW)}$$

---

c) regressione

$$\frac{Y_i - \alpha - \beta x_i}{\sigma} \sim \mathcal{N}(0,1)$$

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

$$\frac{S_e^2}{\sigma^2}(n-2) = \frac{SS_R}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{Y_i - A - B x_i}{\sigma}\right)^2 \sim \chi^2(n-2)$$

* A, S_e indip B, S_e indip.

MMSD (B) ora 5

Note Title 09/03/2011

$S_e := \sqrt{\frac{SS_R}{n-2}}$ $\quad S_e^2(n-2) = SS_R$ $\quad \frac{S_e^2}{\sigma^2}(n-2) = \frac{SS_R}{\sigma^2} \sim \chi^2(n-2)$

$A \sim \mathcal{N}(\alpha, \sigma^2 k_A)$ $\quad B \sim \mathcal{N}(\beta \sim \sigma^2 k_B)$ $\quad A, S_e$ indip. $\quad B, S_e$ indip

INFERENZA SUI PARAMETRI DI REGRESSIONE

funzione ancillare → int. di conf.
→ test statistici

Def. Si dice funz. ancillare di un parametro θ una v.a. che dipende da: θ, il campione, altre grandezze note di cui sia nota esattamente la distribuzione

→ Esempio: $X_i \sim \mathcal{N}(\mu, \sigma^2)$ $i = 1, 2, \dots, n$ σ nota μ no

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

funz. anc. per μ?

$\bar{X}$ no perché la legge non è nota del tutto (μ)

$$\bar{X} - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n})$$

Questa (volendo) è una funz. anc. per μ

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Questa pure: è quella canonica

→ Esempio: stessa situazione, ma σ incognita

$\bar{X} - \mu$ no perché la legge non è nota del tutto (σ)

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ no, perché dipende anche da σ

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

campione — noto

perché sia funz. anc. di μ occorre conoscerne la distribuzione e che questa sia nota

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

inoltre S e $\bar{X}$ indipendenti

$$Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$W := \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1) \qquad k := n-1$$

indipendenti

$$Z\sqrt{\frac{k}{W}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \sqrt{\frac{n-1}{\frac{S^2}{\sigma^2}(n-1)}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S}$$

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

Cos'è la t di Student?

$Z \sim \mathcal{N}(0,1)$ $\quad W \sim \chi^2(k)$ indip

Allora $\dfrac{Z}{\sqrt{W/k}} \sim t(k)$

(per definizione)

- Torniamo alla regressione lineare semplice

– funz. ancillare per β e α:

$$B \sim \mathcal{N}(\beta, \sigma^2 k_B) \qquad Z := \frac{B-\beta}{\sigma\sqrt{k_B}} \sim \mathcal{N}(0,1)$$

$$W := \frac{S_e^2}{\sigma^2}(n-2) \sim \chi^2(n-2) \qquad k = n-2$$

indip.

HW: conto

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

funz. anc. per β

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

funz. anc. per α

HW: tutto

- funz. anc. per σ

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

- funz. ancillare per la retta azzurra : $\alpha + \beta x_0$

fissato un valore x_0 qualsiasi, $\alpha+\beta x_0$ è l'ordinata della retta per quel valore di ascissa

i. Stimatore : $\alpha + \beta x_0 \approx A + B x_0$

ii. Distribuzione $A + B x_0 \sim \mathcal{N}\left(\alpha+\beta x_0 \; ; \; \sigma^2 k_m(x_0)\right)$

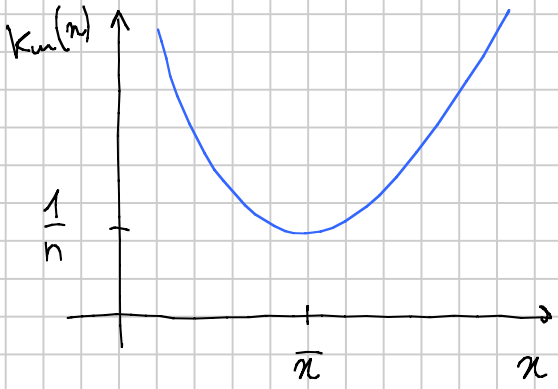
$$E(A+Bx_0) = E(A) + x_0 E(B) = \alpha + \beta x_0$$

$$\begin{aligned} Var(A+Bx_0) &= Cov(A+Bx_0 \; ; \; A+Bx_0) = Cov(A;A) + \\ &Cov(A;Bx_0) + Cov(Bx_0;A) + Cov(Bx_0;Bx_0) \\ &= Cov(A;A) + x_0 Cov(A;B) + x_0 Cov(B;A) + x_0^2 Cov(B;B) \\ &= Var(A) + 2x_0 Cov(A;B) + x_0^2 Var(B) \end{aligned}$$

$$= \sigma^2 k_A - 2x_0\sigma^2 \frac{\bar{x}}{n(\overline{x^2}-\bar{x}^2)} + x_0^2\sigma^2 k_B = \sigma^2 \cdot \frac{\overline{x^2} - 2x_0\bar{x} + x_0^2}{n(\overline{x^2}-\bar{x}^2)}$$

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

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



$$A + Bx_0 \sim \mathcal{N}\left(\alpha+\beta x_0 ;\ \sigma^2 k_m(x_0)\right)$$

iii. funz. anc. $$Z = \frac{A + Bx_0 - (\alpha + \beta x_0)}{\sigma\sqrt{k_m(x_0)}} \sim \mathcal{N}(0,1)$$

Se indip da A, da B e quindi da $A+Bx_0$ perciò

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{Se\sqrt{k_m(x_0)}} \sim t(n-2)$$

Funzione ancillare per la risposta media a livello di ingresso x_0

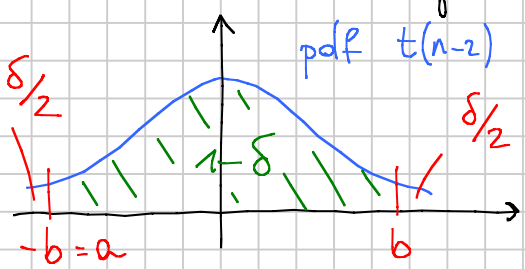
⑨ INTERVALLI DI CONFIDENZA

- per la risposta media

i. fisso il livello di confidenza $1-\alpha$ $1-\delta$ 0,95 0,90 0,99

ii. fisso la f. anc. $\frac{A+Bx_0-(\alpha+\beta x_0)}{Se\sqrt{k_m(x_0)}} \sim t(n-2)$

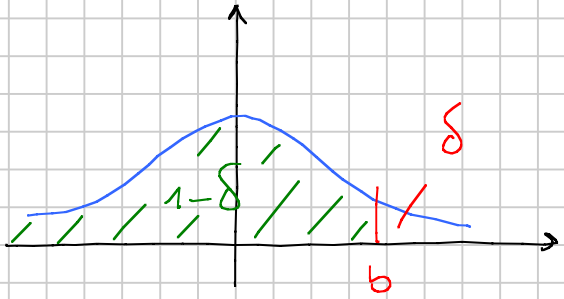
iii. trovo il/i quantili della legge relativa, per il livello di confidenza assegnato



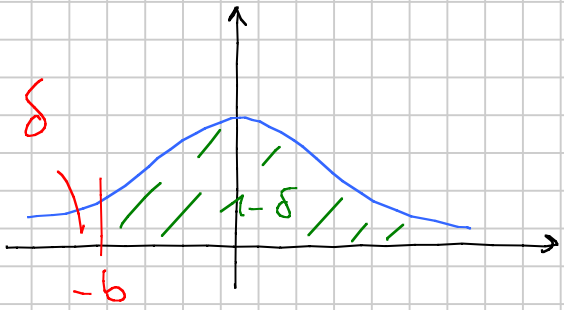
$$F_{t(n-2)}(b) := P(t(n-2) \le b) = 1 - \frac{\delta}{2}$$

$$b = F^{-1}_{t(n-2)}\left(1-\frac{\delta}{2}\right) \overset{\text{TINV}}{=} INV.T(\delta ; n-2)$$

$-b$



$$b = F^{-1}_{t(n-2)}(1-\delta) = INV.T(2*\delta\,;\,n-2)$$



oppure $\quad -b = -INV.T(2*\delta\,;\,n-2)$

iv. Scrivo $1-\delta = P(\dots)$ e ricavo il parametro $\alpha+\beta x_0$

$$1-\delta = P\left(-b \le \frac{A+Bx_0-(\alpha+\beta x_0)}{Se\sqrt{km(x_0)}} \le b\right) = P\left(-b\,Se\sqrt{\ } \le A+Bx_0-(\alpha+\beta x_0) \le b\,Se\sqrt{\ }\right)$$

$\frac{\cdot}{Se\sqrt{km(x_0)}}$: $t(n-2)$

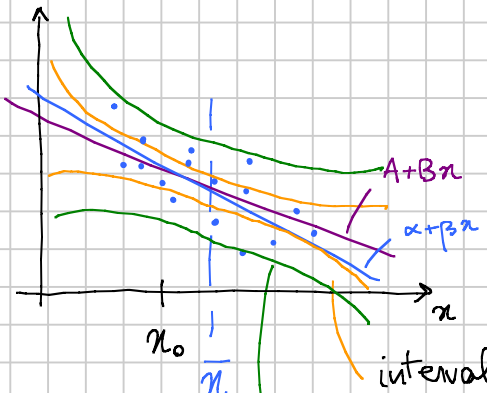
$$= P\left(-(A+Bx_0)-b\,Se\sqrt{\ } \le -(\alpha+\beta x_0) \le -(A+Bx_0)+b\,Se\sqrt{\ }\right)$$

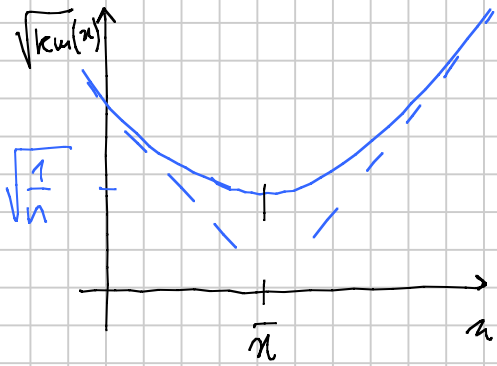
$$= P\left(A+Bx_0+b\,Se\sqrt{\ } \ge \alpha+\beta x_0 \ge A+Bx_0-b\,Se\sqrt{\ }\right)$$

$$= P\left(A+Bx_0-b\,Se\sqrt{\ } \le \alpha+\beta x_0 \le A+Bx_0+b\,Se\sqrt{\ }\right)$$

$$\alpha+\beta x_0 \in \underbrace{A+Bx_0}_{\text{centro}} \pm \underbrace{b\,Se\sqrt{km(x_0)}}_{\text{raggio}} \quad \text{con lvl di conf } 1-\delta$$

v. Grafico





intervallo di confidenza per la risposta media

$$\alpha+\beta x \in A+Bx \pm \text{raggio}(x)$$

intervallo di predizione per una risposta futura

$$\alpha+\beta x+e \in A+Bx \pm \text{raggio\_maggiore}(x)$$

MMSD (B)

lez 7

Note Title 16/03/2011

INTERVALLI DI PREDIZIONE

Simile all'intervallo di confidenza, ma invece che contenere un parametro incognito, contiene una v.a. futura

- Campione gaussiano classico

$X_i \sim \mathcal{N}(\mu, \sigma^2)$ iid $\quad i = 1, 2, \dots, n \quad$ $1-\alpha$ liv di conf

→ int. conf media $\quad \mu \in \bar{X} \pm q \frac{S}{\sqrt{n}} \quad q = F^{-1}_{t(n-1)}\left(1-\frac{\alpha}{2}\right)$

→ int. di predizione per X_{n+1} $\quad \boxed{X_{n+1} \in \bar{X} \pm q\, S \sqrt{\frac{1}{n}+1}}$

$$\bar{X} - X_{n+1} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} + \sigma^2\right) \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n}+1\right)\right)$$

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad$ $X_{n+1} \sim \mathcal{N}(\mu, \sigma^2)$ indip perché X_{n+1} è nuovo

HW: ripasso, approfondire

$$\frac{\bar{X} - X_{n+1}}{\sigma\sqrt{\frac{1}{n}+1}} \sim \mathcal{N}(0,1)$$

non è funz. ancillare ma si usa similmente

$$\frac{\bar{X} - X_{n+1}}{S\sqrt{\frac{1}{n}+1}} \sim t(n-1)$$

HW: verificare

q quantile t.c. $\quad 1-\alpha = P\left(-q \le t(n-1) \le q\right)$

$$1-\alpha = P\left(-q \le \frac{\bar{X} - X_{n+1}}{S\sqrt{\frac{1}{n}+1}} \le q\right) = P\left(\bar{X} - qS\sqrt{\frac{1}{n}+1} \le X_{n+1} \le \bar{X} + qS\sqrt{\frac{1}{n}+1}\right)$$

HW

- Regressione

$Y_i = \alpha + \beta x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2)$ iid $\quad i = 1, 2, \dots, n$

$A + Bx$ retta di regressione

$S_e \approx \sigma$ errore standard

$$A + Bx \sim \mathcal{N}\left(\alpha + \beta x \,;\, \sigma k_m\right) \qquad k_m = \frac{1}{n} + \frac{(x - \bar{x})^2}{n(\bar{x^2} - \bar{x}^2)}$$

(x_{n+1}, Y_{n+1}) punto futuro $\qquad x_{n+1}$ noto

↑ 180 nell'esempio

$$Y_{n+1} = \alpha + \beta x_{n+1} + e_{n+1} \sim \mathcal{N}\left(\alpha + \beta x_{n+1} \,;\, \sigma^2\right)$$

$$A + B x_{n+1} - Y_{n+1} \sim \mathcal{N}\left(0 \,;\, \sigma^2 k_m(x_{n+1}) + \sigma^2\right) \sim \mathcal{N}\left(0 \,;\, \sigma^2\left(k_m(x_{n+1}) + 1\right)\right)$$

$A + B x_{n+1} \sim \mathcal{N}\left(\alpha + \beta x_{n+1} \,;\, \sigma^2 k_m(x_{n+1})\right)$

$-Y_{n+1} \sim \mathcal{N}\left(-\alpha - \beta x_{n+1} \,;\, \sigma^2\right)$ indip

$$\frac{A + B x_{n+1} - Y_{n+1}}{\sigma \sqrt{k_m(x_{n+1}) + 1}} \sim \mathcal{N}(0, 1)$$

$1 - \delta$ livello di confidenza

$$\frac{A + B x_{n+1} - Y_{n+1}}{S_e \sqrt{k_m(x_{n+1}) + 1}} \sim t(n-2)$$

HW: verificare

q quantile t.c. $1 - \delta = P\left(-q \le t(n-2) \le q\right)$

$$Y_{n+1} \in A + B x_{n+1} \pm q\, S_e \sqrt{k_m(x_{n+1}) + 1}$$

...

- Approfondimento: int. di predizione quando i dati futuri sono tanti

→ domanda di una bibita: prevedere tutto il mese prossimo non si possono fare intervalli singoli e "metterli assieme" in qualche modo.

$X_1, X_2, \dots, X_n$ $X_{n+1}, \dots, X_{n+m}$ iid $\mathcal{N}(\mu, \sigma^2)$

passato futuro

$T := \sum_{i=1}^{m} X_{n+i}$ somma

$U := \frac{1}{m} T$ media

$T \sim \mathcal{N}(m\mu ; m\sigma^2)$

$U \sim \mathcal{N}(\mu ; \frac{\sigma^2}{m})$

HW: check

$T \in ?$ $\quad m\mu \approx m\bar{X}$ $\quad \bar{X} = \frac{1}{n}\sum_1^n x_i$ $\quad m\bar{X} \sim \mathcal{N}\left(m\mu ; \frac{m^2}{n}\sigma^2\right)$ HWC

$$m\bar{X} - T \sim \mathcal{N}\left(0, \frac{m^2}{n}\sigma^2 + m\sigma^2\right) \sim \mathcal{N}\left(0, m^2\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

....

$$T \in m\bar{X} \pm q\, S\, m\sqrt{\frac{1}{n} + \frac{1}{m}}$$

HWC

quantile $t(n-1)$

$$\bar{X} - U \sim \mathcal{N}\left(0 ; \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right) \sim \mathcal{N}\left(0 ; \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

...

$$U \in \bar{X} \pm q\, S \sqrt{\frac{1}{n} + \frac{1}{m}}$$

HWC

---

ora 8

→ int. conf media $\quad \mu \in \bar{X} \pm q\frac{S}{\sqrt{n}}$

→ int. di predizione per X_{n+1} $\quad X_{n+1} \in \bar{X} \pm q\, S\sqrt{\frac{1}{n} + 1}$

L'int. di predizione per U generalizza entrambe $(m = 1, \infty)$

★ L'int. di pred. per T è semplicemente quello di U "espanso" di un fattore m :

→ $U \in [2{,}2 \ ; \ 2{,}5]$ $\quad m = 30$ gg $\quad \rightarrow \quad T \in [66 \ ; \ 75]$

- Stessa cosa si può fare per la regressione
(cfr i due anni precedenti per avere esempi di T e U
con $x_{n+1} = x_{n+2} = \dots = x_{n+m}$ costante)

TEST STATISTICI SUI PARAMETRI DI REGRESSIONE

- Esempio

Acceptable Quality Level

i. β parametro incognito

ii. β_0 parametro di confronto (in altri casi target / AQL)
di solito $\beta_0 = 0$ nella regressione

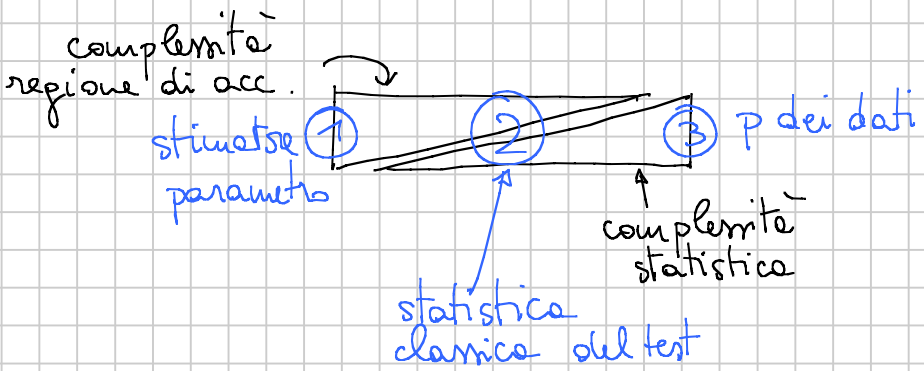
iii. ipotesi bilaterali o unilaterali (e nel II caso, H_0 / H_1)

$H_0: \beta = 0$ — "Y non dipende da x" (non c'è regressione) — $Y = \alpha + \beta x + e$

$H_1: \beta \neq 0$ — c'è regressione — conoscere x ci aiuta a prevedere Y

Nell'esempio, p dei dati 10^{-81} quindi nettamente H_1

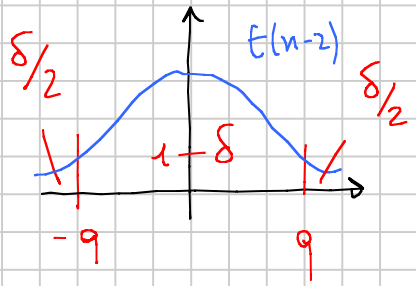
iv.



(2) stat. del test: funz. ancillare e cambio β con β_0

$$\frac{\beta - \beta}{s_e \sqrt{k_\beta}} \sim t(n-2) \quad \Rightarrow \quad \boxed{T_\beta = \frac{\beta}{s_e \sqrt{k_\beta}} \overset{H_0}{\sim} t(n-2)}$$

v. Trovo i quantili bi- o uni-laterali per la distrib. $t(n-2)$
(per lvl di significatività δ assegnato)



$\pm q$ $\qquad$ $q = \text{INV.T}(\delta\ ;\ n-2)$

vi. Se H_0 è vera $T_B \sim t(n-2)$ quindi $T_B \in [-q;q]$ quasi sempre
Allora, se $T_B \in [-q;q]$ dico H_0, altrimenti dico H

$$RA_{T_B} := [-q;q]$$ regione di accettazione

① stimatore T_B → B statistica $\qquad RA_B = ?$

vii. $$T_B \in RA_{T_B} \iff \frac{B}{Se\sqrt{k_B}} \in [-q;q] \iff B \in \left[-q\,Se\sqrt{k_B}\ ;\ q\,Se\sqrt{k_B}\right]$$

$$RA_B := \left[-q\,Se\sqrt{k_B}\ ;\ q\,Se\sqrt{k_B}\right]$$

③ p dei dati $\qquad \alpha^*$ $\qquad$ (o δ^* ?)

$$T_B \in RA_{T_B} \iff T_B \in [-q;q] \iff |T_B| \le q = F^{-1}_{t(n-2)}\left(1-\frac{\delta}{2}\right)$$

$$\iff F_{t(n-2)}(|T_B|) \le 1-\frac{\delta}{2} \iff \frac{\delta}{2} \le 1 - F_{t(n)}(|T_B|)$$

↑ F monotona crescente

$$\iff \delta \le 2 - 2F_{t(n-2)}(|T_B|) =: \delta^*$$

p dei dati

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

$$RA_{\delta^*} := [\delta\ ;\ 1]$$

★ Se non ho fissato il liv di significatività ha senso a maggior ragione calcolare i pdd: se è molto ($\sim 10^{-4}$) piccolo si sceglie H_1; se è molto ($\sim 30\%$) grande si sceglie H_0; negli altri casi si discute

- Altri test: $H_0: \alpha = 0$; bi/unilaterali su σ ; bi/unilaterali su $\alpha + \beta x_0$ analoghi → non approfondiamo

VALUTAZIONE DEL MODELLO

→ Ci sono errori nel modello o nelle ipotesi iniziali?

→ È un modello utile? Quanto è vale conoscere x?

ERRORE STANDARD E COEFFICIENTE DI DETERMINAZIONE

★ $Se := \sqrt{\frac{SS_R}{n-2}} \approx \sigma$ margine di errore su Y nota x

→ ha la stessa unità di misura di Y

9,3 kg nell'esempio è l'incertezza ineliminabile su Y conoscendo "solo x"

★ Coeff. di determinazione

$R^2 := 1 - \frac{SS_R}{S_{YY}}$ calo percentuale di incertezza sulle Y_i che si ottiene "aggiungendo" le x_i

$SS_R = \sum_i (Y_i - A - Bx_i)^2$ incertezza residua, note le x_i

$S_{YY} = \sum_i (Y_i - \bar{Y})^2$ incertezza sulle Y, noto niente

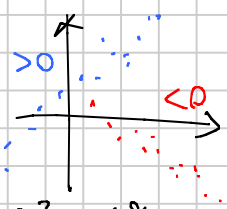
$SS_R \leq S_{YY}$

→ Più grande è migliore è la regressione, e più utili sono le x_i

→ Nel caso della regr. lin. semplice $R^2 = r^2$

dove r è il coeff di correl. campionaria (cfr Cap 2)

$-1 \leq r \leq 1$ stimatore di $\rho(X;Y) = \frac{Cov(X;Y)}{\sqrt{Var(X)\,Var(Y)}}$



★ Se due var hanno $r = 0{,}4$ non è un gran che, perché $R^2 = 16\%$

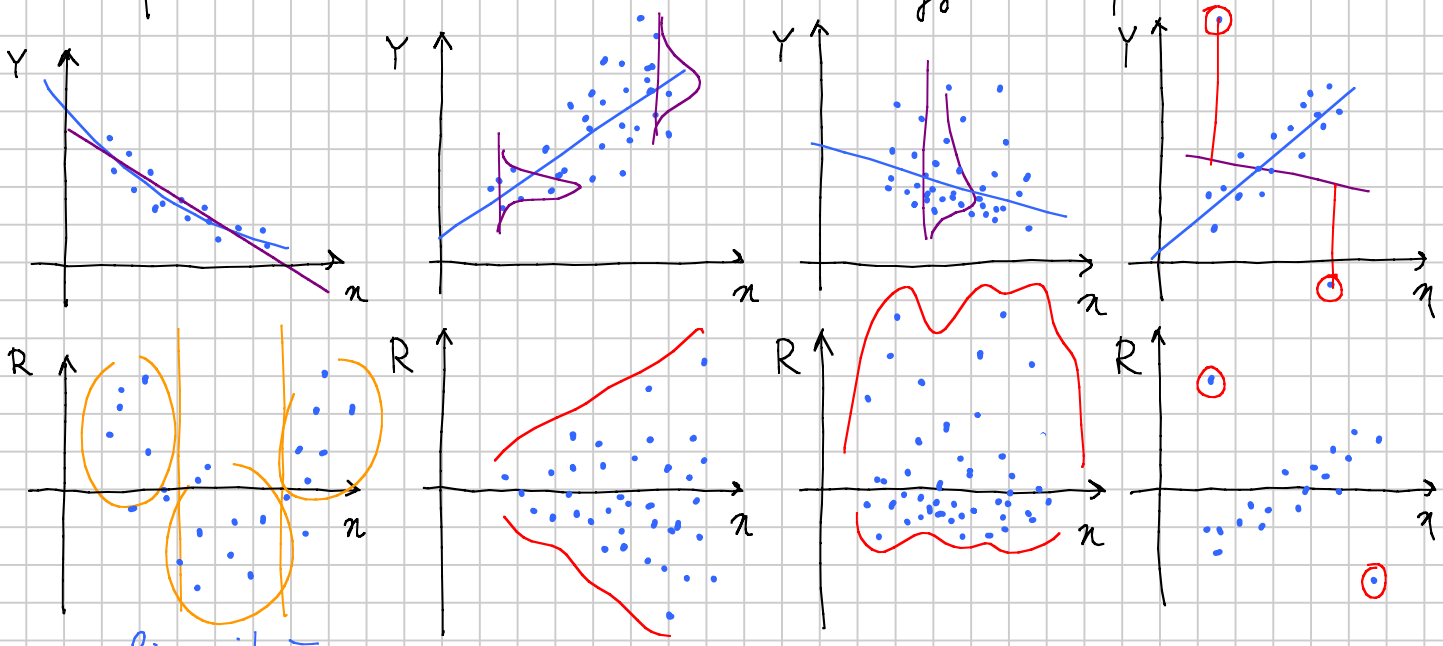
ANALISI DEI RESIDUI

La regressione è un modello corretto se davvero:

$$Y_i = \alpha + \beta x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2) \text{ indip.}$$

funzione lineare — errori normali — omoschedasticità

★ Occorre diagnosticare se ci sono problemi rispetto a questi aspetti. Successivamente vedremo come correggerli in qualche caso



nonlinearità

$Y = f(x) + e$

f nonlineare

eteroschedasticità

$\sigma = \sigma(x)$

$\text{Var}(e_i) = \sigma(x_i)$

errori non normali

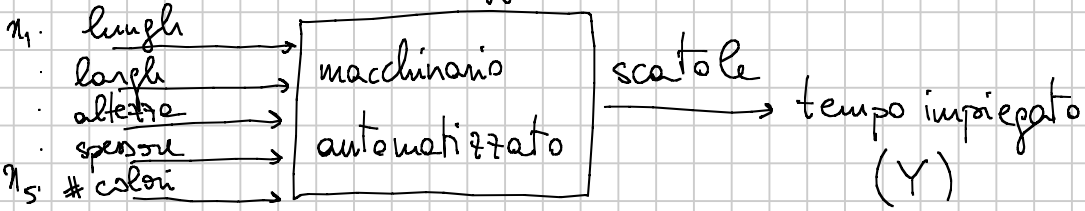
presenza di outliers

HW: Ripassare le matrici (prodotto riga/colonna; inversa; determinante ....)

REGRESSIONE LINEARE MULTIPLA

→ tecnica multivariata

→ esempio: imballaggi di cartone



$$Y = f(x_1, x_2, x_3, x_4, x_5) + e$$

✱ Non occorre a priori ragionare su quali variabili x_i potrebbero non incidere su Y: si fa a posteriori.

✱ Regressione: si fanno n esperimenti $i = 1, 2, \dots, n$ con configurazioni variabili $x_j(i)$ $j = 1, 2, \dots, p$ ($p = 5$ nell'esempio)

$x_j(i) =: x_{ij}$: valore della var. ingr. j nell'esperimento i

→ si ricavano n valori della var. di risposta

$Y(i) =: Y_i$: risposta nell'esperimento i

→ da tutti questi dati stimo il modello $Y = f(\vec{x}) + e$ da usare in futuro e per l'inferenza su qualunque parametro

✱ Ipotesi di lavoro:

i. modello lineare: $f(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ ($x_0 \equiv 1$)

a volte suppongo di definire $x_0 \equiv 1$, così: $f(\vec{x}) = \sum_{j=0}^{p} \beta_j x_j$

ii. modello omoschedastico: $Y(i) = f(x_1(i), x_2(i), \dots, x_p(i)) + e(i)$

$e(i) \sim \mathcal{N}(0, \sigma^2)$ σ^2 costante, non dipende da i

iii. $e(i)$ indipendenti

$$Y(i) = \sum_{j=0}^{p} \beta_j x_j(i) + e(i) \qquad e(i) \sim \mathcal{N}(0, \sigma^2) \text{ indip.}$$

- Stimatori di β_j sono B_j $\quad j = 0, 1, \dots, p$

Residui : $R(i) = Y(i) - \hat{Y}(i) = Y(i) - \sum_{j=0}^{p} B_j x_j(i)$

Previsti : $\hat{Y}(i) = \sum_{j=0}^{p} B_j x_j(i) = B_0 + B_1 x_1(i) + B_2 x_2(i) + \dots$

$A + Bx(i)$ nella r. lin. semplice

★ Notazione matriciale

→ X matrice $n \times (p+1)$ $\qquad X_{i,j} = x_{i,j} = x_j(i)$

(occhio che $i = 1, 2, \dots, n$ ma $j = 0, 1, \dots, p$)

$$X = \begin{pmatrix} 1 & x_1(1) & x_2(1) & \dots & x_p(1) \\ 1 & x_1(2) & & & \\ 1 & & & & \\ 1 & \vdots & \vdots & & \vdots \\ \vdots & & & & \\ 1 & x_1(n) & x_2(n) & & x_p(n) \end{pmatrix} \qquad Y = \begin{pmatrix} Y(1) \\ \vdots \\ Y(n) \end{pmatrix} \qquad e = \begin{pmatrix} e(1) \\ \vdots \\ e(n) \end{pmatrix}$$

→ Y, e vettori $\quad n \times 1$

→ β e B vettori $\quad (p+1) \times 1$

$$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T \qquad B = (B_0, B_1, \dots, B_p)^T$$

→ $Y = X\beta + e$ HWC

R vettore $n \times 1$ dei residui

$R = Y - XB$

→ $SS_R = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} (Y(i) - XB(i))^2 = \|R\|^2$ (scalare; vettore)

→ Equazioni normali (minimizzare SS_R in funzione di B)

$$\left[\nabla_B SS_R\right]_j = \frac{\partial}{\partial B_j} \sum_{i=1}^{n} \left(Y_i - [XB]_i\right)^2 = 2\sum_{i=1}^{n} \left(Y_i - [XB]_i\right) \frac{\partial}{\partial B_j}\left(Y_i - \sum_k X_{ik} B_k\right)$$

$$= -2\sum_{i=1}^{n} \left(Y_i - [XB]_i\right) X_{ij} = -2\left[\sum_i Y_i X_{ij} - \sum_i [XB]_i X_{ij}\right]$$

$$= -2\left[Y^T X - [XB]^T X\right]_j = -2\left[Y^T X - B^T X^T X\right]_j = 0$$

$$B^T X^T X = Y^T X \qquad (X^T X)^T B = X^T Y$$

→ $(X^T X)^T = X^T X^{T^T} = X^T X$ quindi $X^T X$ è simmetrica

$(p+1) \times (p+1)$

$$\boxed{X^T X B = X^T Y}$$ equazioni normali

$$\boxed{B = (X^T X)^{-1} X^T Y}$$ se $X^T X$ è invertibile, posso ricavare B

★ $X^T X$ è invertibile se X ha rango massimo (ovvero $p+1$) cosa che è quasi sempre vera, soprattutto se le $x_{i,j}$ sono misure del mondo reale

★ Se $\det(X^T X) = 0$ non si può procedere ma se $\det(X^T X)$ è anche solo piccolo la formula per B è malcondizionata, numericamente instabile, quindi errori approssimazioni e incertezze anche piccole su Y vengono amplificate su B

ora 11

- Funzioni da usare su Excel $(X^T X)^{-1}$ $B = (X^T X)^{-1} X^T Y$
  - MATR.PRODOTTO (mat1 ; mat2) (MMULT)
  - MATR.INVERSA (mat1) (MINVERSE)
  - MATR.TRASPOSTA (mat1) (TRANSPOSE)

  } Ctrl-shift-invio
  - MATR.SOMMAPRODOTTO (vet#1; vet#2) (SUMPRODUCT) prodotto scalare

STIMATORE PER σ^2

$$S_e^2 := \frac{SS_R}{n-(p+1)} \approx \sigma^2$$

DISTRIBUZIONE STIMATORI

- $$\frac{S_e^2}{\sigma^2}(n-p-1) = \frac{SS_R}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i - \sum_j B_j X_{ij}}{\sigma}\right)^2 \sim \chi^2(n-p-1)$$

$\mathcal{N}(0,\sigma^2)$ indip

$$\sum_{i=1}^{n}\left(\frac{Y_i - \sum_j \beta_j X_{ij}}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{e(i)}{\sigma}\right)^2 \sim \chi^2(n)$$

inoltre S_e indipendente da B
(tutto per la solita versione del thm di Cochran che vedremo più avanti)

$$\frac{S_e^2}{\sigma^2}(n-p-1) \sim \chi^2(n-p-1)$$

S_e^2 stimatore corretto di σ^2 HWC

- Distribuzione di $B = (B_0, B_1, \dots, B_p)^T$

(X,Y) $\quad f_{X,Y}(s,t)$

B $\quad f_B(s_0,s_1,\dots,s_p) = (2\pi)^{-\frac{p+1}{2}} \exp\{ \_\_\_\_ \}$

Non occorre scrivere f_B. Ci basta osservare che

→ ciascuno dei B_j ha legge normale

$$B_j = \left[(X^TX)^{-1}X^TY\right]_j = \sum_{k=1}^{n}\underbrace{\left[(X^TX)^{-1}X^T\right]_{jk}}_{N} Y_k = \sum_k N_{jk} Y_k$$

$\mathcal{N}$; numeri

$$N := (X^TX)^{-1}X^T \qquad \text{matrice } (p+1)\times n$$

→ quindi B ha legge normale multivariata
per fissare la legge di B bisogna conoscerne:

i. vettore delle medie

ii. matrice di covarianza

★ $E(B) = E\left((X^TX)^{-1}X^TY\right) = E(NY) = NE(Y) = NE(X\beta + e)$ (N deterministica)

$= N(X\beta + 0) = NX\beta = (X^TX)^{-1}X^TX\beta = \beta$

quindi B è uno stimatore corretto di β

★ Matrice di covarianza Σ matrice $(p+1)\times(p+1)$

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

i. Σ è simmetrica

ii. $\Sigma_{ii} = Cov(B_i ; B_i) = Var(B_i)$

iii. $\Sigma = \sigma^2 (X^TX)^{-1}$

iv. Sia $V = \sum_j \alpha_j B_j = \alpha \cdot B$ una combinazione lineare dei B_j

$(\alpha_0, \alpha_1, \dots, \alpha_p)^T = \alpha$

Allora $Var(V) = \alpha^T \Sigma \alpha$, ovviamente $E(V) = \alpha \cdot E(B) = \alpha \cdot \beta$

$(1, p+1) \times (p+1, p+1) \times (p+1, 1) = (1, 1)$

Per giustificare iii. e iv. diamo la proprietà più generale

X, Y vettori aleatori $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^n$

$S = Cov(X ; Y)$ $\quad S_{ij} = Cov(x_i ; Y_j)$ $\quad S$ è $m \times n$

M matrice $r \times m$

N matrice $s \times n$

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

$$= \sum_k \sum_\ell M_{ik} N_{j\ell} Cov(X_k ; Y_\ell) = \sum_k \sum_\ell M_{ik} N_{j\ell} S_{k\ell}$$

$$= \sum_k \sum_\ell M_{ik} S_{k\ell} N^T_{\ell j} = [MSN^T]_{ij}$$

$$\boxed{Cov(MX; NY) = M S N^T}$$

★ $\Sigma = Cov(B;B) = Cov(NY;NY)$

$= N\,Cov(Y;Y)N^T = N\sigma^2 I N^T$

$= \sigma^2 NN^T = \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1}$

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

iii. verificato

$N = (X^TX)^{-1}X^T$

Y_i indipendenti

$Y_i \sim \mathcal{N}(X\beta; \sigma^2)$

$Cov(Y;Y) = \sigma^2 I = \begin{pmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ & & \ddots \\ 0 & & & \sigma^2 \end{pmatrix}$

★ $V = \alpha \cdot B = \alpha^T B$ $\quad \alpha^T$ matrice $1 \times (p+1)$

$Var(V) = Cov(V;V) = Cov(\alpha^T B; \alpha^T B) = \alpha^T Cov(B;B)\alpha = \alpha^T \Sigma \alpha$

iv. verificato

ora 12

TUTTE LE DISTRIBUZIONI DELLA REGRESSIONE LIN. MULTIPLA

$Y_i \sim \mathcal{N}(\sum_j \beta_j x_j(i), \sigma^2)$ indipendenti

- Y vettore $\mathcal{N}(X\beta, \sigma^2 I)$
- B vettore $\mathcal{N}(\beta, \Sigma)$ stimatore β $\quad \Sigma = \sigma^2 (X^TX)^{-1}$
- $\frac{S_e^2}{\sigma^2}(n-p-1) \sim \chi^2(n-p-1)$ Se stimatore σ

} indip

- $\tilde{x} = (1, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_p)^T$ un punto fissato per le var di ingresso

$\tilde{Y} = B \cdot \tilde{x}$ valore previsto per la risposta Y a lvl di ingresso $\tilde{x}$

$\tilde{Y} \sim \mathcal{N}(\beta \cdot \tilde{x}; \tilde{x}^T \Sigma \tilde{x})$ si usa per stimare risposte medie e future

è indipendente da S_e

- Inferenza:

- su $\beta_j \approx B_j$ $\quad \dfrac{B_j - \beta_j}{\sigma\sqrt{\left[(X^TX)^{-1}\right]_{jj}}} \sim \mathcal{N}(0,1)$

$$\boxed{\frac{B_j - \beta_j}{S_e\sqrt{\left[(X^TX)^{-1}\right]_{jj}}} \sim t(n-p-1)}$$ funz. ancillare per β_j

- su $\sigma \approx S_e$ $\quad \boxed{\dfrac{S_e^2}{\sigma^2}(n-p-1) \sim \chi^2(n-p-1)}$ f. anc. per σ

★ Permettono di eseguire tutti i test e gli intervalli di confidenza su β_j e σ

- Sulla risposta futura $\tilde{x} = (1, \tilde{x}_1, \dots, \tilde{x}_p)^T$

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

$$\tilde{Y} \sim \mathcal{N}(\beta \cdot \tilde{x}\,;\, \sigma^2)$$

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

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

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

$$\rightarrow P\left(\tilde{Y} \in B\cdot\tilde{x} \pm q\, S_e\sqrt{1+\tilde{x}^T(X^TX)^{-1}\tilde{x}}\right) = 1-\delta \qquad \text{dove } q = F^{-1}_{t(n-p-1)}\left(1-\frac{\delta}{2}\right)$$

questo esempio è il caso bilaterale

- Sulla risposta media $\beta\cdot\tilde{x} \approx B\cdot\tilde{x}$

$$\boxed{\frac{B\cdot\tilde{x} - \beta\cdot\tilde{x}}{S_e\sqrt{\tilde{x}^T(X^TX)^{-1}\tilde{x}}} \sim t(n-p-1)}$$

BONTÀ DEL MODELLO

a) Selezione delle variabili
b) Coefficiente di determinazione
c) Errore standard

* A posteriori della regressione si può valutare che una o più variabili non abbiano un effetto significativo su Y
In questi casi è naturale esplorare diversi modelli, ottenuti scegliendo di volta in volta quali var inserire e quali no.
La bontà di questi modelli può essere valutata sotto diversi aspetti.

SELEZIONE DELLE VARIABILI

Fatta la regressione e trovati β, $(X^TX)^{-1}$ usando tutte le var di ingresso, è possibile eseguire i test:

$H_0: \beta_j = 0$ $\quad$ $H_1: \beta_j \neq 0$

$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_p x_p + \varepsilon$

"Y non dipende da x_j" $\quad$ "Y dipende da x_j"

posso togliere x_j $\quad$ la devo tenere

* Può succedere (succede spesso) la situazione seguente:
eseguo la regressione e trovo che x_3 e x_4 non sono significative $\alpha_3^* = 87\%$ $\quad$ $\alpha_4^* = 72\%$
Se tolgo x_3 e rifaccio, trovo $\alpha_4^* = 0{,}4\%$
Se tolgo x_4 e rifaccio, trovo $\alpha_3^* = 0{,}03\%$
Se tolgo entrambe ho un modello peggiore ma non me ne accorgo.

→ Il p dei dati del test $\beta_j = 0$ dipende tantissimo da quali altre var sono nel modello

→ Questo succede in particolare quando le var in questione sono tra loro correlate.

✯ Quando si usa Excel, la cosa più ragionevole è partire con tutte le variabili e andare per passi, togliendo di volta in volta una sola variabile tra quelle non significative. Di solito si sceglie quella con p dei dati maggiore, ma se c'è tempo non è male fare diversi tentativi

Questa tecnica si chiama stepwise backward e si arresta quando tutte le var residue risultano significative

Soglie ragionevoli per dire che x^* è significativo sono:

~~~30%~~

~~~5%~~

~1‰ = $\frac{5\%}{p+1}$

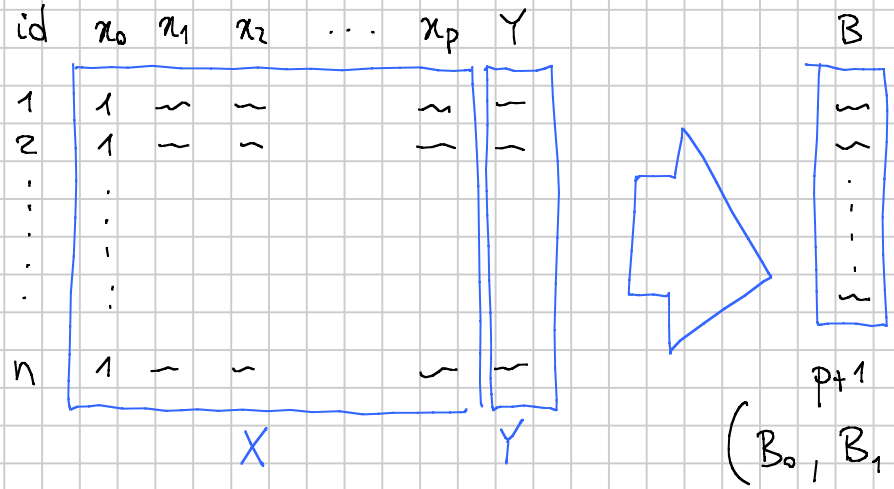
per modello predittivo (butto sob le var che non sembrano proprio dire niente)

per sapere quali var impattano di sicuro (tengo sob le var sicure)

MMSD (B) ora 13

Note Title 30/03/2011

REGRESSIONE MULTIPLA : BONTÀ DEL MODELLO



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

$p+1$ coefficienti $(B_0, B_1, \dots, B_p)$

- Selezione delle variabili

a) Stepwise backward :

i. inizio con tutte,

ii. faccio la regressione,

iii. tolgo la peggiore variabile se è il caso

iv. se l'ho tolta, torno a ii. altrimenti ho finito

ora 15

$SS_R = Y^T (Y - XB)$ HWC

(c'è qualcosa di simile sul Ross)

★ $Y = \beta_0 + \cancel{\beta_1 x_1} + \dots + \beta_p x_p + e$

va tolta se $\beta_1 = 0$ o comunque molto piccolo

esiste un test $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

Excel dà automaticamente il p dei dati di ciascuno

(Oppure $\text{DISTRIB.T}(\text{ASS}(\text{stat}); n-p-1)$

$$\text{stat} = \frac{B_1}{S_e \sqrt{[X^T X^{-1}]_{1,1}}}$$

HWC)

$\alpha_0^*, \alpha_1^*, \alpha_2^*, \dots, \alpha_p^*$ più è alto, più sono convinto di buttare una variabile

★ Di solito si butta quella con α^* maggiore, ma non è detto che sia ottimale come scelta

(Sotto H_0 $\alpha^* \sim unif[0,1]$)

★ Soglia per fermarsi: quando tutti gli α_i^* residui sono inferiori a:

- se voglio un modello predittivo da usare su casi futuri
  soglia intorno al 30%
- se voglio giudicare su quali variabili agire per ottenere effetti su Y
  soglia intorno a $\frac{5\%}{p+1} \approx 1‰ - 1\%$

◉ PARENTESI: LA CORREZIONE DI BONFERRONI

$H_0: \beta_0 = 0$ $H_1: \beta_0 \neq 0$
$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$
. . . .
$H_0: \beta_p = 0$ $H_1: \beta_p \neq 0$

} $p+1$ test che vanno fatti contemporaneamente { $\alpha_0^*, \alpha_1^*, \vdots, \alpha_p^*$

Voglio un lvl di sign del 5%
la prob di dire H_1 quando non è vera $\leq 5\%$
Se faccio i test al 5% ho sistematicamente dei falsi H_1
Se abbasso la sign dei singoli test a $\frac{5\%}{p+1}$, quella complessiva risulta $\leq 5\%$

b) Stepwise forward:

i. si fanno p regressioni lineari semplici $Y = \beta_0 + \beta_i x_i + e$
ii. tengo la variabile risultata migliore (α^* più piccolo) (sia k)
iii. si fanno $p-1$ regr. con due variabili

$$Y = \beta_0 + \beta_k x_k + \beta_i x_i + e \qquad i = 1, 2, \dots, p \text{ tranne } i = k$$

iv. tengo la var migliore come seconda dopo x_k
v. continuo fino a che $\alpha^* <$ soglia

★ Irrealistico farlo con Excel, tranne i parr i.e.m. che si riescono ad automatizzare usando bene il comando REGR.LIN

c) Cosa succede a SS_R, R^2, S_e quando cambio le variabili?

$$SS_R := \sum_{i=1}^{n} \left(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - B_3 x_{i3} \right)^2$$

cerca (B_0, B_1, B_2, B_3) che minimizza SS_R

test $\beta_3 = 0$ dice H_0 $\alpha_3^* = 87\%$, tolgo x_3

$$SS_R^* = \sum_{i=1}^{n} \left(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} \right)^2$$

cerca $(B_0, B_1, B_2, 0)$ che minimizza SS_R
cerca (B_0, B_1, B_2) che minimizza SS_R^* } stessa cosa

★ $SS_R^* > SS_R$ sempre e comunque

★ $R^2 := 1 - \frac{SS_R}{S_{YY}}$ questo aumenta $S_{YY} = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 = (n-1) S_Y^2$

$R_*^2 < R^2$ sempre e comunque

★ $S_e^2 := \frac{SSR}{n-p-1}$ Se a volte aumenta, a volte diminuisce

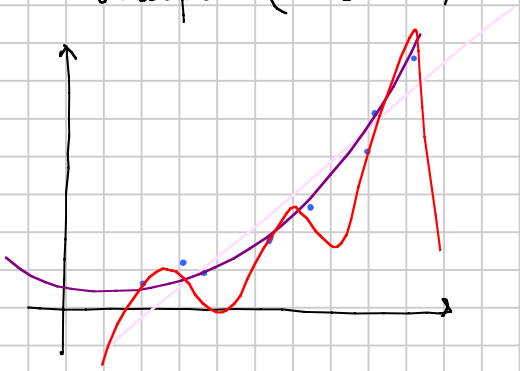
→ Se è un indicatore ragionevole del fatto che la selezione delle var abbia migliorato o no il modello

★ $\bar{R}^2 := 1 - \frac{S_e^2}{S_Y^2}$ "R^2 corretto" (adjusted) si comporta come Se (al contrario)

● Quando si tolgono variabili, guardare se R^2 aumenta (o se Se diminuisce) per capire se il modello è migliorato. Piccole variazioni non sono importanti. In ogni caso non è un metodo ottimale neanche questo.

OVERFITTING

Esempio (remoto)



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

$$Y = \alpha + \beta x + \gamma x^2$$

$$Y = \alpha + \beta x + \gamma x^2 + \delta x^3 + \dots + \zeta x^k$$

Nella regressione se i coefficienti stimati $(p+1)$ sono "troppi" rispetto alle osservazioni (n) si ha lo stesso fenomeno: si trova un modello che si adatta fin troppo bene al campione in esame, ma che sarà pochissimo adattato alla popolazione, quindi alle osservazioni future

- $p+1 = n$ il modello ha residui nulli, passa esattamente per i punti $R_i \equiv 0$, $SS_R = 0$, $S_e = 0$, DIV/0
- $p+1 > n$ errore
- $p+1$ poco minore di n il modello è sbagliatissimo
- $\frac{n}{p+1} \approx 5$ siamo al limite (grossi problemi, ma non da buttare)

  ↑

  ★ in mezzo: buon senso

  ↓

- $\frac{n}{p^2} \approx 2+$ va benissimo

★ Se occorre migliorare n/p, o si spende per aumentare n, oppure si usa la selezione delle var per diminuire p

★ Grosso problema: un effetto naturale dell'overfitting è che conosciamo i coefficienti con precisione molto bassa

- $\beta_i \in B_i \pm$ roba → è grande se c'è overfitting

- $H_0: \beta_i = 0 \quad H_1: \beta_i \neq 0$ ha potenza molto bassa

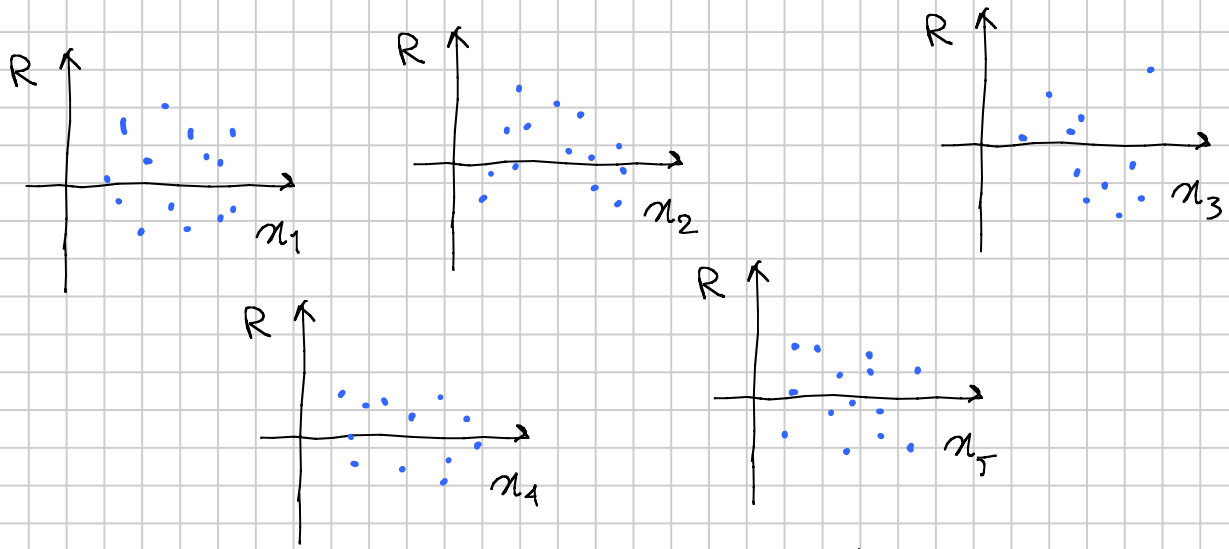
  si dice quasi sempre H_0 (a meno che $|\beta_i|$ sia enorme)
  anche se è vera H_1
  viene α^* spesso grande

→ Questo rende poco affidabile la stepwise backward (si buttano via all'inizio variabili forse buone)
Usare in questi casi la forward, o almeno fare con la backward vari tentativi.

ora 15

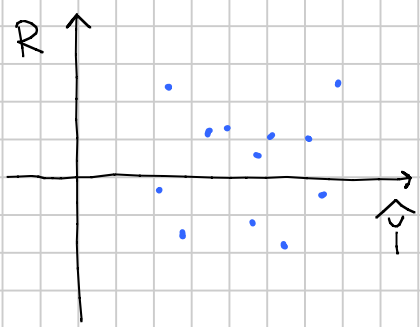
ANALISI DEI RESIDUI

→ Come nella regr lin semplice, va fatta sempre
Purtroppo non c'è una x soltanto
Si fanno tanti grafici: uno per ogni variabile di ingresso e si analizzano tutti



★ Spesso è utile aggiungerne uno:
quello dei residui vs $\hat{Y}$

$\hat{Y}_i = B_0 + B_1 x_{i1} + \dots + B_p x_{ip}$



CURA DELLA REGRESSIONE (quando i residui mostrano stranezze)

OUTLIERS → correggibili
→ non correggibili

0	1	1	0	1	1
0	0	1	1	0	0
0	0	0	~~0~~ 1	0	0
0	1	0	1	1	0

era un errore + era correggibile

Circoscrizioni di Philadelphia: le tre del centro andavano escluse perché troppo diverse

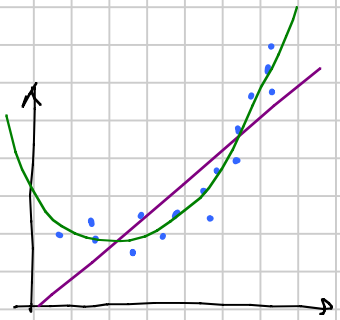
★ Vanno comunque eliminati o corretti

CORREGGERE LA NONLINEARITÀ

Metodo principale: aggiungere dummy variables

→ esempio 1

id	x	Y
1	~	~
2	~	~
⋮	⋮	⋮
n	~	~



$$Y = \beta_0 + \beta_1 x + e$$

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

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

id	x	x^2	Y
1	2,1	$2,1^2$	~
2	1,7	$1,7^2$	~
⋮	⋮	⋮	⋮
n	~	n^2	~

dummy variable

faccio la regr lin multipla sulle due variabili → i coeff che ottengo vanno interpretati come quelli del polinomio

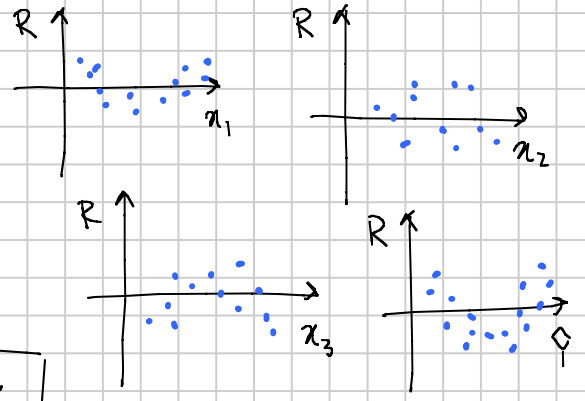
★ Posso alzare il grado progressivamente aggiungendo dummy variables. (Occhio all'overfitting.)

→ Se si esagera con il grado del polinomio, tipicamente i p dei dati vengono tutti (quasi) uguali

→ esempio 2

id	x_1	x_2	x_3	Y
1				
2				
⋮				
n				

id	x_1	x_1^2	x_2	x_3	x_3^2	Y



★ Nella regr multivariata potrei dover aggiungere parecchie dummy variables. (Overfitting!)

x_1 x_2 x_3 x_1^2 x_3^2 x_1x_3

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

in un generico polinomio di II grado in più variabili ci possono essere questi prodotti che si chiamano termini di interazione.

★ Sono molto importanti. Approfondiremo sul DOE

p variabili $x_1, x_2, \dots, x_p$

p quadrati $x_1^2, x_2^2, \dots, x_p^2$

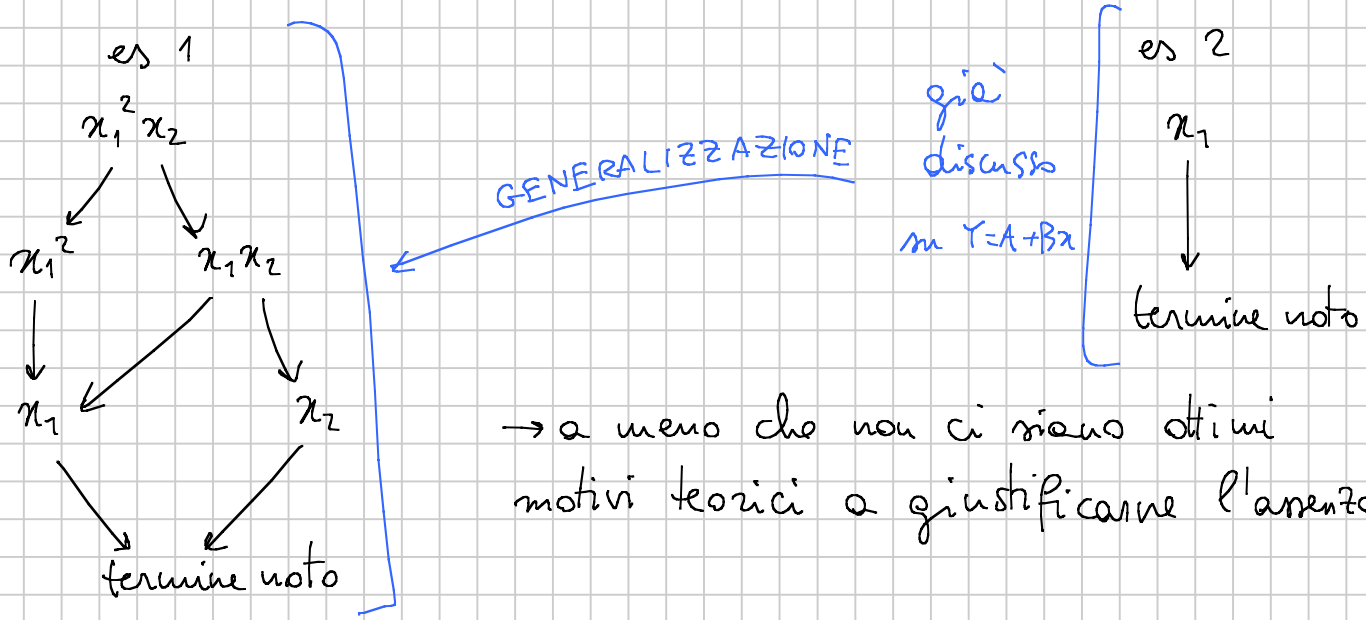
$\binom{p}{2}$ interazioni $x_1x_2, x_1x_3, x_1x_4 \dots x_1x_p, x_2x_3, x_2x_4 \dots x_2x_p, \dots x_{p-1}x_p$

La regr. lin. multipla necessaria ha $p' = p + p + \binom{p}{2} = \frac{p^2+3p}{2}$ variabili → overfitting sicuro

★ Viceversa, se sono dalla parte opposta rispetto all'overfitting (ad es $n = 10000$, $p = 3$) è legittimo e doveroso aggiungere un po' di termini (verranno significativi perché i test sono molto potenti)

HW: capire!

- Regola gerarchica: se si inserisce un termine di grado ≥ 1, vanno inseriti, anche se non significativi, tutti i termini "contenuti"



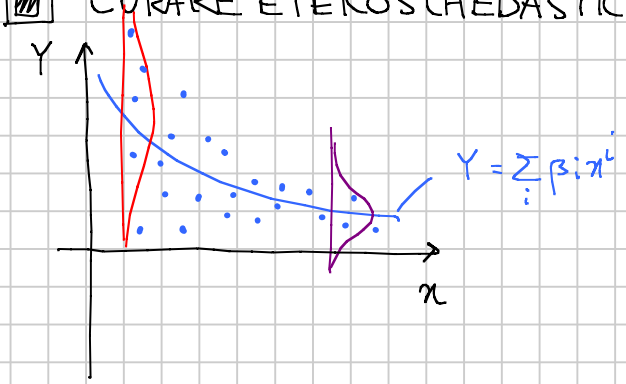
→ a meno che non ci siano ottimi motivi teorici a giustificarne l'assenza

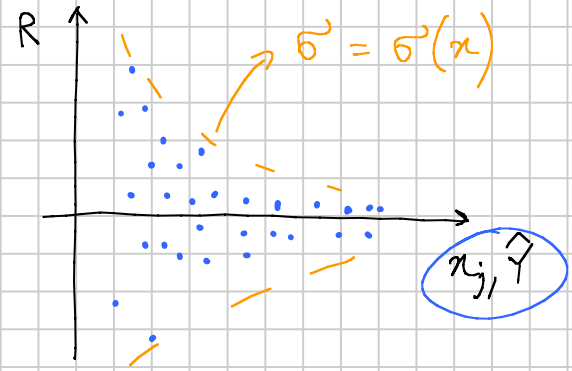
MMSD (B)

ora 16

Note Title 06/04/2011

CURARE ETEROSCHEDASTICITÀ





$$Y(i) = \sum_j \beta_j x_j(i) + e_i$$

$$e_i \sim \mathcal{N}(0, \sigma_i^2) \qquad \sigma_i^2 \equiv \sigma^2 \quad \text{OMOSCHEDASTICITÀ}$$

$$\sigma_i \neq \sigma_j \quad \sigma = \sigma(x) \quad \text{ETEROSCHEDASTICITÀ}$$

Per correggerla occorre conoscere come σ dipende da x

★ Basta stabilire $\sigma(x)$ a meno di una costante moltiplicativa

$\sigma(x) = 7{,}2 \cdot x$ vera $\qquad \sigma(x) \propto x \Rightarrow \sigma(x) = x$ ipotizzata

funziona perfettamente $\sigma(x) = x$, $\sigma(x) = 2x$, $\sigma(x) = 7{,}2x$ ...

★ Andamenti tipici:

a) $\sigma \propto x_j$

b) $\sigma^2 \propto x_j \iff \sigma \propto \sqrt{x}$

a) $\sigma \propto \hat{Y}$

b) $\sigma^2 \propto \hat{Y} \iff \sigma \propto \sqrt{\hat{Y}}$

valori di risposta previsti

$\hat{Y} \to \sigma, \sigma^2 \propto E(Y) = \sum_j \beta_j x_j$

$$\hat{Y} = \sum_j B_j x_j$$

c) vedi oltre

★ Due situazioni tipiche

a) L'errore è "in percentuale" sul valore di Y

$\sigma \approx 25\% \, E(Y)$

quindi $\sigma \propto \hat{Y}$

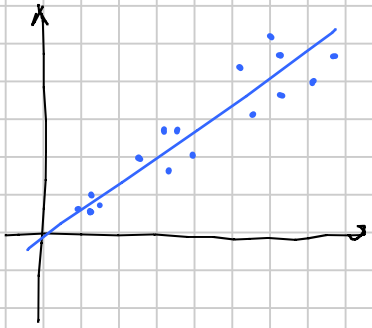
> NB. anche $\sigma \propto x$ è di questo tipo, infatti di solito succede quando $Y \propto x$
>
> $Y = \alpha + \beta x$ se $\alpha \ll 1$ o $\alpha \approx 0$
>
> $Y \propto \beta x$ $\quad Y \propto x$ $\quad \sigma \propto Y \iff \sigma \propto x$
>
> Rappresenta un sottocaso di a), ma se vi siamo conviene approfittarne: x è nota, "Y" no (ovvero non occorre conoscere β)

b) L'errore è la somma di tanti contributi indipendenti, e quanti sono è $\propto Y$ (o $\propto x$)

x : km percorsi (tracciato cittadino)

Y : tempo

$\underbrace{T_1 + T_2 + \ldots + T_x}_{\Sigma \text{ iid}} = Y$



$\sigma^2 = \mathrm{Var}(Y) = x \, \mathrm{Var}(T_1) \propto x$

N.B. Come sopra

★ Difficile distinguere tra a), b) e tra NB sì/no, anche guardando i residui. Comunque se non è chiaro, a parte ciò i risultati saranno simili con qualunque modello

c) Situazione atipica: abbiamo stima di $\sigma(x(i))$ per ogni $x(i)$

- Esempio 1: $x(i) \longrightarrow$ 5 risposte, 4 risposte, 3 risposte $\Big\}$ $Y(i) =$ media

$$Y(i) = \frac{Y_{i,1} + Y_{i,2} + \dots + Y_{i,m}}{m(i)} \qquad m = 3,4,5 \text{ a seconda di } i$$

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{\mathrm{Var}(X_1)}{n}$$

HWC

$$\sigma^2(x(i)) = \mathrm{Var}(Y(i)) = \frac{1}{m(i)} \mathrm{Var}(Y_{i,1}) \propto \frac{1}{m(i)}$$

$\sigma^2 \propto \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$ a seconda dell'esperimento

— Esempio 2: $x_1, x_2, \dots, x_p \longrightarrow$ probabilità $p = \sum_j \beta_j x_j$

$m(i)$

farò un po' di esperimenti per ciascuna combinazione delle var di ingresso

$Y(i)$: # di successi tra gli $m(i)$ esperimenti di tipo i

$$Y(i) \sim \mathrm{bin}(m(i), p(i))$$

$$E(Y(i)) = m(i)p(i) = m(i) \sum_j \beta_j x_j(i)$$

posso fare la regressione

$$\hat{p}(i) = \frac{Y(i)}{m(i)} \qquad E(\hat{p}(i)) = p(i) = \sum_j \beta_j x_j(i)$$

$$\mathrm{Var}(\hat{p}(i)) \equiv \sigma \qquad \text{No}$$

uso $\hat{Y}$ al posto di Y

$$\mathrm{Var}(Y(i)) = m(i)p(i)(1-p(i)) \quad \Rightarrow \quad \mathrm{Var}(\hat{p}(i)) = \frac{p(i)(1-p(i))}{m(i)}$$

$$\sigma^2(i) = \mathrm{Var}(\hat{p}(i)) \propto \frac{1}{m(i)} \bar{p}(i)(1-\bar{p}(i))$$

dove $\bar{p}(i) = \sum_j B_j x_j(i)$ è il previsto

REGRESSIONE ETEROSCHEDASTICA / PESATA (WEIGHTED)

$\sigma \propto$ qualcosa

$$SS_R = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} \left(Y_i - \sum_j B_j x_{ij} \right)^2$$

$R_i \sim e_i$

questo non fa la cosa giusta se le varianze non sono uguali

$$SS_R^W := \sum_{i=1}^{n} R_i^2 w_i = \sum_{i=1}^{n} \left(Y_i - \sum B_j x_{ij} \right)^2 w_i = \sum_{i=1}^{n} \left(\sqrt{w_i} Y_i - \sum_j B_j \sqrt{w_i} x_{ij} \right)^2$$

dove w_i è il peso dell'esperimento / dato i $\quad w_i \propto \frac{1}{\sigma_i^2}$

Se conosco $\sigma \propto \dots$ posso determinare i pesi (a meno di una costante)

$$Y_i \sim \mathcal{N}\left(\sum \beta_j x_{ij} \,,\, \sigma_i^2 \right) \qquad \sqrt{w_i} Y_i \sim \mathcal{N}\left(\sum_j \beta_j \sqrt{w_i} x_{ij} \,,\, w_i \sigma_i^2 \right)$$

$w_i \sigma_i^2$ non dipende da i

✱ Posso risolvere questa regressione, o (tecnica sconsigliata) minimizzando $SS_R^W = SS_R^W(B_1, B_2, \dots, B_p)$; oppure trasformo tutti i dati moltiplicando per $\sqrt{w_i}$

Y	x_0	x_1	x_2	$\dots$	x_p
Y_1	1	$x_1(1)$			$x_p(1)$
Y_2	1	$x_1(2)$			
$\vdots$	$\vdots$	$\vdots$			$\vdots$
Y_n	1	$x_1(n)$			

Y'	x_0'	x_1'	$\dots$	x_p'
$\sqrt{w_1} Y_1$	$\sqrt{w_1}$	$\sqrt{w_1} x_1(1)$		$\sqrt{w_1} x_p(1)$
$\sqrt{w_2} Y_2$	$\sqrt{w_2}$	$\vdots$		$\vdots$
$\vdots$	$\vdots$			
$\sqrt{w_n} Y_n$	$\sqrt{w_n}$	$\sqrt{w_n} x_1(n)$		

$$Y_i' \sim \mathcal{N}\left(\sum_j \beta_j x_{ij}' \,,\, \sigma_*^2 \right) \qquad Y' = \sum \beta_j x_j' + e$$

è una regr. omoschedastica

✱ Lancio la regr su Y', x' e trovo i coeff. B^W che minimizzano SS_R^W

e quindi sono quelli giusti per la regr. eteroschedastica

★ La regr. Y', x' va fatta senza termine noto (spuntare "passa per l'origine")

★ Se servono intervalli di conf/predizione, si calcolano su Y', x' e si trasformano indietro, dividendo per $\sqrt{w(\tilde{x})}$

★ R^2 non è confrontabile con quello della regressione precedente Se non ha senso

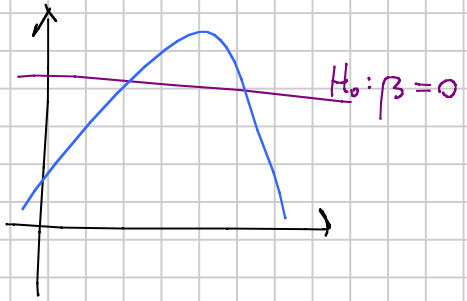
★ Per capire se il modello è migliorato si può solo osservare l'omoschedasticità dei suoi residui.

★ In generale questo "trattamento" va fatto solo in casi eccezionali

---

SCHEMA REGRESSIONE

1) Outliers
2) Selezione variabili
3) Nonlinearità
4) Eteroschedasticità



1, 3, 4) Trasformazioni nonlineari + 5) Residui asimmetrici

TRASFORMAZIONI NONCINEARI

Esempio: prodotto in bottiglia → lavaggio
a che velocità faccio andare la macchina?

$v = [\text{bottiglie}/\text{ora}]$

$v = v(x_1, x_2, \dots, x_p)$

x_1: volume, x_2: profondità, x_p: tipo prodotto, livello di pulizia richiesto

T: tempo di lavaggio $\quad T = [\text{sec}]$

$T = T(x_1, x_2, \dots, x_p) = \sum_j \beta_j x_j + e$

$$T = \sum_j \beta_j x_j + e \qquad e \sim \mathcal{N}(0, \sigma^2)$$

$$v = \frac{3600}{T} = \frac{3600}{\sum_j \beta_j x_j + e}$$

→ relazione non lineare
→ residui non $\mathcal{N}$ / simmetrici
→ eteroschedasticità

$$\frac{a}{b+\mathcal{N}} \sim \frac{a}{\mathcal{N}'} \sim \frac{1}{\mathcal{N}^4} \not\sim \mathcal{N}$$ → fitti outliers

$e \sim \pm 1$

960 ± 240

$\sum \beta_j x_j \approx 4$ den : 3-5 v : 720 ~ 1200

$\sum \beta_j x_j \approx 9$ den : 8-10 v : 360 ~ 450

405 ± 45

- Se la variabile che dipende in maniera semplice dalle x_i non è Y (v) ma una sua riscrittura o trasformazione allora posso migliorare tantissimo la regressione se al posto di Y lavoro su di essa (T)

★ Se faccio una trasformazione sbagliata posso anche peggiorare

★ Trasformazioni lineari non servono quasi a nulla

★ Anche trasformazioni non lineari di alcune delle x_i possono in certi casi aiutare.

ora 18

TRASFORMAZIONI LINEARI NELLA REGRESSIONE

Non hanno quasi effetto!

$$Y' = \lambda + \mu Y \quad , \quad x'_j = \lambda_j + \mu_j x_j$$

- var € → k€ $\lambda = 0$ $\mu = \frac{1}{1000}$
- var % → [0,1] $\lambda = 0$ $\mu = \frac{1}{100}$
- T [Fahr] → T [°C] λ , μ
- var {1,2} → {0,1} $\lambda = -1$ $\mu = 1$

α_j^* $j = 1, 2, \dots, p$, R^2 } non cambiano

grafici dei residui

$S_e' = \mu S_e$

$B_j' = \frac{B_j}{\mu_j}$ $j = 1, 2, \dots, p$ } cambiano in modo prevedibile

★ Standardizzazione $N(\mu, \sigma^2) \to N(0,1)$

$X \longrightarrow \frac{X - \mu}{\sigma}$ HWC

$\mathbb{E}(X) = \mu$ $\mathbb{E}(Y) = 0$

$Var(X) = \sigma^2$ $Var(Y) = 1$

$X \longrightarrow \frac{X - \mu}{\sigma} =: X'$

$X_1, X_2, \dots, X_n$ $x_1', x_2', \dots, x_n'$

$\bar{X}$, S_X $\bar{x'} = 0$, $S_{x'} = 1$

$$x_i \longrightarrow \frac{x_i - \bar{X}}{S_X} =: x_i' = -\frac{\bar{X}}{S_X} + \frac{1}{S_X} x_i = \lambda + \mu x_i$$

operazione che si fa per colonne

→ Per la regressione, questo significa $B_0' = 0$

e inoltre i B_j' sono confrontabili

$B_3' = 2,4$ $B_2' = -0,7$

x_3 o x_3' ha un impatto su Y più forte di x_2 o x_2'

$\alpha_3^* \ll \alpha_2^*$

VARIABILI CATEGORICHE

- la regressione per funzionare vuole tutte variabili numeriche "misurate"
- esempi di var non numeriche:

NOMINALI
- i. tipo di materiale {PET, PVC, VETRO}
- ii. fornitore {6 fornitori}
- iii. tipo di esca {4 categorie}

} non codificare

DICOTOMICA
- iv. sesso {m, f} ⟶ codificare come si vuole

ORDINALI
- v. titolo di studio {∅, III media, maturità, laurea, dottorato}
- vi. fascia di età {18-25, 25-35, ...}

} forse codificare

$x \xrightarrow{!} Y$

NON sono adatte alla regressione!

$T = T(x) + e$ $\quad x$: tipo di materiale {1 PET, 2 PVC, 3 VETRO}

Errore classico: codificare arbitrariamente le categorie con numeri

E(T)	x
9	PET
8	PVC
5	VETRO

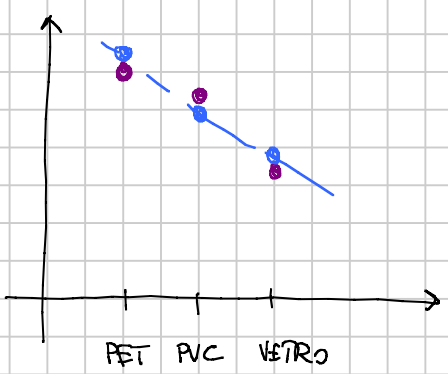
la dipendenza da x è una tabella, non una formula

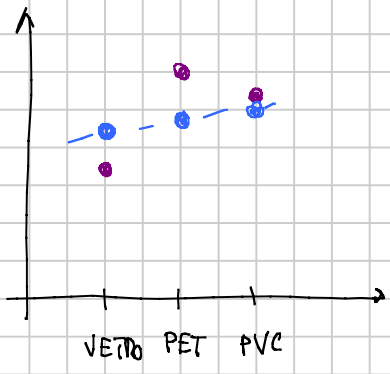
$T = \cancel{\alpha + \beta x + e} = \beta_x + e$ dove $\begin{cases} \beta_{PET} = 9 \\ \beta_{PVC} = 8 \\ \beta_{VETRO} = 5 \end{cases}$

c'è un coefficiente da stimare per ogni categoria

Errore: $T = \alpha + \beta x + e$ (x = 1, 2, 3)

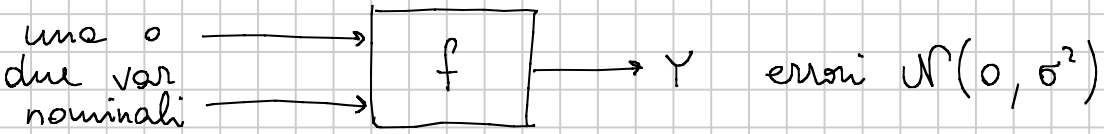
E(T)	x
$\alpha + \beta$	PET
$\alpha + 2\beta$	PVC
$\alpha + 3\beta$	VETRO





- Variabili dicotomiche (e ordinali) si possono (forse) codificare quelle nominali NO

★ Il modo corretto per gestire:



e basta

è usare la analisi della varianza (Cap. 10)

- Negli altri casi c'è una tecnica un po' fragile che si può usare:

Le var nominali (e forse certe ordinali) si possono "esplodere" in tante dicotomiche quanto è il numero di categorie meno 1

Es: PVC, PET, VETRO → 2 dicotomiche

↑ scelgo la preferita / default / più comune / più simile alle altre

id	x	x_{PVC}	x_{VETRO}
1	PVC	1	0
2	PVC	1	0
3	PET	0	0
.	PVC	1	0
.	VETRO	0	1
.	PET	0	0
.	VETRO	0	1
	...		
	..		

	x_1	x_2
PET	0	0
PVC	1	0
VETRO	0	1

★ Esplodere una dicotomica equivale a codificarla

MMSD (B)

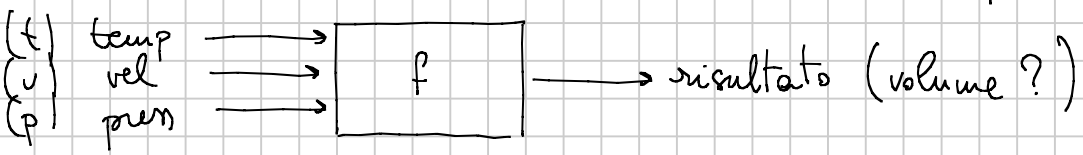
ora 19

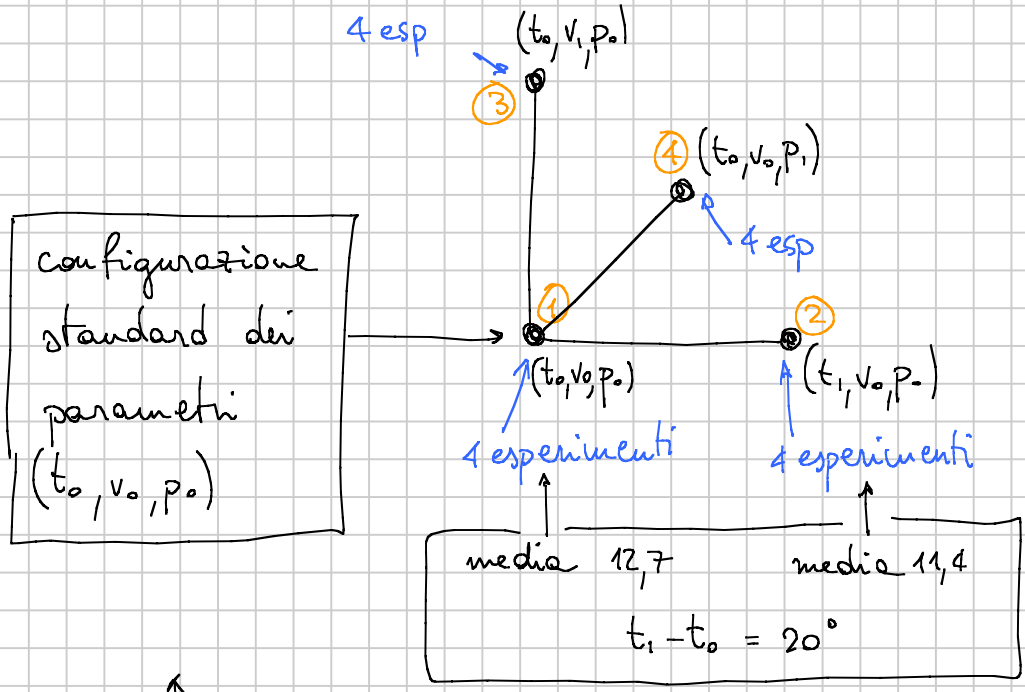
Note Title 13/04/2011

DESIGN OF EXPERIMENT (DoE) (Cap 10 Sleeper)

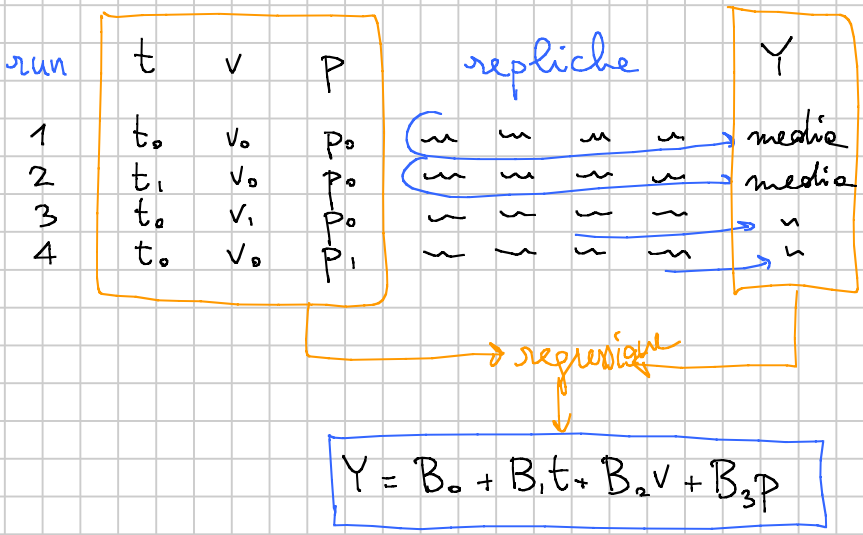
Se devo fare una regressione e devo ancora raccogliere i dati esiste un modo "migliore" per raccoglierli?

Esempio: stampo plastica: regolazioni → temperature, → velocità, → pressione } → risultato

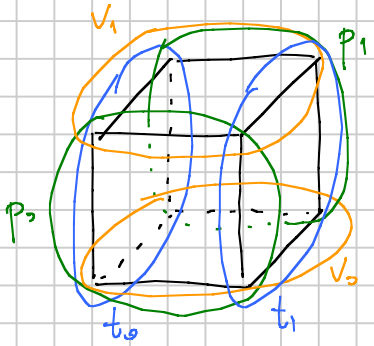


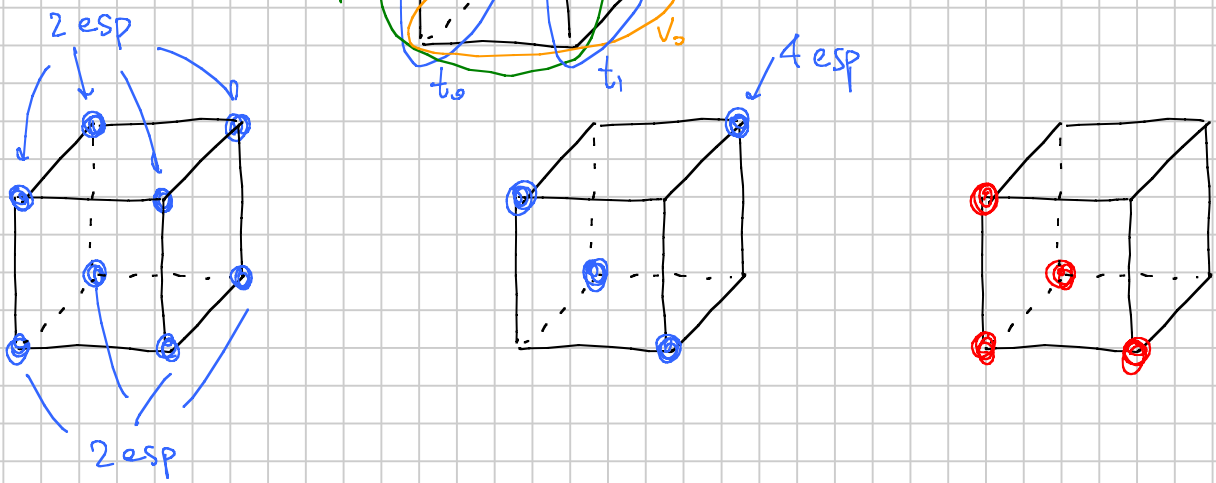


Alla fine riesco a ricavare $Y = B_0 + B_1 t + B_2 v + B_3 p$



Change one Setting a Time } n COST n



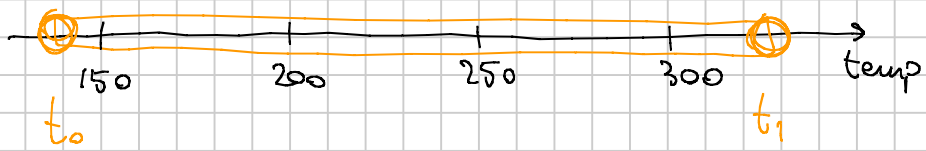


design completo design frazionario

- Come scegliere i valori di ingresso
  - Due livelli soli per ogni variabile di ingresso vanno quasi sempre bene
  - Si usano 3 livelli quando il modello deve essere fortemente non lineare
    → è possibile alla fine di una analisi su 2 livelli verificare con un ulteriore esperimento (center point) se sono sufficienti
    → con 2 livelli si prendono modelli lineari, anche con interazioni
    $Y = \beta_0 + \beta_1 t + \beta_2 v + \beta_3 p + \beta_4 tv + \beta_5 tp + \beta_6 vp + \beta_7 tvp$

non si arriva ai "quadrati"

- La scelta dei due livelli è critica



occorre sempre decidere in brainstorming con ingegneri e tecnici tenendo conto degli aspetti industriali pratici

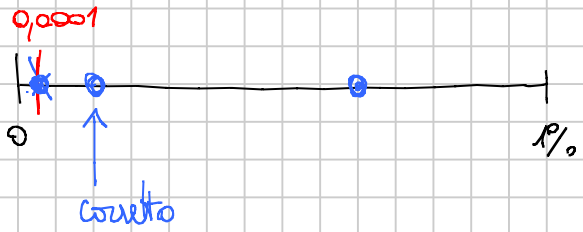
★ criteri statistici

a) se l'intervallo è troppo largo il modello può essere nonlineare

b) se l'intervallo è troppo largo è plausibile che il "sistema" non funzioni come previsto

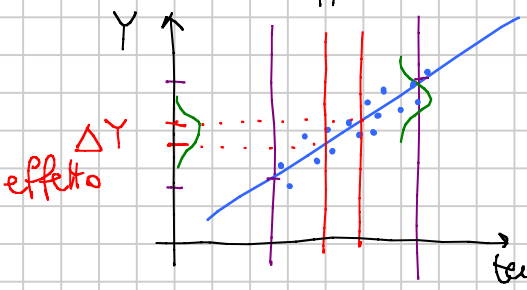
$t_0 = 200$, $t_1 = 250$ tutto ok

$t_0 = 150$, $t_1 = 300$ settaggi un pò al limite per la macchina qualcosa non funziona



(senti audio)

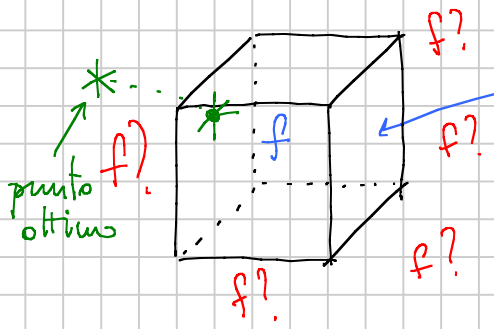
c) se l'intervallo è troppo stretto può darsi che io non veda variazioni apprezzabili



rapporto segnale/rumore insufficiente

rapporto segnale/rumore adeguato

d) se l'intervallo è troppo stretto il design space può essere inutile



volume del cubo si chiama DESIGN SPACE o MODEL SPACE

$Y = f(x_i)$ è valido solo all'interno del D.S.

ora 20

- Variabili di risposta
  - Possono essere più di una (facendo varie regressioni)
    Questo può portare a problemi nella scelta delle var di ingresso e nei loro livelli
  - Devono essere variabili numeriche misurabili con precisione su un range adeguato
    esempio: riempimento buono/non buono è dicotomica
    ↳ devo trovare un modo di misurarla
    grammatura? scala di qualità?
  → Se Y dicotomica è l'unica soluzione allora saranno necessarie centinaia o migliaia di esperimenti
- Variabili categoriche
  → tipo di plastica
  → tipo di iniettore

* solo come var di ingresso e solo con 2 categorie ciascuna (dicotomiche)
in ogni caso impediscono il center point

- Scelta delle var di ingresso e del design
  esempio: 11 variabili selezione delle var necessarie $\frac{n}{p}$ piccolo
  nel DOE siamo costretti comunque a $n = n(p)$
  → fattoriale completo $n = 2^p = 2^{11} = 2048$ run
  poi da moltiplicare per le repliche (almeno 2)
  → fattoriale frazionario $2^3 \to 2^2 = 4$ run $= 2^{p-k}$ k piccolo

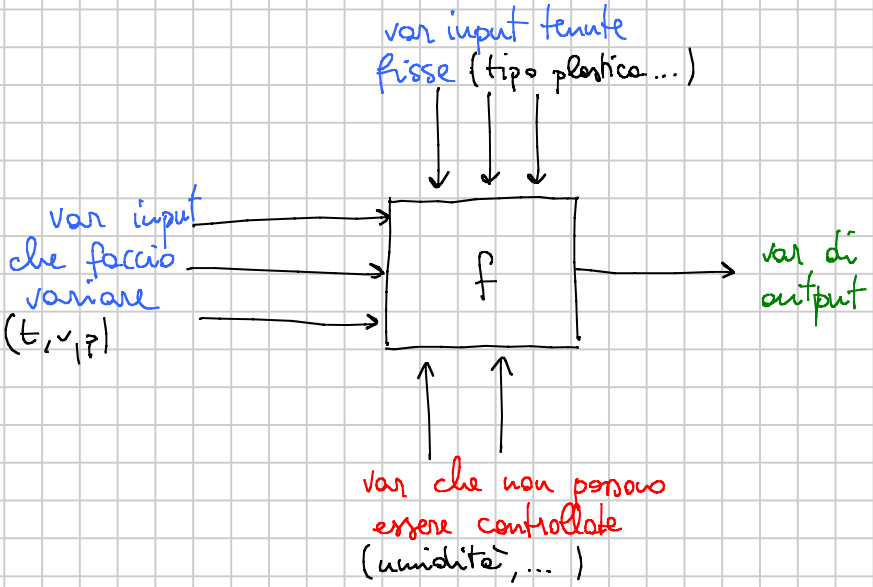
$k=1$ dimezzo le prove del completo

$k=2$ ... $\frac{1}{4}$

$k=3$ ... $\frac{1}{8}$

★ Le var di ingresso da far variare vanno scelte con cura, parlando con tutte le persone coinvolte

diagramma IPO



Le var da far variare vanno ridotte all'osso per tenere basso il numero degli esperimenti (vedi oltre esperimenti di screening / modeling)

DESIGN FRAZIONARI & ALIASING

Esempio: esperimento con 3 variabili : completo L8 e fraz L4

	A	B	C	~~A^2~~	AB	AC	BC	ABC
	-1	-1	-1	+1	+1	+1	+1	-1
•	-1	-1	+1	+1	+1	-1	-1	+1
•	-1	+1	-1	+1	-1	+1	-1	+1
	-1	+1	+1	+1	-1	-1	+1	-1
•	+1	-1	-1	+1	-1	-1	+1	+1
	+1	-1	+1	+1	-1	+1	-1	-1
	+1	+1	-1	+1	+1	-1	-1	-1
•	+1	+1	+1	+1	+1	+1	+1	+1

unità codificate : $t_0 = -1$ $t_1 = +1$ $V_0 = -1$ $V_1 = +1$ ..

I	A	B	C	AB	AC	BC	ABC
+1	-1	-1	-1	+1	+1	+1	-1
+1	-1	-1	+1	+1	-1	-1	+1
+1	-1	+1	-1	-1	+1	-1	+1
+1	-1	+1	+1	-1	-1	+1	-1
+1	+1	-1	-1	-1	-1	+1	+1
+1	+1	-1	+1	-1	+1	-1	-1
+1	+1	+1	-1	+1	-1	-1	-1
+1	+1	+1	+1	+1	+1	+1	+1

$= X$

design di Taguchi L8

X è ortogonale a meno di un fattore 8

$X^TX = 8\,I$ $\qquad (X^TX)^{-1} = \frac{1}{8} I$

L4 a 3 fattori

I	A	B	C	AB	AC	BC	ABC
+1	-1	-1	+1	+1	-1	-1	+1
+1	-1	+1	-1	-1	+1	-1	+1
+1	+1	-1	-1	-1	-1	+1	+1
+1	+1	+1	+1	+1	+1	+1	+1

A è in alias con BC

B è in alias con AC

C è in alias con AB

I è in alias con ABC

Alias structure:

3 I + ABC

3 A + BC

3 B + AC

3 C + AB

dell' L4 a 3 fattori

Se faccio la regressione con I, A, B, C e non gli altri ....

$Y = \beta_0 + \beta_1 A + \beta_3 C + \beta_2 AC \rightarrow B_0(1+ABC) + B_1(A+\cancel{BC}) + B_2(B+AC) + B_3(C+\cancel{AB})$

$Y = B_0 + B_1 A + B_3 C + B_2 B$

A B AC C

.... trovo 4 coefficienti che sarei erroneamente portato ad interpretare come quelli di I, A, B, C mentre invece sono quelli di I, A, AC, C

Y non dipende davvero da B, ma non ho modo di capirlo

- Nell' alias structure ho tante righe quante le run = coefficienti stimabili e tanti termini per riga quanto 2^k

"design di IV risoluzione"

Fractional Factorial Design → $2^6 = 64$ run il completo

Factors:	6	Base Design:	6, 16	Resolution:	IV
Runs:	16	Replicates:	1	Fraction:	1/4
Blocks:	1	Center pts (total):	0		

$k = 2$ $16 = 2^{6-2} = 2^4$ 2^{-k}

Design Generators: E = ABC, F = BCD

Alias Structure

4	I + ABCE + ADEF + BCDF
4	A + BCE + DEF + ABCDF
4	B + ACE + CDF + ABDEF
4	C + ABE + BDF + ACDEF
4	D + AEF + BCF + ABCDE
4	E + ABC + ADF + BCDEF
4	F + ADE + BCD + ABCEF
4	AB + CE + ACDF + BDEF
4	AC + BE + ABDF + CDEF
4	AD + EF + ABCF + BCDE
4	AE + BC + DF + ABCDEF
4	~~AF~~ + DE + ABCD + BCEF
4	BD + CF + ABEF + ACDE
4	BF + CD + ABDE + ACEF
6	ABD + ACF + BEF + CDE
6	ABF + ACD + BDE + CEF

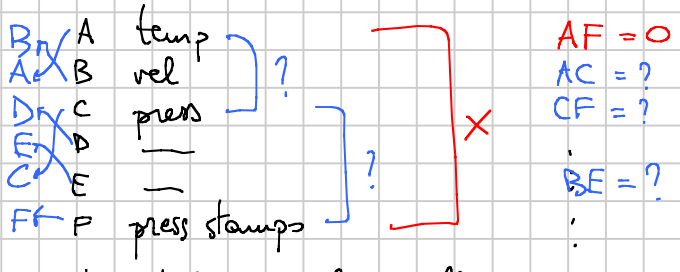
L 16
16 righe
16 run
16 coefficienti stimabili

2 coefficienti inutili

Figure 10-53 Alias Structure for an $L16$ Treatment Structure with Six Factors

Adeguato (forse eccessivo) per determinare gli effetti singoli

solo in parte adeguato per determinare le interazioni



Risoluzione: il minimo tra tutti i coeff della somma dei gradi dei due effetti di grado più basso in alias tra di loro

Available Factorial Designs (with Resolution)

	Factors													
Runs	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	Full	III												
8		Full	IV	III	III	III								
16			Full	V	IV	IV	IV	III	III	III	III	III	III	III
32				Full	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV
64					Full	VII	V	IV	IV	IV	IV	IV	IV	IV
128						Full	VIII	VI	V	V	IV	IV	IV	IV

scuro = III ris
giallo = IV ris
verde ≥ V ris

- Siccome di solito si considerano importanti solo gli effetti singoli e le interazioni a due

III : $A + BC + \dots$ singoli in alias con interazioni a due

V : $A + BCDE + \dots$, $AB + CDE + \dots$ } singoli da soli; inter a due da sole

es : $p = 11$ $2^{11} = 2048$ si può fare il fractional con $k = 4$

$2^{11-4} = 2^7 = 128$ L128 è di V ris per 11 fattori

- EFFICIENZA DEI DESIGN

Fractional Factorial Design

```
Factors:    6    Base Design:          6, 16    Resolution:    IV
Runs:      16    Replicates:               1    Fraction:     1/4
Blocks:     1    Center pts (total):       0

Design Generators: E = ABC, F = BCD

Alias Structure

1  I + ABCE + ADEF + BCDF

1  A + BCE + DEF + ABCDF
1  B + ACE + CDF + ABDEF
1  C + ABE + BDF + ACDEF
1  D + AEF + BCF + ABCDE
1  E + ABC + ADF + BCDEF
1  F + ADE + BCD + ABCEF
1  AB + CE + ACDF + BDEF
1  AC + BE + ABDF + CDEF
½  AD + EF + ABCF + BCDE
½  AE + BC + DF + ABCDEF
½  AF + DE + ABCD + BCEF
½  BD + CF + ABEF + ACDE
½  BF + CD + ABDE + ACEF
0  ABD + ACF + BEF + CDE
0  ABF + ACD + BDE + CEF
```

Figure 10-53 Alias Structure for an $L16$ Treatment Structure with Six Factors

O : E.O.I. effect of interest

universale : effetti singoli + interazioni a due

ottimale : sottoinsieme ✕ vanno tolti quelli ritenuti irrealistici

$\frac{10}{16} = 0{,}625$ $\frac{13}{16} = 0{,}8125$

$$eff = \frac{1}{run} \sum_{i=1}^{run} \begin{cases} 1 \\ \frac{1}{2} \\ 0 \end{cases}$$

1 eoi nella run

2 eoi " "

altrimenti

Table 10-9 Modeling Efficiency for Two-Level Factorial Treatment Structures

Treatment Structure	Number of Factors						
	3	4	5	6	7	8	9
$L4$	0.375						
$L8$	0.875	0.813	0.500	0.125	0.125		
$L16$		0.688	1.000	0.625	0.500	0.563	0.313
$L32$			0.500	0.688	0.766	0.781	0.750
$L64$				0.344	0.453	0.578	0.641
$L128$					0.227	0.289	0.359

- Esperimenti di screening

  permettono di screenare un elevato numero di variabili con pochi esperimenti (non poco maggiore di p)

  la III risoluzione è adeguata, gli alias tipo A+BC portano solo ad includere qualche var di troppo (falsi positivi)

* Esperimenti di modeling

  cercano un modello preciso → servono (alcune) interazioni

  ↳ V risolu7 o IV strutturata bene ... efficienza ...

  sono fattibili se le variabili sono poche

* Per esperimenti di screening, ancora meglio dei Taguchi di III ci sono i design di PLACKETT - BURMAN che non sono affetti da aliasing ma da confounding.

  (Meno falsi positivi.)

MMSD (B)

ora 22

Note Title 20/04/2011

DOE → gestione variabili
→ struttura design

- Repliche

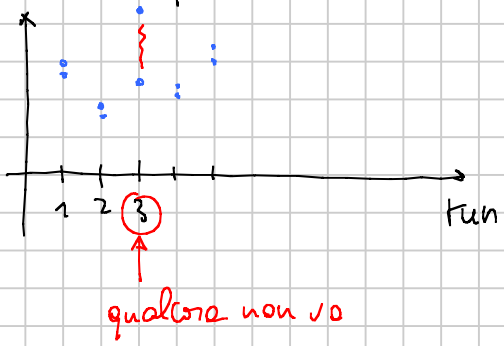
5 fattori → L16 : 16 run (su 32 combinazioni possibili)

16 run = 16, 32, 48, 64 ... esperimenti a seconda di quante repliche faccio per ogni run

run	A	B	C	D	E	rep1	rep2	rep3	Y
1	-1	-1	+1	-1	-1	11,4	11,7	10,8	media

→ Perché servono diverse repliche?

a) Per scoprire errori e outlier servono sempre almeno 2 repliche



b) Per alzare il rapporto segnale/rumore

rep 1 $\sim N(\mu, \sigma^2)$ media(3 repliche) $\sim N(\mu, \frac{\sigma^2}{3})$ ← varianza minore

Al diminuire del rumore diventano rilevabili via via effetti più piccoli

$$Y = 7,4 + \underset{6,8}{2,3}\,A + e \qquad e \sim N(0, (3,1)^2)$$

→ Per determinare scientificamente il numero di repliche serve conoscere:

- σ il rumore dell'esperimento
- ordine di grandezza degli effetti che voglio poter rilevare
- probabilità di rilevare un simile effetto (= potenza del test)

→ Siccome spesso non si conosce σ, esiste la regola empirica:

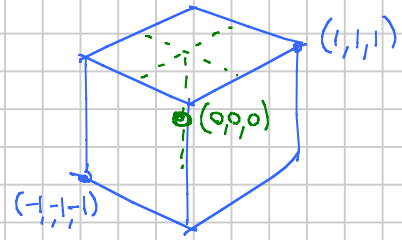
$$\left\lceil \frac{run + 32}{run} \right\rceil = \text{repliche per run}$$

L4	9 repliche
L8	5 "
L16	3 "
L32	2 "
L64	2 "
⋮	⋮ "
⋮	⋮ "

} per i Taguchi

- Center point

  È una run aggiuntiva



run	A	B	C	D	E	rep1	rep2	rep3	media
⋮	±1	⋮	⋮	⋮	⋮	~	~	~	~
17	0	0	0	0	0	~	~	~	~

→ stesso numero di repliche

→ si fa la regressione sulle altre run, lasciandolo fuori

→ si controlla se il valore osservato per il cp è compatibile con il modello di regressione trovato

$\hat{Y} = B_0 + B_1 x_1 + \dots + B_p x_p$ valore previsto generico

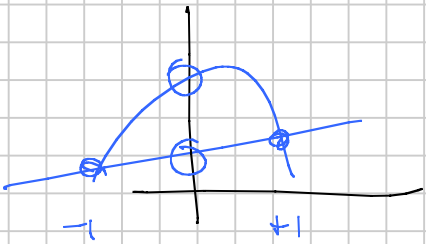
$\hat{Y} = B_0$ valore previsto per il c.p.

confronto $B_0 \pm Se$ o $B_0 \pm 2Se$ con il valore misurato

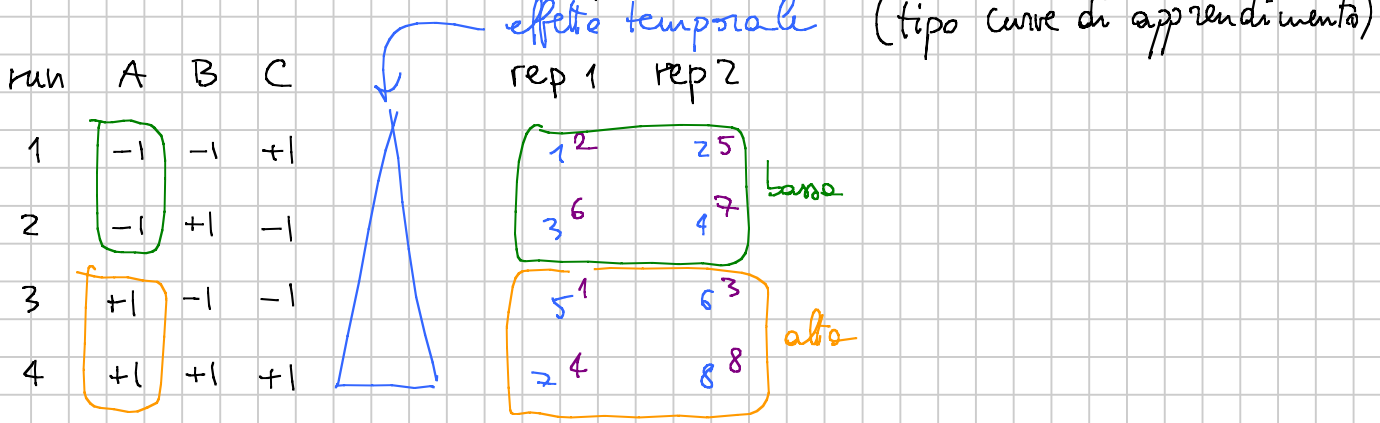
→ se è compatibile, il modello va bene

se non è compatibile mi dice che il fenomeno presenta forti nonlinearità

★ Va rifatto tutto il DOE passando a 3 livelli.



- Randomizzazione delle repliche



Se faccio la regressione, A viene significativa, importante
Se faccio le repliche in ordine casuale, risolvo il problema
→ Occhio che la rand. costa, quindi si trovano resistenze

- Ortogonalità

Per evitare di perderla, la matrice delle risposte deve essere completa, senza buchi, anche dopo aver tolto gli outliers
↳ casomai si rifanno i singoli esperimenti problematici
↳ occhio ai difetti cronici del design

- Regressione vera e propria
↳ 1 punto per ogni run → Y = media (repliche)
↳ oppure 1 punto per ogni replica → Y = singola replica

run	rep	A	B	C	Y
1	1	-1	-1	+1	~
1	2	-1	-1	+1	~
2	1	-1	+1	-1	~
2	2	-1	+1	-1	~
-	-	-	-	-	-

★ Si trovano gli stessi identici valori per i coefficienti di regressione
★ Cambia se e la potenza dei test

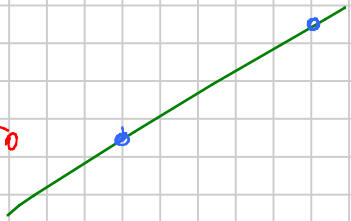
↳ $\sim \frac{\sigma}{\sqrt{r}}$ $\sim \sigma$ r : # di repliche

L16 5 fattori

1	I + ABCDE
2	A + BCDE
3	B + ACDE
.	C + ABDE
.	D ...
.	E ...
.	AB + CDE
.	AC + BDE
	AD + ...
	AE .
	BC .
	BD .
	BE .
	CD .
	CE .
16	DE .

16 coefficienti con 16 punti

residui nulli → Se non definito



★ Basta togliere qualche variabile

ora 23

● Selezione delle variabili

Si fa in modo completamente diverso dal solito.

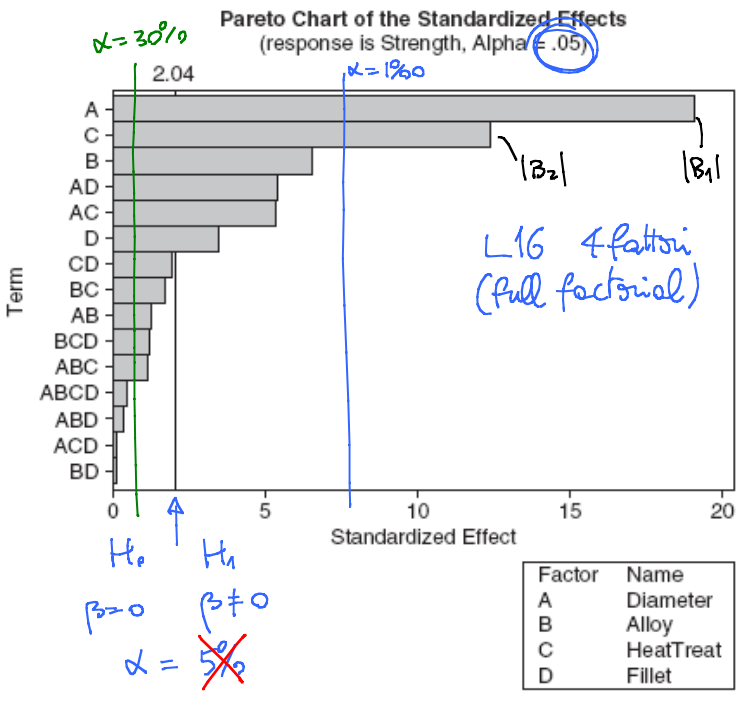


Figure 10-31 Pareto Chart of Effects

Si fa il diagramma di Pareto dei valori assoluti dei coefficienti e si analizza qualitativamente

$\frac{B_i}{Se\sqrt{\Sigma_{ii}}} = stat \qquad H_0: \beta_i = 0$

$|B_i|$

$Y = B_0 + B_1 A + B_2 C + B_3 B + \dots$

A, C A, C, B A, C, B, AD, D, AC

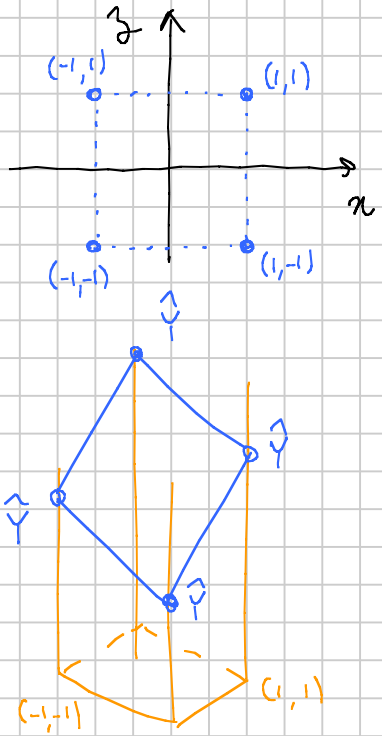
★ Posso togliere le variabili che non mi interessano tutte insieme perché il design è ortogonale e rilanciare la regressione

Ottengo solo una stima più giusta di σ, con un Se più preciso

Termini di interazione

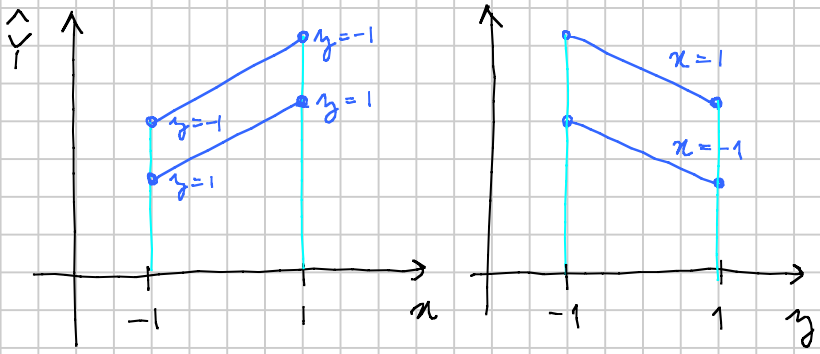
$$\hat{Y} = B_0 + B_1 x + B_2 y + B_3 xy$$

↑ termine di interazione

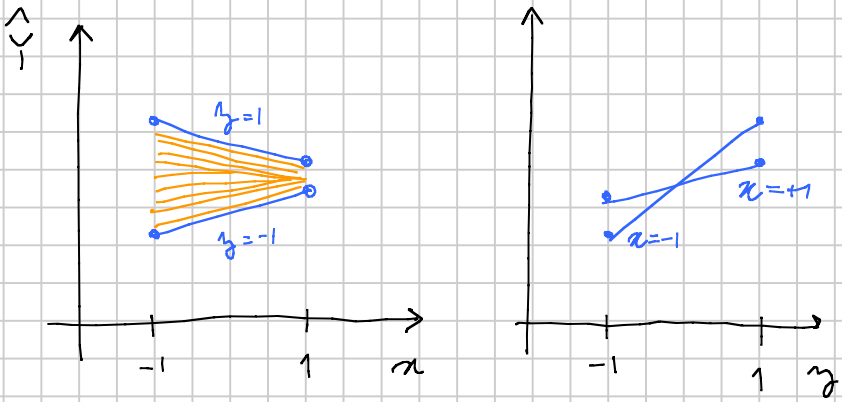


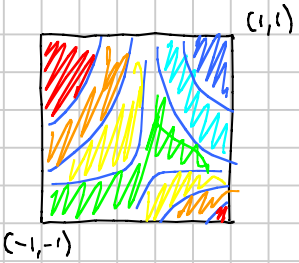
caso $B_3 = 0$

$\hat{Y} = B_0 + B_1 x + B_2 y$



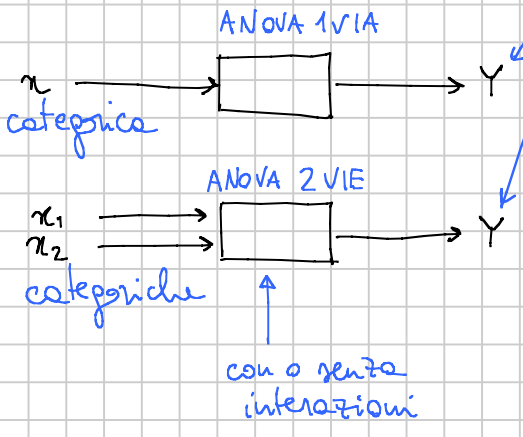
caso $B_3 \neq 0$





contour plot

ANALISI DELLA VARIANZA (Cap 10 Ross)



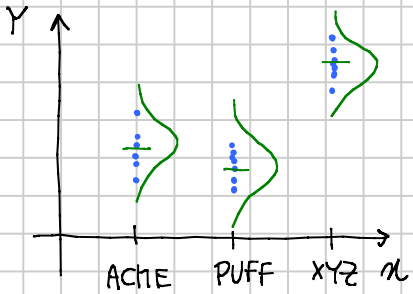
↙ ANOVA omosch

- Y: % di polvere < 5 μm
- x: fornitore lattosio: 3 ditte

- Y: consumi di carburante
- x_1: diesel normale, 3 varianti sperimentali (4 carburanti)
- x_2: tipo di auto: utilitaria, berlina, suv

★ Si usa quando le categorie sono 2–5 (mai 10+)

ANOVA A 1 VIA



$$Y = f(x) = a + bx + cx^2 \cdots$$

$$= \begin{cases} \mu_{ACME} & x = ACME \\ \mu_{PUFF} & x = PUFF \\ \mu_{XYZ} & x = XYZ \end{cases}$$

$\mathcal{N}(0, \sigma^2)$ → errore

$$Y = \mu_x + e$$

- Struttura dei dati : m campioni di lunghezze variabili

$$\{ACME, PUFF, XYZ\} = \{1, 2, \dots, m\}$$

$Y_{1,1}, Y_{1,2}, \dots Y_{1,n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ iid $\quad \mu_1 \approx Y_{1,*} := \frac{1}{n_1}\sum_i Y_{1,i} \quad \sigma^2 \approx S_1^2 = \frac{1}{n_1 - 1}\sum_i (Y_{1,i} - Y_{1,*})^2$

$Y_{2,1}, Y_{2,2}, \dots, Y_{2,n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$ iid $\quad \mu_2 \approx Y_{2,*} := \dots \quad \sigma^2 \approx S_2^2 = \dots$

$\dots\dots$

$Y_{m,1}, Y_{m,2}, \dots, Y_{m,n_m} \sim \mathcal{N}(\mu_m, \sigma^2)$ iid $\quad \mu_m \dots \quad \sigma^2 \approx \dots$

ora 24

$Y_{i,*} \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right)$ stimatore μ_i corretto e consistente → HWC

Ho poi m stimatori diversi e indipendenti per σ^2

S_i^2 : $\quad \frac{S_i^2}{\sigma^2}(n_i - 1) \sim \chi^2(n_i - 1)$

Voglio riassumerli in uno stimatore solo per σ, migliore di tutti

$S_M^2 = \theta_1 S_1^2 + \theta_2 S_2^2 + \dots + \theta_m S_m^2 \qquad \theta_1 + \theta_2 + \dots + \theta_m = 1$

media pesata $\qquad$ pesi

i. $E(S_M^2) = \sigma^2$

ii. $Var(S_M^2)$ dipende dai θ_i → il valore minimo si ha quando

$$\theta_i = \frac{gdl_i}{gdl_{TOT}} = \frac{n_i - 1}{\sum_j (n_j - 1)}$$

iii. Quando i θ_i sono questi si usa il simbolo S_p^2, si

diamo stimatore pooled della varianza e inoltre:

$$\frac{S_p^2}{\sigma^2} \cdot gdl_{TOT} \sim \chi^2(gdl_{TOT})$$

HWC

$$S_p^2 := \frac{n_1-1}{N-m} S_1^2 + \frac{n_2-1}{N-m} S_2^2 + \dots + \frac{n_m-1}{N-m} S_m^2$$

dove $N = \sum_{i=1}^{m} n_i$ e quindi $gdl_{TOT} = \sum_i (n_i - 1) = N - m$

★ $S_p^2 = S_W^2$ varianza "within" (=entro i campioni)

★ S_B^2 varianza "between" (=tra i campioni)

campione 1 $\longrightarrow Y_{1,*}$

" 2 $\longrightarrow Y_{2,*}$

.....

$\longrightarrow Y_{m,*}$

} faccio la varianza campionaria

$$S_B^2 := \frac{1}{m-1} \sum_{j=1}^{m} n_j (Y_{j,*} - Y_{*,*})^2$$

dove $Y_{*,*} = \frac{1}{N} \sum_{i,j} Y_{i,j} = \frac{1}{N} \sum_{j=1}^{m} n_j Y_{j,*}$ media globale

★ Se Y non dipende da x, ovvero se $\mu_1 = \mu_2 = \dots = \mu_m$, allora si può verificare (HWC) che

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

Se Y dipende da x, S_B^2 è tipicamente più grande

→ Inoltre S_B^2 è sempre indipendente da S_W^2

- L'anova a 1 via formalmente esegue il test:

$H_0: \mu_1 = \mu_2 = \dots = \mu_m$ (Y non dipende da x) H_1: non tutte uguali (Y dipende da x)

il test si basa su S_B^2: se è grande è un indizio per H_1

in concreto calcolo la statistica:

$$F := \frac{S_B^2}{S_W^2}$$

se è vera H_1 F tipicamente > 1

se è vera H_0 F vicina a 1

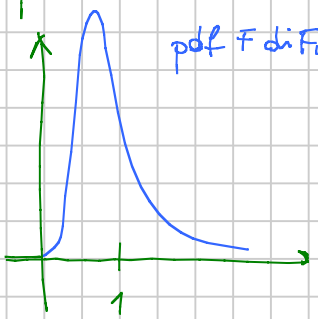
★ distribuzione sotto H_0: se S_B^2 e S_W^2 sono entrambi stimatori di σ^2 basati su χ^2 con a e b gdl, allora

$$\frac{S_B^2}{S_W^2} \sim F(a; b)$$

a: gdl numeratore; b: gdl denominatore; F di Fischer

F di Fischer

Classe di v.a. continue, concentrate intorno a 1, positive, con 2 parametri

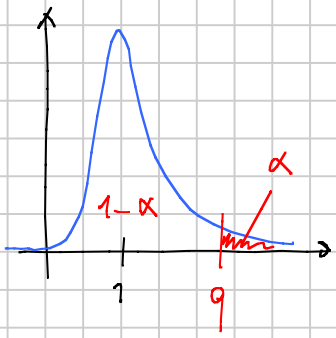


Se S_1^2, S_2^2 sono indip con $\frac{S_1^2}{\sigma^2} a \sim \chi^2(a)$ $\frac{S_2^2}{\sigma^2} b \sim \chi^2(b)$

Allora $\frac{S_1^2}{S_2^2} \sim F(a; b)$

$$F := \frac{S_B^2}{S_W^2} \overset{H_0}{\sim} F(m-1; N-m)$$

sotto H_1 F è più grande



$F > q$ dico H_1

$F \leq q$ dico H_0

★ Altra inferenza:

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

$$\mu_i \approx Y_{i,*} \qquad \frac{Y_{i,*} - \mu_i}{S_W / \sqrt{n_i}} \sim t(N-m)$$

● Identità utili

$$S_W^2 = \frac{1}{N-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{i,j} - Y_{i,*}\right)^2 \qquad \text{(HWC)}$$

$$S_W^2,\ S_B^2,\ S_Y^2 = \frac{1}{N-1} \sum_{i,j} \left(Y_{i,j} - Y_{*,*}\right)^2 \qquad \text{varianze}$$

$SS_W,\ SS_B,\ SS_Y$ devianze

$$SS_W = S_W^2(N-m) \qquad SS_B = S_B^2(m-1) \qquad SS_Y = S_Y^2(N-1)$$

$$SS_Y = SS_B + SS_W \qquad \text{identità delle devianze:}$$

MMSD (B)

ora 25

Note Title 04/05/2011

Info pratiche:

- oggi solo 2 ore
- domani scambio Nicolodi / Morandin (laboratorio 8:30 - 10:30, Nicolodi 14:30 - 16:30) → aula Loli!
- domani niente ricevimento (per cose brevi e urgenti, subito dopo lezione)
- mercoledì 11 maggio niente lezione

ANOVA A 1 VIA

- Cfr analisi
- Se $m = 2$ $\quad Y_{1,1}, Y_{1,2}, \dots Y_{1,n_1}$; $Y_{2,1}, Y_{2,2}, \dots, Y_{2,n_2}$

$$\downarrow \qquad\qquad \downarrow$$

$$N(\mu_1, \sigma^2) \qquad N(\mu_2, \sigma^2)$$

Anova a 1 via: $H_0: \mu_1 = \mu_2$ $\qquad H_1: \mu_1 \neq \mu_2$

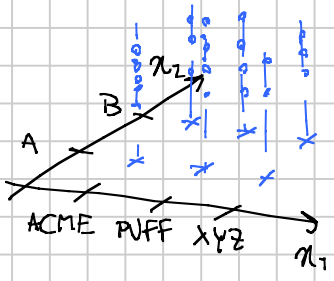
stesse ipotesi di lavoro del test per il confronto delle medie di due popolazioni normali

$$\frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad \frac{Y_{1,\cdot} - Y_{2,\cdot}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1 + n_2 - 2)$$

☆ Sono lo stesso test: il p-dei-dati viene identico.

- Vedi ora 20 anno scorso: test globale di regressione

ANOVA A DUE VIE



x_2: tipo di mescolamento

Y: % < 5 μm dopo mescolato

x_1 \ x_2	1	2
1	$x_{1,1,1}, x_{1,1,2}, \dots$	$x_{1,2,1}, x_{1,2,2}, \dots$
2	...	...
3		$x_{3,2,1} \dots$

$x_{i,j,k}$ dato k-esimo dell'esperimento fatto con $x_1 = i$ e $x_2 = j$

$x_{i,j,1}, x_{i,j,2}, \dots$ sono repliche

- In ogni casella serve lo stesso numero di repliche se no non funziona.
  - → se è possibile ricondursi a numero di repliche uguali buttando via non troppi dati, è utile farlo (se no, regressione)
  - → se una casella è vuota o ha comunque troppo pochi dati si può valutare se eliminare la riga o la colonna

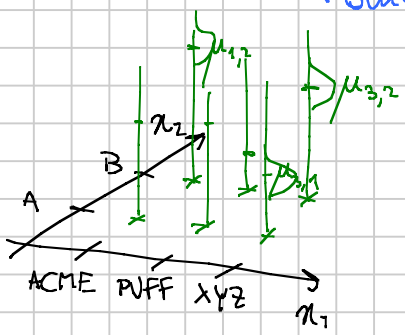
- $i = 1, 2, \dots, m$ valori di x_1
  $j = 1, 2, \dots, n$ valori di x_2
  } var categoriche
  $k = 1, 2, \dots, l$ repliche
  → $l = 1$ anova 2 vie classica
  → $l > 1$ anova a 2 vie con repliche/con interazioni

  → NB Se $l > 1$ posso ottenere info in più, ma comunque posso anche ricondurmi a $l = 1$ facendo la media su ogni casella.

SENZA REPLICHE ($l = 1$)

$$x_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2)$$

omoschedastico



ipotesi fondamentale di linearità:

$$\mu_{i,j} = \mu + \alpha_i + \beta_j$$

- μ: media globale
- α_i: contributo dovuto ad x_1
- β_j: contributo dovuto ad x_2

B	72	55	61 ←
A	65	48	54
	ACME	PUFF	XYZ

← forzati dal modello lineare

se è troppo restrittivo, serve un modello con interazioni : $\ell > 1$

→ ipotesi tecnica (non restrittiva)

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

si usa (HWC)

- $X_{ij} \sim \mathcal{N}(\mu + \alpha_i + \beta_j ; \sigma^2)$

$$X_{i,*} = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \sim \mathcal{N}\left(\mu+\alpha_i ; \frac{\sigma^2}{n}\right) \qquad X_{*,j} = \frac{1}{m}\sum_{i=1}^{m} X_{i,j} \sim \mathcal{N}\left(\mu+\beta_j ; \frac{\sigma^2}{m}\right)$$

$$X_{*,*} = \frac{1}{nm}\sum_{i,j} X_{ij} = \frac{1}{m}\sum_{i} X_{i,*} = \frac{1}{n}\sum_{j} X_{*,j} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{mn}\right)$$

ora 26

- Stimatori

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

HWC ★

$$\alpha_i \approx X_{i,*} - X_{*,*} \sim \mathcal{N}\left(\alpha_i ; \sigma^2 \frac{m-1}{mn}\right) \quad \text{effetto riga} \quad i = 1,2,\dots,m$$

$$\beta_j \approx X_{*,j} - X_{*,*} \sim \mathcal{N}\left(\beta_j ; \sigma^2 \frac{n-1}{mn}\right) \quad \text{effetto colonna} \quad j = 1,2,\dots,n$$

$$\mu+\alpha_i \approx X_{i,*} \sim \mathcal{N}\left(\mu+\alpha_i ; \frac{\sigma^2}{n}\right) \quad \text{risp. media riga} \quad i = 1,\dots,m$$

$$\mu+\beta_j \approx X_{*,j} \sim \mathcal{N}\left(\mu+\beta_j ; \frac{\sigma^2}{m}\right) \quad \text{risp. media colonna} \quad j = 1,\dots,n$$

$$\mu_{ij} = \mu+\alpha_i+\beta_j \approx X_{i,*} + X_{*,j} - X_{*,*} \sim \mathcal{N}\left(\mu_{ij} ; \sigma^2 \frac{m+n-1}{mn}\right) \quad (i,j)\dots$$

μ_{ij}: media che mi aspetto (i,j) — previsto (i,j)

HWC ★★

$$R_{ij} := X_{ij} - \left(X_{i,*} + X_{*,j} - X_{*,*}\right) \sim \mathcal{N}\left(0 ; \sigma^2 \frac{(m-1)(n-1)}{mn}\right) \quad (i,j)\dots$$

residuo (i,j) ovvero osservato meno previsto

$$SS_e = \cancel{SS_R} = \sum_{ij} R_{ij}^2$$

R area confusione

$$S_e^2 := \frac{SSe}{(m-1)(n-1)} \approx \sigma^2$$

$$\frac{S_e^2}{\sigma^2}(m-1)(n-1) = \frac{SSe}{\sigma^2} \sim \chi^2\left((m-1)(n-1)\right)$$

$$\sum_{i,j} \frac{(X_{ij} - \mu_{ij})^2}{\sigma^2} \sim \chi^2(mn) \qquad \text{per def di } \chi^2$$

gdl da togliere

$$\sum_{i,j} \frac{\left[X_{ij} - \left(X_{i*} + X_{*j} - X_{**}\right)\right]^2}{\sigma^2} = \sum_{i,j} \frac{R_{ij}^2}{\sigma^2} = \frac{SSe}{\sigma^2} \sim \chi^2\left(mn - (m+n-1)\right)$$

$$\sim \chi^2\left((m-1)(n-1)\right)$$

★ $\sigma^2 \approx S_e^2$ stimatore che funziona sempre

● Test fondamentali dell'anova a 2 vie

→ Effetto riga: Y dipende da x_1 ?

→ Effetto colonna: Y dipende da x_2 ?

analogamente alla regressione (per la selezione delle variabili)

★ Eff riga: $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$ (Y non dipende da x_1) $\quad H_1$: non tutti nulli (Y dipende da x_1)

$$F_R := \frac{S_R^2}{S_e^2} \overset{H_0}{\sim} F\left(m-1; (m-1)(n-1)\right)$$

↳ piccola dico H_0

↳ grande dico H_1

dove $S_R^2 := \frac{SS_R}{m-1}$ varianza per righe

$SS_R := \sum_{i=1}^{m} n\left(X_{i*} - X_{**}\right)^2$ devianza per righe

★ Eff colonna: $H_0: \beta_1 = \dots = \beta_n = 0$ (Y non dipende da x_2) $\quad H_1$: non tutti nulli (Y dipende da x_2)

$$F_C := \frac{S_C^2}{S_e^2} \overset{H_0}{\sim} F\left(n-1; (m-1)(n-1)\right)$$

↳ piccola dico H_0

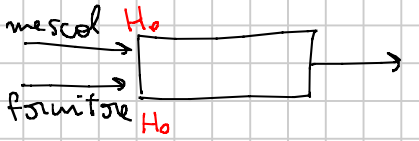
↳ grande dico H_1

dove $S_C^2 := \frac{SS_C}{n-1}$ varianza per colonne

$SS_C := \sum_{j=1}^{n} m\left(X_{*j} - X_{**}\right)^2$ devianza per colonne

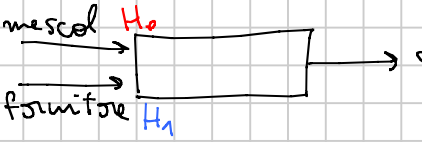
- Se una delle due var risulta non significativa, può a volte essere rimossa (ovvero si rifà l'analisi come ANOVA a 1 via)

→ qui mi fermo



→ qui rifaccio come ANOVA 1 via

→ Mi interessa rifare il test perché spero di trovare un H_1



→ qui rifaccio come ANOVA 1 via

→ Non mi interessa rifare il test (è già H_1), ma posso trovare stime più precise dei parametri e dei previsti

guardare il test t-paired nel Cap 8 del Ross

* Senti anche ora 27

MMSD (B) ora 27

Note Title 18/05/2011

★ Settimana prossima niente lezioni
Ultima volta teoria : 1 giugno
Ultima volta lab : da definire, ma quella settimana

● Legami ANOVA test classici

ANOVA 1 via con 2 campioni soli $\Leftrightarrow$ test $\dfrac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$

ANOVA 2 vie con $m=2$, n qualsiasi $\Leftrightarrow$ test t-Paired

$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$
$Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma^2)$

$X_i \sim \mathcal{N}(\mu_i, \sigma_X^2) \quad Y_i \sim \mathcal{N}(\mu_i+\delta, \sigma_Y^2)$

ad esempio n persone pesate prima e dopo una dieta

$Z_i := Y_i - X_i \sim \mathcal{N}(\delta, \sigma_Z^2)$
$Z_1, Z_2, \dots, Z_n \sim \mathcal{N}(\delta, \sigma_Z^2)$

$H_0: \delta = 0 \qquad H_1: \delta \neq 0$

$$\frac{\bar{Z}-\mu_0}{S_Z/\sqrt{n}} \overset{H_0}{\sim} t(n-1)$$

★ In entrambi i casi i test sembrano differenti, ma si ottiene esattamente lo stesso p dei dati.

▨ CON REPLICHE $(\ell \geq 2)$

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

$\mu_{i,j} = \mu + \alpha_i + \beta_j$ NO REPLICHE
$\mu_{i,j}$ qualsiasi REPLICHE

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

↑ scostamento dal modello lineare = interazione (effetto simile ai termini $x_i x_j$ nella regressione e nel DoE)

- Test che vengono effettuati:

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

se dice H_0 si può fare la media in ogni casellina e passare all'anova a 2 vie

$$X_{ij} := X_{ij*} = \frac{1}{e}\sum_{k=1}^{e} X_{ijk} \sim \mathcal{N}\left(\mu+\alpha_i+\beta_j, \frac{\sigma^2}{e}\right)$$

↳ ANOVA 2 VIE SENZA REPLICHE

attenzione che questo modello ha una varianza fittizia

→ Se dice H_1:

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

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

---

TEST DI ADATTAMENTO

→ Test del chi-quadro elementare
→ Test del chi-quadro con stima di parametri
→ Tabelle di contingenza → varie questioni

} test di adattamento ad una distribuzione

TEST χ^2 ELEMENTARE

esempio: va. discreta (→ incluse le categoriche) con una legge teorica nota

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

dati Istat → 4% 75% 7% 14%

Prendo un campione forse ha legge diversa : residenti a Feriso

$n = 300$ $\quad 6_{2\%} \quad 183_{61\%} \quad 60_{20\%} \quad 51_{17\%}$

Q: questo campione può sensatamente venire da quella distribuzione (H_0), oppure vi è evidenza che la distrib sia diversa (H_1) ?

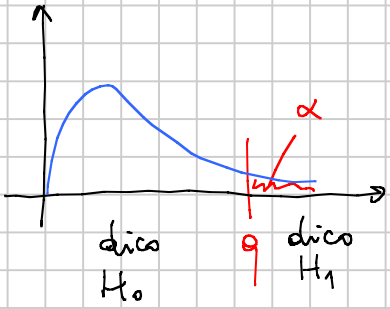
ora 28

$$\frac{(6-12)^2}{12} + \frac{(183-225)^2}{225} + \frac{(60-21)^2}{21} + \frac{(51-42)^2}{42} \rightarrow \text{statistica}$$

($12 = 4\% \cdot 300$, $225 = 75\% \cdot 300$, $21 = 7\% \cdot 300$, $42 = 14\% \cdot 300$)

↳ se è grande dirò H_1, se è piccola dirò H_0

↳ uso quantili χ^2 con 3 (categorie − 1) g.d.l.



$q = \text{INV.CHI}(\alpha\,;\,3)$

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

H_0: $P(X=1) = \varphi_0(1)$, $P(X=2) = \varphi_0(2)$, …, $P(X=k) = \varphi_0(k)$

H_1: non tutte uguali

→ Si chiama test di adattamento a distribuzione discreta completamente specificata

→ Statistica del test

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

dove O_i : osservati → quanti del campione $X_1, X_2, \dots, X_n$ sono i

dove A_i: atteso $\rightarrow$ $A_i = n \cdot \varphi_0(i)$

↳ NB. Non sempre gli A_i sono interi, non occorre che lo siano

$\rightarrow$ $W \dot{\sim} \chi^2(k-1)$ sotto H_0

W assume valori maggiori sotto H_1

✱ L'approssimazione è buona quando gli attesi A_i sono tutti numeri grandi

↳ n grande
↳ k piccolo
↳ φ_0 abbastanza uniforme

} tutti aspetti auspicabili

$\rightarrow$ "Rule of thumb": tutti gli $A_i > 1$ e almeno i $\frac{4}{5}$ degli $A_i > 5$

✱ Quando l'approssimazione non è buona, si dovrebbe usare la vera legge di W sotto H_0. (serve la distrib. multinomiale)
È calcolabile con metodi sofisticati
oppure si può fare una simulazione Monte Carlo.

Ad esempio, genero $N = 1000$ volte le 300 variabili $X_1, \dots, X_{300}$ con legge discreta $\varphi_0 = (4\%, 75\%, 7\%, 14\%)$
Per 1000 volte calcolo W $\rightarrow$ ottengo 1000 valori $W_1 \dots W_{1000}$ che hanno la distribuzione vera di W sotto H_0

$\alpha = 5\%$ $\quad q = \text{PERCENTILE}([W_1, W_2, \dots, W_{1000}]; 0{,}95)$

95% $\quad$ 5%

950 W_i $\leftarrow q \rightarrow$ 50 W_i

✱ NB (O_1, O_2, O_3, O_4) sotto H_0 ha legge multinomiale di parametri $n = 300$; $p_1 = 4\%$, $p_2 = 75\%$, $p_3 = 7\%$, $p_4 = 14\%$

★ Ad esempio si può usare questo test per verificare se la distanza in mesi dal compleanno al decesso ha legge uniforme. Si scopre che (per i vip) non è così.

Generalizzazione a TANTI o INFINITI valori

→ Si raggruppa X in pochi valori e si fa il test elementare

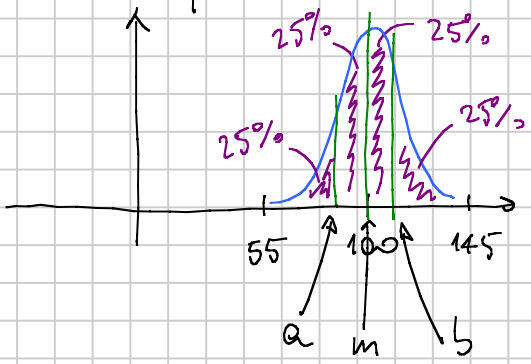
Ad esempio: vedi file Excel

ora 29

Generalizzazione a variabili CONTINUE

→ Si dividono i valori possibili in fasce (poche e possibilmente equiprobabili) e si fa il test elementare

ad esempio: $N(100, 15^2)$ ⟷ quoziente di intelligenza.



$a = \text{INV.NORM}(25\%; 100; 15)$

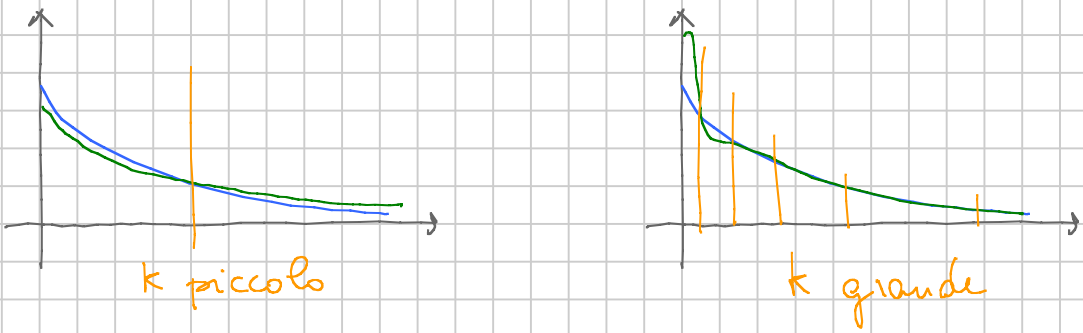
$b = \text{INV.NORM}(75\%; 100; 15)$

$m = 100 \quad (= \text{INV.NORM}(50\%; 100; 17))$

★ Se i bin sono più o meno equiprobabili l'approssimazione è migliore e il test è più potente

↳ Se X è continua, conoscendo F^{-1} è possibile (e raccomandato) ottenere bin esattamente equiprobabili.

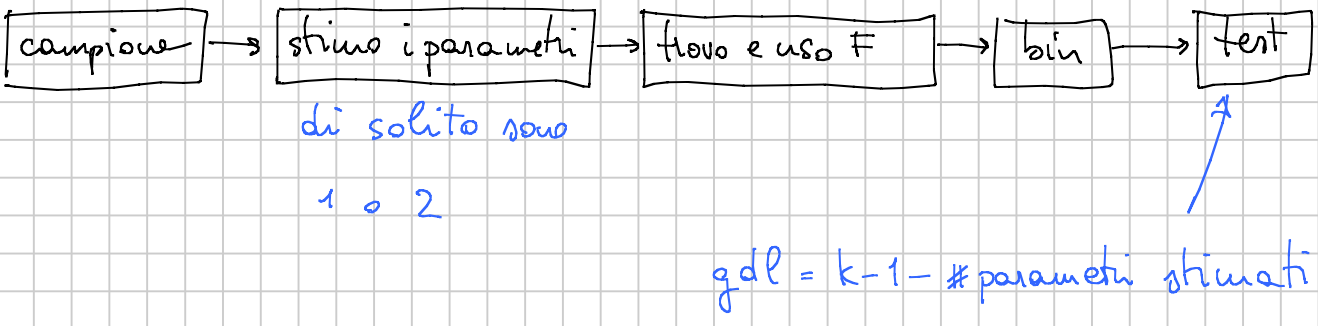
- Quanti bin bisogna fare:
  i. maggiore n, maggiore k
  ii. k mai troppo grande (gli attesi devono rispettare la "rule of thumb")
  iii. più k è piccolo più il test è potente (differenze anche piccole, ma globali)
  iv. più k è grande più il test è sensibile (differenze locali, non troppo piccole)



★ In mancanza di orientamenti precisi, consiglio k piccolo, dispari ($k = 5, 7, 9$)

- Generalizzazione a distribuzione specificata a meno di parametri

Q: Questo campione è normale?



★ Scegliere k poco più grande per non trovarsi con troppo pochi gdl

TABELLE DI CONTINGENZA

Esempio: parti cesarei / ospedale

	P.C.	P.N.
E.R.	22	74
H.	18	45
G.R.	33	71

Q1: La propensione al P.C. è la stessa in tutti gli ospedali?
Q2: La variabile "tipo di parto" è influenzata da quella "ospedale"?
Q3: La variabile "ospedale" è influenzata da quella "tipo di parto"?
Q4: Le variabili sono indipendenti?

→ C'è un solo test:

H_1: NO, SÌ, SÌ, NO
H_0: SÌ, NO, NO, SÌ

} risponde a tutte le Q (perché in realtà sono una sola)

- Come si esegue il test:

	P.C.	P.N.	
E.R.	22	74	96
H.	18	45	63
G.R.	33	71	104
	73	190	263

$27{,}8\% \cdot 96 = 26{,}6 = A_{ER,PC}$

atteso nell'ipotesi "di indipendenza"

$$A_{ij} = \frac{T_{i*} \cdot T_{*j}}{T_{**}}$$

dove $T_{i*} = \sum_j o_{ij}$ $\quad T_{*j} = \sum_i o_{ij}$ $\quad T_{**} = \sum_i \sum_j o_{ij}$

$27{,}8\% = \frac{73}{263}$ stima dell'incidenza dei cesarei

	P.C.	P.N.	
E.R.	22	74	96
H.	18	45	63
G.R.	33	71	104
	73	190	263

} osservati

atteso

	P.C.	P.N.	
E.R.	26,6	69,4	96
H.	17,5	45,5	63
G.R.	28,9	75,1	104
	73	190	263

$$W = \sum_{i,j} \frac{(O_{i,j} - A_{i,j})^2}{A_{i,j}} \overset{H_0}{\sim} \chi^2[(m-1)(n-1)]$$

$\sum_{i,j}$: somma su tutte le caselle

$(m-1)$: # di righe

$(n-1)$: # di colonne

MMSD (B)

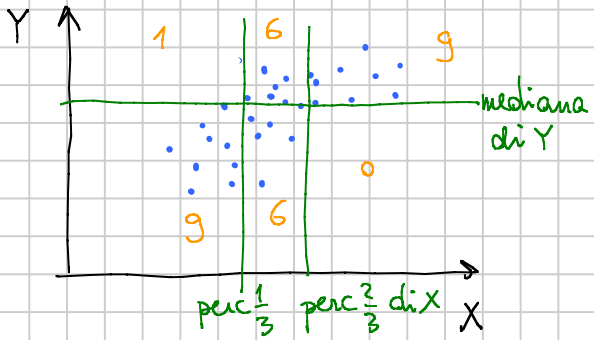
ora 30

Note Title 01/06/2011

Esame : mar 7 18 + ritardo round 0 (crocette)
mer 8 8.30 round 1 (Excel)
entro sab 4 8.30 inviare file desiderati al round 1
entro lun 6 li rimetto online corretti e ripuliti

TABELLE DI CONTINGENZA

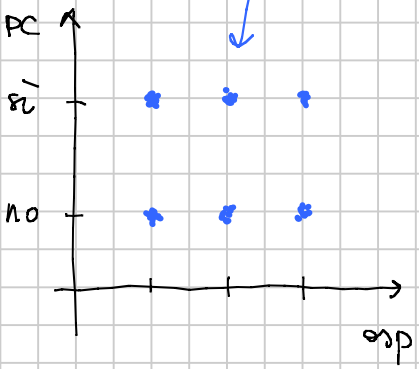
- 2 Variabili numeriche con tanti valori



Y \ X	1	6	9
9		6	0
1			

dirà H_1

var sono categoriche

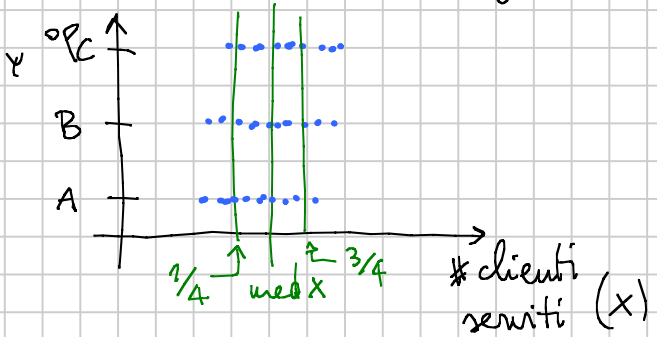


★ Il test mi dice se X e Y sono indipendenti (H_0) o correlate (H_1)

→ Test non parametrico analogo alla regr. semplice
non ha ipotesi (omosch, gauss, linearità, ...)
ma è meno potente e non fornisce un modello quantitativo

→ Conviene di solito conviene "affettare" ciascuna variabile in modo equiprobabile. Di solito 2×2, 2×3 o 3×3 sono numeri adeguati

- 1 variabile categorica e una numerica



→ Test non parametrico analogo alla ANOVA a 1 via

☆ Il test mi dice se la distrib di X è diversa per A, B, C (H_1) (non solo la media ma "tutta" la distribuzione) o no (H_0)

↑ dipende da quante fette faccio

= ... se X dipende da Y (H_1) o no (H_0)

N.B. Situazione interessante

X categorica : campione

φ_0 assegnata χ^2 sempl. 3gdl

	A	B	C	D	
campione	17	6	2	12	37
	50%	10%	10%	30%	10%
	18,5	3,7	3,7	11,1	attesi
campione 2	22	9	10	19	60

tab cont 4×2 → 3gdl

Test esatto di Fisher o test di Fisher-Irwin

Permette di ottenere un p dei dati esatto nel caso di tabelle di contingenza 2×2 (qualunque numerosità)

↳ Se le numerosità sono elevate non è pratico usarlo

7	2	9
4	12	16
11	14	25

4 (attesi)	5	9
7	9	16
11	14	25

legge V		9
		16
11	14	25

(25) palline: (9) rosse, 16 blu

Ne pesco (11) a caso: quante rosse escono?

Questa legge è detta ipergeometrica, una legge discreta che ha 3 parametri interi

$P(V=k) = \varphi_V(k) = \text{DISTRIB.IPERGEOM}(k; 11; 9; 25)$

★ Sotto H_0 (indipendenza delle 2 var) V ha legge ipergeom.
Se osservo valori di V troppo estremi ho il diritto di scegliere H_1

→ Si fa sempre con il p dei dati:

α^* : somma delle probabilità $\varphi_V(k)$ per quei k che sono almeno altrettanto estremi/rari del valore effettivamente osservato

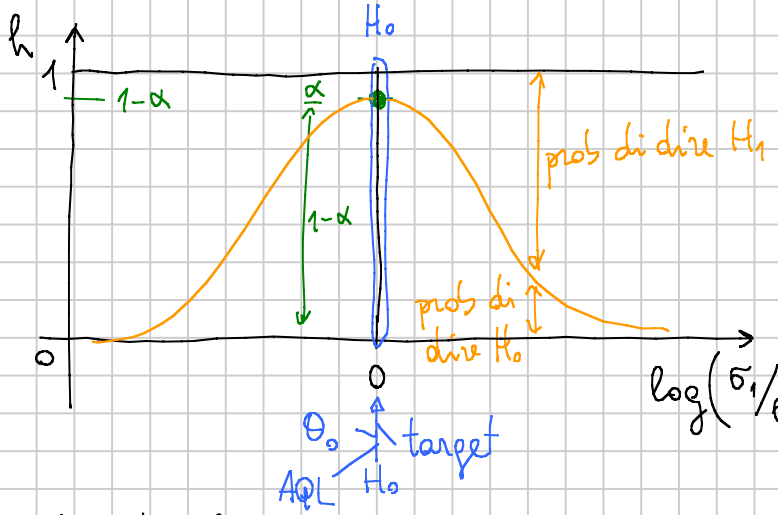
ora 31

TEST STATISTICI E CURVE OC

① stimatore
② statistica
③ p dei dati

} bilaterali e unilaterali

{ su parametri: $\mu, \sigma, p, \lambda, \mu_1-\mu_2$, $\beta, \alpha, \beta_i, \sigma, \sigma_1/\sigma_2$
(non su parametri: anova, χ^2)



$h = h(\text{parametro}) = P(\text{dire } H_0)$

$\log(\sigma_1/\sigma_2) = \theta$ incognito

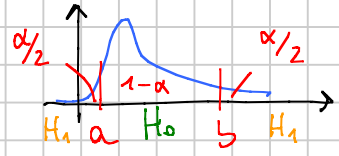
test bilaterale $H_0: \sigma_1 = \sigma_2$ $H_1: \sigma_1 \neq \sigma_2$ $\left(H_0: \frac{\sigma_1}{\sigma_2} = \eta\right)$

② $1-\alpha = P(\text{dire } H_0 \mid \text{vera } H_0) = \ldots \Rightarrow$ RA

$h(\theta) = P(\text{dire } H_0)$: β prob err II specie (se $\theta \notin H_0$)
$1-h(\theta) = P(\text{dire } H_1)$: potenza del test (se $\theta \notin H_0$)

★ Come si calcola h? Prima devo costruire il test ② o ①

② $\frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(n-1, n-1)$



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

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

$$h(\theta) = P(\text{dire } H_0 \,;\, \theta) = P\left(a \le \frac{S_1^2}{S_2^2} \le b \,;\, \theta\right) = \begin{cases} \rightsquigarrow 1-\alpha \text{ solo se } \theta = \theta_0 = 0 \\ \hookrightarrow \text{e in generale?} \end{cases}$$

$$\frac{S_1^2}{\sigma_1^2}(m-1) \sim \chi^2(m-1) \qquad \frac{S_2^2}{\sigma_2^2}(n-1) \sim \chi^2(n-1)$$

F di Fisher : rapporto tra due χ^2 indip, corrette coi loro gdl

$$\frac{\frac{S_1^2}{\sigma_1^2}(m-1)/(m-1)}{\frac{S_2^2}{\sigma_2^2}(n-1)/(n-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \boxed{\frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \sim F(m-1;n-1)}$$

funz ancillare di σ_1/σ_2

$$\boxed{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)}$$

$$h(\theta) = P\left(a \le \frac{S_1^2}{S_2^2} \le b \,;\, \theta\right) = P\left(a\frac{\sigma_2^2}{\sigma_1^2} \le \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \le b\frac{\sigma_2^2}{\sigma_1^2} \,;\, \theta\right)$$

$$\theta = \log\left(\sigma_1/\sigma_2\right)$$

$$h(\theta) = \left(a e^{-2\theta} \le F(m-1;n-1) \le b e^{-2\theta}\right) = \text{DISTRIB.F}\left(b e^{-2\theta}; m-1; n-1\right) - \text{DISTRIB.F}\left(a e^{-2\theta}; m-1; n-1\right)$$

- Caso unilaterale

campione di n oggetti, X il numero di difettosi nel campione
p la frequenza di difettosi nella popolazione

N.B. di solito nel controllo Q popolazione = fornitura

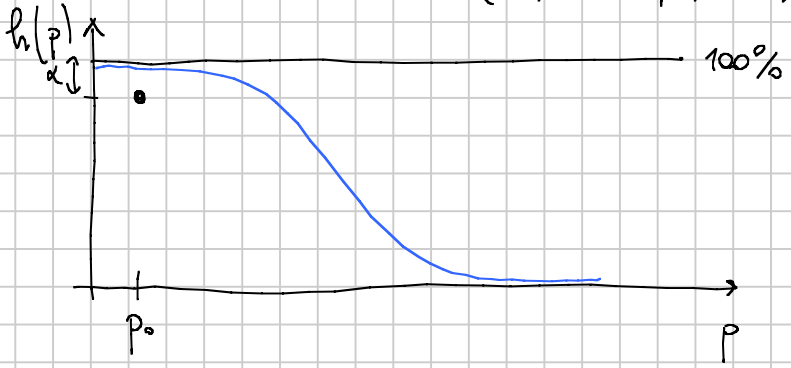
$$p = \frac{\#\text{difettosi fornitura}}{\#\text{fornitura}}$$

$\boxed{X \sim \text{bin}(n,p)}$ test : $\boxed{\begin{array}{ll} X \le a & \text{accetto fornitura } (H_0) \\ X > a & \text{rifiuto fornitura } (H_1) \end{array}}$

$\hookrightarrow RA_X : \{0,1,\dots,a\}$ $\quad H_0 : p \le p_0$ $\quad H_1 : p > p_0$

$$h(p) = P(\text{dire } H_0 ; p) = P(X \in RA_X ; p) = P(X \le a ; p)$$

$$= \text{DISTRIB.BINOM}(a ; n ; p ; 1)$$



Cosa p_0 ?

difettosi nelle forniture

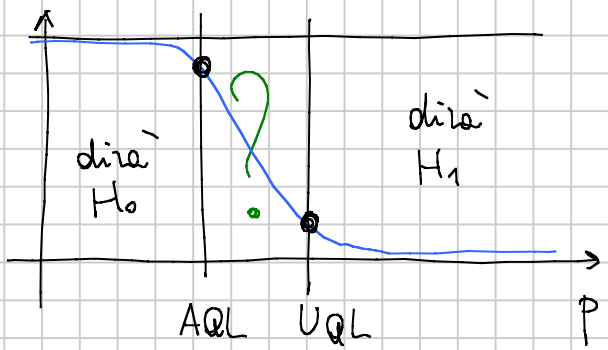
Voglio che se $p \le p_0$ allora accetterò la fornitura con prob elevata $\ge 1-\alpha$

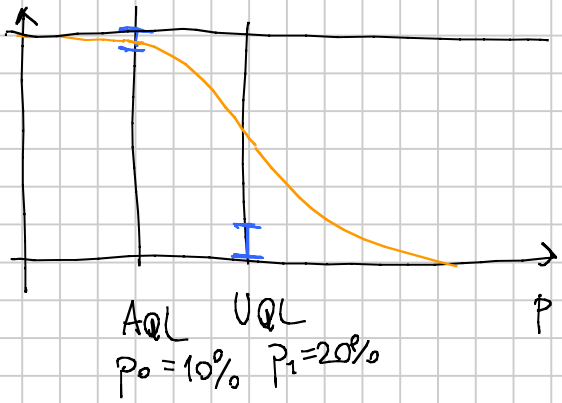
p_0 : AQL acceptable quality level

11% ≤ 5%

p_1 : UQL unacceptable quality level (RQL)

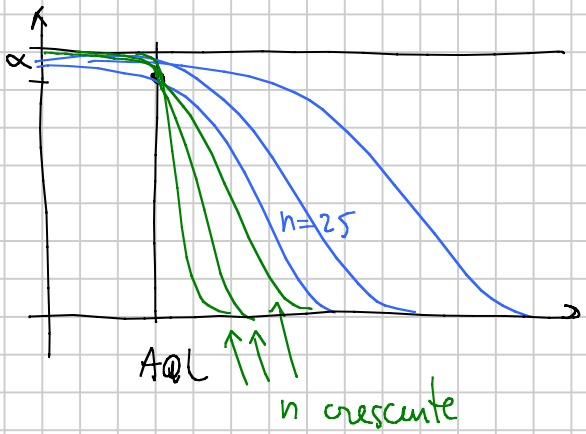
30% ≤ 20%

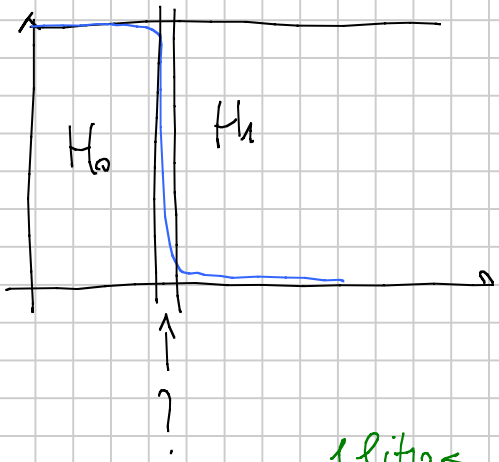




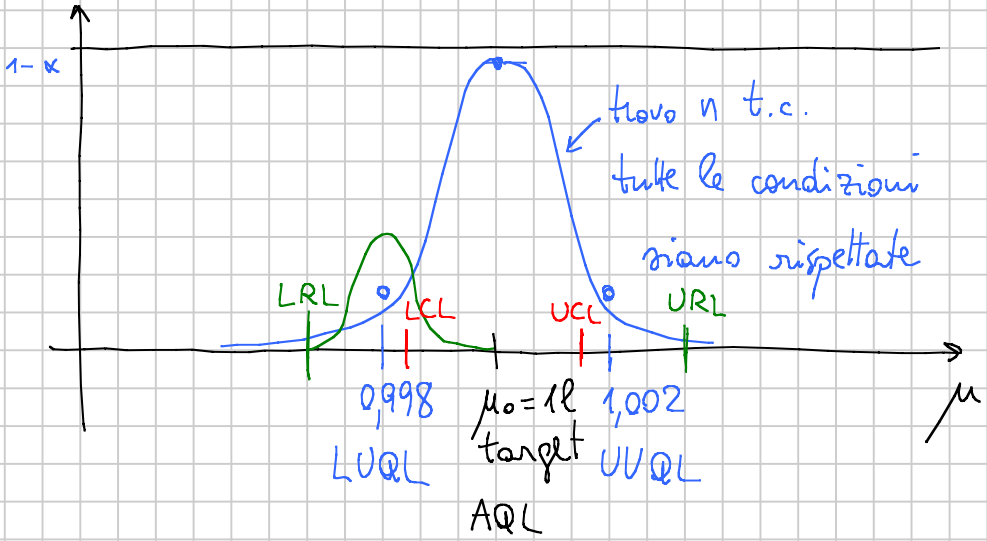
UQL = 20% $\quad \beta = 10\%$

AQL = 10% $\quad \alpha = 5\%$





- UQL bilaterale



1 litro ← c'è

$X \sim \mathcal{N}(\mu, \sigma^2)$

riempimento cartoni di latte

σ nota

$H_0: \mu = \mu_0 \qquad H_1: \mu \neq \mu_0 \qquad X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} \mathcal{N}(0,1) \qquad \alpha = 0{,}0027$

② $|Z| > q$ dico H_1

$|Z| \leq q$ dico H_0

$q = \text{INV.NORM.ST}\left(1 - \frac{\alpha}{2}\right) = 3$

① $\bar{X} \in \mu_0 \pm q\frac{\sigma}{\sqrt{n}}$ LCL , UCL

UQL determinati in funzione di quali livelli di riempimento sui singoli cartoni ritengo inaccettabili LRL , URL

N.B. Teorema di Cochran ora 6 anno scorso

FINE