METODI E MODELLI A SUPPORTO DELLE DECISIONI (B)

Note Title 26/09/2011

ora 1

Programma di massima e bibliografia

- Ross: P.e S.p!.e S. Apogeo
  (1,2) 3 – 8 già fatti
  ↳ ripasso: esercizi cap 8 e 7 con Excel
  9 regressione + appunti + lab
  ↳ DoE (Design of Experiment)
  - Sleeper: Design for Six Sigma Statistics
    59 tools for ...
    capitolo 10 sul DoE
  10 analisi della varianza
  11 test del chi-quadro (nelle diverse varianti)

★ Conoscere meglio i test statistici
  ↳ Curva OC (AQL, UQL)
  ↳ potenza / err. di II specie
  ↳ Test controllo qualità (p di Bernoulliana) esatto, non in

★ Simulazioni Montecarlo (semplici, con Excel)

- Middleton: Analisi statistiche con Excel Apogeo
  ↳ riferimento per Excel

(- Per approfondimenti:
  i. www.statsoft.com "statistics textbooks"
  ii. dispense del prof. Soliani 3000 pag.
  iii. per il DoE, libri di)

A cosa serve:
→ Qualità → Six Sigma
→ Continuous improvement → Analisi dei dati

green belt
black belt
master black belt (matematico)

★ Cercate di capire tutto e di approfondire dove serve

- Lezioni
  - 2 × 10 = 20 ore di laboratorio
  - 32 ore di lezione
  - Tablet → pdf → per gli appunti stampare 2010
    - avi
    - xlsx dei laboratori

  } anche per gli anni scorsi soprattutto i laboratori

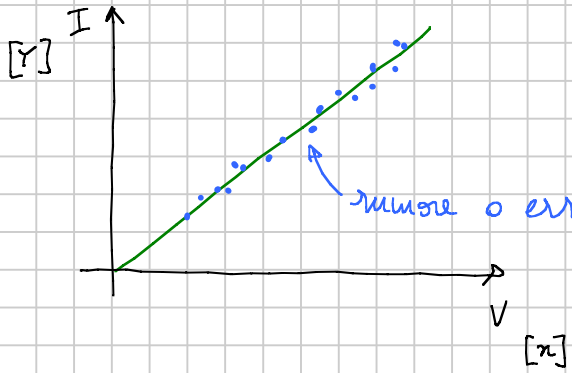
  ★ tutto su lee.unipr.it
    - compiti vecchi
    - forum

- Esame: (6 appelli all'anno)
  "round 0" test a crocette vale 2 sessioni (2/3 anno)
  ↳ c'è il programma su lee. (domande elementari su tutto)
  "round 1" compito su Excel $5\frac{1}{2}$ ore si può usare i libro
  + Nicolodi
  + Orale tutto assieme

- Ricevimento:
  mar: 9.00 – 10.30 (mio studio)

REGRESSIONE

REGRESSIONE LINEARE SEMPLICE



$I = \frac{1}{R} V$ relazione lineare deterministica

$I = \frac{1}{R} V + e$ (casuale)

e: errore casuale additivo

var di ingresso — pensata deterministica

var di risposta — casuale

★ idea: ricavare f dai dati ottenuti da alcuni esperimenti
↳ si ricavano anche informazioni su e

• Modello considerato: f lineare, $e \sim \mathcal{N}(0, \sigma^2)$

σ^2: incognita, ma costante

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

β_0, β_1: incognite

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

• Dati: si fanno n esperimenti e si misurano x_i e Y_i

$i = 1, 2, \dots, n$

(a volte $x_i = x(i)$ $\quad Y_i = Y(i)$)

x	Y
x_1	Y_1
x_2	Y_2
...	...
x_n	Y_n

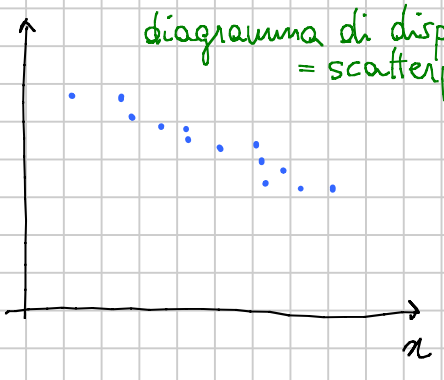
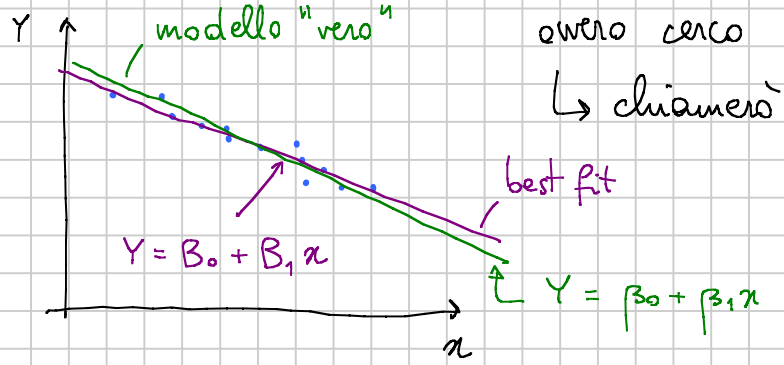


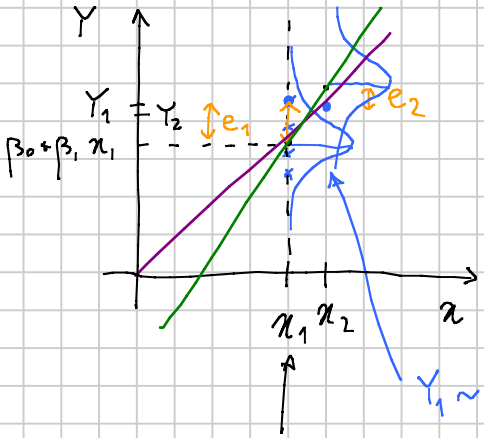
diagramma di dispersione = scatterplot

- Si cerca la funzione che approssima meglio i dati (= best fit)



ovvero cerco di stimare β_0 e β_1 (incognite)
↳ chiamerò B_0 e B_1 gli stimatori

- Vediamo in ordine causa-effetto



$Y = \beta_0 + \beta_1 x + e$ ← $e \sim \mathcal{N}(0, \sigma^2)$ modello

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

$Y = B_0 + B_1 x$ regressione

$Y_1 \sim \mathcal{N}(\beta_0 + \beta_1 x ; \sigma^2)$

fisso x_1 e misuro $Y_1 = \beta_0 + \beta_1 x_1 + e_1$

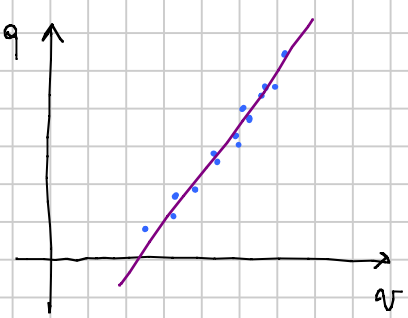
x_1	e_1	Y_1
x_2	e_2	Y_2
.....		
x_n	e_n	Y_n

x_i determ.
$e_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.
$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i ; \sigma^2)$ indipendenti tra loro
media e hanno medie diverse

Ripassare legge normale

- Qualche esempio

1) Ripartitore polveri v velocità cocleа
q quantità polvere

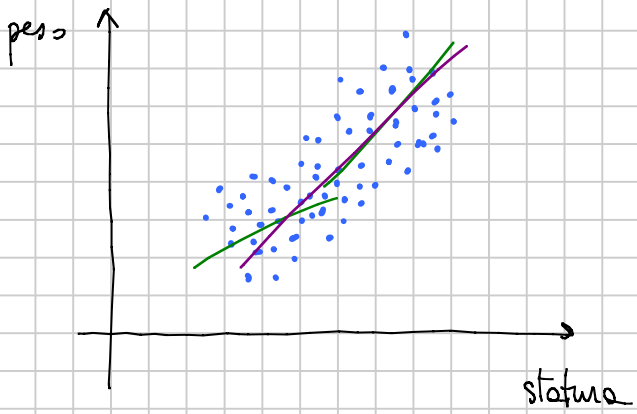


2) Degradazione principio attivo

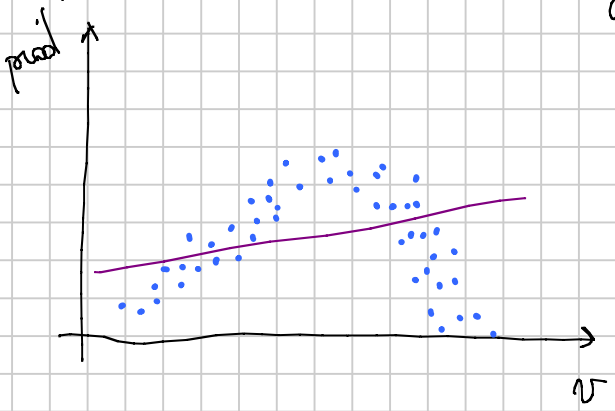
a concentrazione attivo

t concentrazione tossico

3) Legame peso - statura



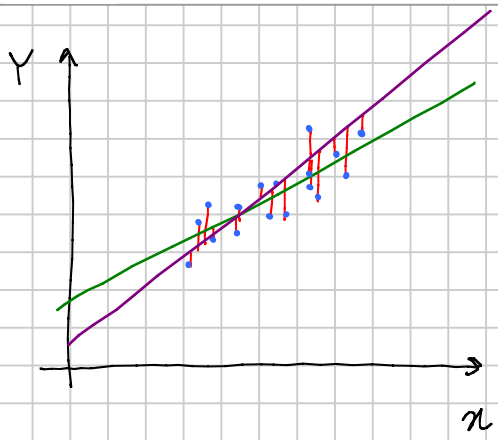
4) Produzione linea di montaggio



MMSD (B)

ord 3

Note Title 27/09/2011



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

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

$Y = B_0 + B_1 x$ retta "best fit"

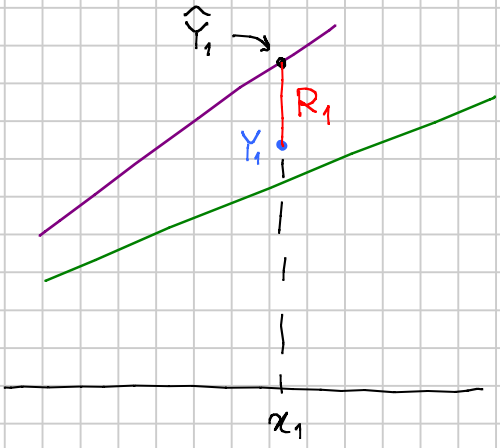
$S_e \approx \sigma$ $B_0 \approx \beta_0$ $B_1 \approx \beta_1$

stimatori

- Previsti e residui

→ $\hat{Y}_i$ il previsto medio per valore di ingresso x_i

$$\hat{Y}_i := B_0 + B_1 x_i$$



→ R_i il residuo corrispondente

$$R_i := Y_i - \hat{Y}_i = Y_i - B_0 - B_1 x_i$$

★ I residui sono un po' positivi e un po' negativi

- Come si trova la retta di regressione

È quella alla quale corrispondono i residui più piccoli

$$SS_R := SS := \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i)^2$$

somma dei quadrati dei residui

$SS = SS(B_0, B_1)$ dipende dalla retta

→ cerco B_0 e B_1 che minimizzano SS

→ siccome la struttura di SS è semplice (rispetto a B_0, B_1) basta annullare le derivate parziali

$$\frac{\partial SS}{\partial B_0} = \sum_{i=1}^{n} \frac{\partial}{\partial B_0}\left(Y_i - B_0 - B_1 x_i\right)^2 = \sum_{i=1}^{n} 2\left(Y_i - B_0 - B_1 x_i\right)(-1)$$

$$\frac{\partial SS}{\partial B_1} = \sum_{i=1}^{n} \frac{\partial}{\partial B_1}\left(Y_i - B_0 - B_1 x_i\right)^2 = \sum_{i=1}^{n} 2\left(Y_i - B_0 - B_1 x_i\right)(-x_i)$$

le pongo $=0$

$$\begin{cases} \sum_{i=1}^{n} \left(Y_i - B_0 - B_1 x_i\right) = 0 \\ \sum_{i=1}^{n} \left(Y_i - B_0 - B_1 x_i\right) x_i = 0 \end{cases}$$

$\sum_{i=1}^{n} R_i = 0$ ovvero $\bar{R} = 0$

equazioni normali

$$B_1 = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} = \frac{\text{cov. camp}\,(x_i ; Y_i)}{\text{var. camp}\,(x_i)}$$

HW

$$B_0 = \frac{\overline{x^2}\,\bar{Y} - \bar{x}\,\overline{xY}}{\overline{x^2} - \bar{x}^2} = \bar{Y} - \bar{x} B_1$$

- Come si stima σ^2 ?

$Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i ; \sigma^2\right)$ $\quad S_Y^2$ non va bene

hanno media diversa

$e_i \sim \mathcal{N}(0, \sigma^2)$ non li conosco!

$R_i \sim ?$ posso tentare

$$S_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(R_i - \bar{R}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} R_i^2 = \frac{SS}{n-1}$$

$$S_R^2 = \frac{SS}{n-1}$$

è sempre zero!!

I eq normale

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - B_0 - B_1 x_i\right) = 0$$

lo stimatore "giusto" si chiama errore standard

$$S_e^2 := \frac{SS}{n-2}$$

- Caratteristiche degli stimatori

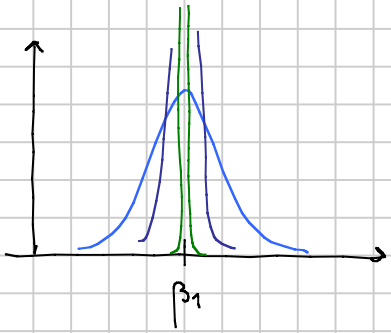
$$E(B_1) = \beta_1 \qquad E(B_2) = \beta_2 \qquad E(S_e^2) = \sigma^2$$

$$E(S_e) < \sigma$$

R: stimatore
" corretto
" consistente

$$B_1^{(n)} \xrightarrow[n \to \infty]{} \beta_1$$

$$Var(B_1^{(n)}) \xrightarrow[n \to \infty]{} 0$$



- Distribuzione degli stimatori

$$B_1 = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} = \frac{1}{D}\left[\frac{1}{n}\sum_i x_i Y_i - \bar{x}\frac{1}{n}\sum_i Y_i\right] \sim \mathcal{N}(?\,;?)$$

$Y_i = \beta_0 + \beta_1 x_i + e_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i\,;\,\sigma^2)$ indipendenti

R: trasf lineari di $\mathcal{N}$
somma di $\mathcal{N}$

$$E(B_1) = \frac{1}{D}\left[\frac{1}{n}\sum_i x_i E(Y_i) - \bar{x}\frac{1}{n}\sum_i E(Y_i)\right]$$

$$= \frac{1}{D}\left[\frac{1}{n}\sum_i x_i(\beta_0 + \beta_1 x_i) - \bar{x}\frac{1}{n}\sum_i(\beta_0 + \beta_1 x_i)\right]$$

$$= \frac{1}{Dn}\left[\beta_0 \sum x_i + \beta_1 \sum x_i^2 - \beta_0 \bar{x} n + \beta_1 \bar{x} \sum x_i\right]$$

$$= \frac{\beta_1}{Dn}\left[\sum x_i^2 - \bar{x}\sum x_i\right] = \frac{\beta_1}{D}\left[\overline{x^2} - \bar{x}^2\right] = \beta_1$$

R: proprietà della media "E"

$$Var(B_1) = Var\left\{\frac{1}{D}\left[\frac{1}{n}\sum_i x_i Y_i - \bar{x}\frac{1}{n}\sum_i Y_i\right]\right\}$$

passaggio importante: raccogliere più possibile le Y_i

R: proprietà della varianza "Var"

$$\mathrm{Var}(a+bX) = b^2\,\mathrm{Var}(X)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \qquad \text{se } X \text{ e } Y \text{ sono indipendenti}$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X;Y) \qquad \text{in generale}$$

$$\mathrm{Var}(X) = \mathrm{Cov}(X;X)$$

$$\mathrm{Cov}(X;Y) = \mathrm{Cov}(Y;X)$$

$$\mathrm{Cov}(aX+bY;Z) = a\,\mathrm{Cov}(X;Z) + b\,\mathrm{Cov}(Y;Z) \qquad \text{bilinearità}$$

$$\mathrm{Var}(aX+bY) = \mathrm{Cov}(aX+bY;aX+bY)$$

$$= a\,\mathrm{Cov}(X;aX+bY) + b\,\mathrm{Cov}(Y;aX+bY)$$

$$= a^2\,\mathrm{Cov}(X;X) + ab\,\mathrm{Cov}(X;Y) + ba\,\mathrm{Cov}(Y;X) + b^2\,\mathrm{Cov}(Y;Y)$$

$$= a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X;Y) + b^2\,\mathrm{Var}(Y)$$

$$\mathrm{Var}(B_1) = \mathrm{Var}\left\{\frac{1}{D}\left[\frac{1}{n}\sum_i x_i Y_i - \bar{x}\,\frac{1}{n}\sum_i Y_i\right]\right\}$$

$$= \frac{1}{D^2 n^2}\,\mathrm{Var}\left\{x_1 Y_1 + x_2 Y_2 + \dots + x_n Y_n - \bar{x}\left(Y_1 + Y_2 + \dots + Y_n\right)\right\}$$

$$= \frac{1}{D^2 n^2}\,\mathrm{Var}\left\{Y_1(x_1-\bar{x}) + Y_2(x_2-\bar{x}) + \dots + Y_n(x_n-\bar{x})\right\} \qquad \text{indip}$$

$$= \frac{1}{D^2 n^2}\left\{\mathrm{Var}\left(Y_1(x_1-\bar{x})\right) + \mathrm{Var}(\dots) + \dots + \mathrm{Var}\left(Y_n(x_n-\bar{x})\right)\right\}$$

$$= \frac{1}{D^2 n^2}\left\{(x_1-\bar{x})^2\sigma^2 + (x_2-\bar{x})^2\sigma^2 + \dots\right\}$$

$$= \frac{\sigma^2}{D^2 n^2}\left\{x_1^2 - 2\bar{x}x_1 + \bar{x}^2 + x_2^2 - 2\bar{x}x_2 + \bar{x}^2 + \dots + x_n^2 - 2\bar{x}x_n + \bar{x}^2\right\}$$

$$= \frac{\sigma^2}{D^2 n^2}\left\{\underbrace{\sum_i x_i^2}_{n\overline{x^2}} \overbrace{- 2\bar{x}\underbrace{\sum_i x_i}_{n\bar{x}}}^{-2n\bar{x}^2} + n\bar{x}^2\right\} = \frac{\sigma^2}{D^2 n^2}\left\{n\overline{x^2} - n\bar{x}^2\right\}$$

$$\mathrm{Var}(B_1) = \frac{\sigma^2}{Dn} = \frac{\sigma^2}{n\left(\overline{x^2} - \bar{x}^2\right)} =: \sigma^2 k_1$$

$$k_1 = \frac{1}{n\left(\overline{x^2} - \bar{x}^2\right)} \xrightarrow[n\to\infty]{} 0$$

HW: controllare il conto 2010-11

$$B_1 \sim \mathcal{N}\left(\beta_1 ; \sigma^2 k_1\right)$$

HW: Rifare per B_0

$$B_0 = \frac{\overline{x^2}\,\overline{Y} - \bar{x}\,\overline{xY}}{\overline{x^2} - \bar{x}^2} = \frac{1}{D}\left[\overline{x^2} \cdot \frac{1}{n}\sum_i Y_i - \bar{x} \cdot \frac{1}{n}\sum_i x_i Y_i\right]$$

$\rightarrow E(B_0) = \ldots = \beta_0$

$\rightarrow Var(B_0) = \ldots = \dfrac{\sigma^2 \overline{x^2}}{n D} =: \sigma^2 k_0 \qquad k_0 = \dfrac{\overline{x^2}}{n(\overline{x^2} - \bar{x})}$

$\rightarrow B_0 \sim \mathcal{N}$

$$B_0 \sim \mathcal{N}\left(\beta_0 ; \sigma^2 k_0\right)$$

★ B_0 e B_1 non sono indipendenti

HW: $Cov(B_0 ; B_1) = -\sigma^2 \dfrac{\bar{x}}{n(\overline{x^2} - \bar{x}^2)}$

[lo faccio martedì prossimo]

HW: $R_i \sim ?$

$Cov(R_i ; R_j) = ?$

MMSD

ora 5

Note Title 04/10/2011

REGR. LIN. SEMPLICE

$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2) \quad \text{iid}$$

$\beta_0 \to B_0$ $\quad \beta_1 \to B_1$ $\quad \sigma^2 \to S_e$

Distribuzione delle statistiche

$$B_1 = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2} = \frac{1}{nD}\left(\sum_j x_j Y_j - \sum_i x_i \cdot \frac{1}{n}\sum_j Y_j\right) \qquad (\bar{x})$$

$$= \frac{1}{nD}\sum_{j=1}^{n}\left\{x_j Y_j - Y_j \bar{x}\right\} = \frac{1}{nD}\sum_j \left(x_j - \bar{x}\right) Y_j = \sum_{j=1}^{n} c_j Y_j$$

dove $\quad c_j := \dfrac{x_j - \bar{x}}{n\left(\overline{x^2} - \bar{x}^2\right)}$

$$B_0 = \bar{Y} - \bar{x} B_1 = \frac{1}{n}\sum_j Y_j - \sum_j \bar{x} c_j Y_j = \sum_{j=1}^{n}\left\{\frac{1}{n} - \frac{x_j - \bar{x}}{n\left(\overline{x^2} - \bar{x}^2\right)}\bar{x}\right\} Y_j = \sum_{j=1}^{n} d_j Y_j$$

$$\frac{\overline{x^2} - \bar{x}^2 - \bar{x}x_j + \bar{x}^2}{n\left(\overline{x^2} - \bar{x}^2\right)}$$

dove $\quad d_j := \dfrac{\overline{x^2} - \bar{x}x_j}{n\left(\overline{x^2} - \bar{x}^2\right)}$

ho scritto B_1 e B_0 come combinazione lineare delle Y_j

- HW ora 4 : $\mathrm{Cov}\left(B_0, B_1\right) = \mathrm{Cov}\left(\sum_j d_j Y_j ; \sum_i c_i Y_i\right)$

Bilinearità della covarianza

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i ; \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}\left(X_i, Y_j\right)$$

ogni volta che $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ sono v.a. e $a_1, \ldots, a_m, b_1, \ldots, b_n$ sono reali qualsiasi

Forma matriciale

Sia Σ la matrice di covarianza dei vettori aleatori

$\vec{X} = (X_1, X_2, \dots, X_m)$ $\quad \vec{Y} = (Y_1, Y_2, \dots, Y_n)$

$\mathrm{Cov}(\vec{X}; \vec{Y}) := \Sigma \in M_{m,n}$ $\quad \Sigma_{ij} := \mathrm{Cov}(X_i; Y_j)$

Siano $\vec{a} = (a_1, \dots, a_m) \in \mathbb{R}^m$ $\quad \vec{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$

Allora: $\boxed{\mathrm{Cov}(\vec{a}\cdot\vec{X}; \vec{b}\cdot\vec{Y}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i \Sigma_{ij} b_j = \vec{a}^T \Sigma \vec{b}}$

Siano $M \in M_{c,m}$ $\quad N \in M_{d,n}$

Allora $\mathrm{Cov}(M\vec{X}; N\vec{Y}) \in M_{c,d}$ matr di cov di $M\vec{X}$ e $N\vec{Y}$

$\boxed{\mathrm{Cov}(M\vec{X}; N\vec{Y}) = M \Sigma N^T}$

$$\mathrm{Cov}(B_0; B_1) = \mathrm{Cov}\left(\sum_j d_j Y_j; \sum_i c_i Y_i\right) = \sum_{j=1}^{n}\sum_{i=1}^{n} c_i d_j \mathrm{Cov}(Y_j; Y_i)$$

↑ bilinearità

$Y_1, Y_2, \dots, Y_n$ indipendenti $\quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i; \sigma^2)$

(la matr. di cov. tra $\vec{Y}$ e $\vec{Y}$ è $\sigma^2 I$)

$$\sum_{j=1}^{n} c_j d_j \mathrm{Cov}(Y_j; Y_j) = \sum_{j=1}^{n} c_j d_j \sigma^2 = \sigma^2 \sum_{j=1}^{n} \frac{x_j - \bar{x}}{n(\overline{x^2} - \bar{x}^2)} \cdot \frac{\overline{x^2} - \bar{x} x_j}{n(\overline{x^2} - \bar{x}^2)}$$

$$= \sigma^2 \sum_j \frac{-\bar{x} x_j^2 + \bar{x}^2 x_j + \overline{x^2} x_j - \bar{x}\,\overline{x^2}}{n^2(\overline{x^2} - \bar{x}^2)^2} = \frac{\sigma^2}{D^1}\left\{-\bar{x}\underbrace{\sum x_j^2}_{n\overline{x^2}} + (\bar{x}^2 + \overline{x^2})\underbrace{\sum_j x_j}_{n\bar{x}} - n\bar{x}\,\overline{x^2}\right\}$$

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

↑ di solito negativa

- $R_i = Y_i - B_0 - B_1 x_i = Y_i - \sum_k d_k Y_k - x_i \sum_k c_k Y_k$

$$= Y_i\{1 - d_i - c_i x_i\} - \sum_{k \neq i}(d_k + c_k x_i) Y_k = \sum_{k=1}^{n} \rho_k Y_k$$

....

(vengono conti bruttini)

$B_0 \sim \mathcal{N}(\beta_0; \sigma^2 k_0)$ $\quad B_1 \sim \mathcal{N}(\beta_1; \sigma^2 k_1)$ non sono indipendenti

$\mathrm{Cov}(B_0, B_1) = \sigma^2 k_{01}$ $\qquad k_{01} := -\dfrac{\overline{x}}{n(\overline{x^2} - \overline{x}^2)}$

$S_e^2 := \dfrac{SS}{n-2}$ $\qquad SS := \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i)^2$

$\dfrac{S_e^2}{\sigma^2}(n-2) \sim \chi^2(n-2)$ per ora fidiamoci

N.B. Varianza campionaria $S^2 := \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ $\qquad \dfrac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$

Stimatore pooled $S_p^2 := \dfrac{m-1}{m+n-2} S_X^2 + \dfrac{n-1}{m+n-2} S_Y^2$ $\qquad \dfrac{S_p^2}{\sigma^2}(m+n-2) \sim \chi^2(m+n-2)$

Discendono tutte e tre dal thm di COCHRAN

★ Il thm di Cochran dice anche che

S^2 e $\overline{X}$ sono indipendenti

S_p^2 e $(\overline{X}, \overline{Y})$ sono indipendenti

S_e^2 e (B_0, B_1) sono indipendenti

N.B. Ciò è cruciale per la inferenza

Proprietà legge gamma

$\Gamma(\alpha, \lambda)$ ha due parametri

oppure $\Gamma(\alpha, \beta)$ dove $\beta = \frac{1}{\lambda}$

$\Gamma(2, 7)$ cos'è?

→ la gamma è una "somma" di exp iid

$\Gamma(1, \beta) \sim \exp\left(\frac{1}{\beta}\right)$ esponenziale di media β

$\Gamma(n, \beta)$ è la somma di n espon. di media β indip.

occhio che $\Gamma(\alpha, \beta)$ è definita anche per $\alpha > 0$ reale

→ $X \sim \Gamma(\alpha, \beta)$ $\quad Y \sim \Gamma(\eta, \beta)$ indipendenti

$X + Y \sim \Gamma(\alpha + \eta, \beta)$ (riproducibilità)

→ $X \sim \Gamma(\alpha, \beta)$ allora $cX \sim \Gamma(\alpha, c\beta)$

(omotetie)

$\rightarrow \chi^2(k) \sim \Gamma\left(\frac{k}{2}; 2\right)$

$\chi^2(2) \sim \Gamma(1; 2) \sim \text{expo}\left(\frac{1}{2}\right)$ (media 2)

$\rightarrow X \sim \Gamma(\alpha, \beta)$

$E(X) = \alpha\beta \quad Var(X) = \alpha\beta^2$

$$\frac{S_e^2}{\sigma^2}(n-2) \sim \chi^2(n-2) \sim \Gamma\left(\frac{n-2}{2}; 2\right)$$

allora $S_e^2 = \underbrace{\frac{\sigma^2}{n-2}}_{c} \underbrace{\frac{S_e^2}{\sigma^2}(n-2)}_{X \sim \Gamma\left(\frac{n-2}{2}; 2\right)} \sim \Gamma\left(\frac{n-2}{2}; \frac{2}{n-2}\sigma^2\right)$

$$S_e^2 \sim \Gamma\left(\frac{n-2}{2}; \frac{2}{n-2}\sigma^2\right)$$

$E(S_e^2) = \sigma^2$ corretto

$Var(S_e^2) = \frac{2\sigma^4}{n-2}$ consistente

HW: trovare le leggi di S^2, S_p^2

INFERENZA SUI PARAMETRI DI REGRESSIONE

i. Intervalli di confidenza

ii. Test statistici (anche β, potenza, curva OC, AQL, UQL)

iii. Intervalli di predizione

→ Si parte sempre dalla funzione ancillare del parametro incognito

$$\frac{B_0 - \beta_0}{\sigma\sqrt{k_0}} \sim N(0,1)$$

$$\frac{B_1 - \beta_1}{\sigma\sqrt{k_1}} \sim N(0,1)$$

Una funzione ancillare deve dipendere da al più un parametro incognito: quello su cui si fa l'inferenza

Ripasso funzioni ancillari

Dati normali, inferenza su μ

→ σ^2 nota: $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

→ σ^2 incognita: $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

Dati normali, inferenza su σ^2

→ μ nota: $\frac{S_*^2}{\sigma^2} n \sim \chi^2(n)$

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

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

questa è ok

$$\frac{B_0 - \beta_0}{S_e\sqrt{k_0}} \sim t(n-2)$$

$$\frac{B_1 - \beta_1}{S_e\sqrt{k_1}} \sim t(n-2)$$

queste sono ok

Dati bernoulliani, inferenza su p

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

$X \sim bin(n,p)$ (esatta: serve Excel)

Due campioni $\mathcal{N}$ omoschedastici $(\mu_x - \mu_y)$

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$$

Dati expo (λ) inferenza su λ

$$2n\lambda\bar{X} \sim \chi^2(2n)$$

ora 7

Cos'è la t di Student?

sia $Z \sim \mathcal{N}(0,1)$, sia $W \sim \chi^2(k)$ indipend.

allora $\frac{Z}{\sqrt{W/k}} \sim t(k)$

$$\frac{B_0 - \beta_0}{\sigma\sqrt{k_0}} \sim \mathcal{N}(0,1) \quad (= Z)$$

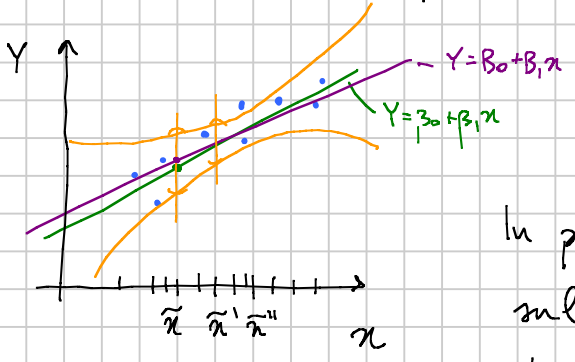
$$\frac{B_0 - \beta_0}{S_e\sqrt{k_0}} = Z\frac{\sigma}{S_e} = \frac{Z}{S_e/\sigma} = \frac{Z}{\sqrt{\frac{S_e^2}{\sigma^2}(n-2)\Big/(n-2)}} \sim t(n-2)$$

($\frac{S_e^2}{\sigma^2}(n-2) = W$, $(n-2) = k$)

↑ perché B_0 e S_e^2 sono indip.

quindi anche Z e W

- "Funzione ancillare" per la retta verde



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

In pratica ci riconduciamo a fare l'inferenza sul "parametro" $\beta_0 + \beta_1\tilde{x}$

dove $\tilde{x}$ lo pensiamo fissato e assegnato

e poi lo facciamo variare

Ad es: interv di conf: $\beta_0 + \beta_1\tilde{x} \in B_0 + B_1\tilde{x} \pm$ raggio

1) Identificare il parametro : $\beta_0 + \beta_1 \tilde{x}$

2) Trovare uno stimatore : $B_0 + B_1 \tilde{x}$

3) Trovare la distribuzione dello stimatore

i. $B_0 + B_1 \tilde{x} \sim \mathcal{N}(?\,;\,?)$

ii. $E(B_0 + B_1 \tilde{x}) = E(B_0) + \tilde{x}\,E(B_1) = \beta_0 + \beta_1 \tilde{x}$ corretto

iii. $\mathrm{Var}(B_0 + B_1 \tilde{x}) = \mathrm{Var}(B_0) + 2\tilde{x}\,\mathrm{Cov}(B_0; B_1) + \tilde{x}^2 \mathrm{Var}(B_1)$

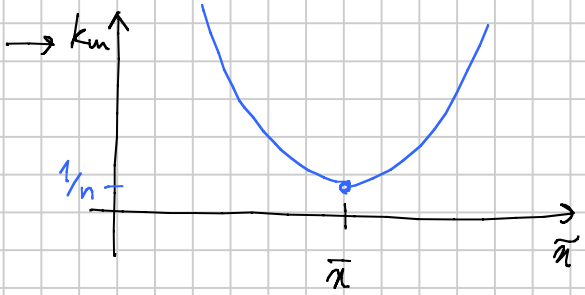
$= \sigma^2 k_0 + 2\tilde{x}\sigma^2 k_{01} + \tilde{x}^2 \sigma^2 k_1 =: \sigma^2 k_m(\tilde{x})$

($a=1$, $X=B_0$; $b=\tilde{x}$, $Y=B_1$)

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

★ $\boxed{B_0 + B_1 \tilde{x} \sim \mathcal{N}(\beta_0 + \beta_1 \tilde{x}\,;\, \sigma^2 k_m(\tilde{x}))}$

→ consistente ($k_m \downarrow 0$ con n)



è sempre $\geq \frac{1}{n}$ e minima per $\tilde{x} = \bar{x}$

4) funzione ancillare

$$\frac{B_0 + B_1 \tilde{x} - (\beta_0 + \beta_1 \tilde{x})}{\sigma \sqrt{k_m(\tilde{x})}} \sim \mathcal{N}(0,1)$$

$$\boxed{\frac{B_0 + B_1 \tilde{x} - (\beta_0 + \beta_1 \tilde{x})}{S_e \sqrt{k_m(\tilde{x})}} \sim t(n-2)}$$

↓ grazie all'indipendenza tra S_e e (B_0, B_1)

INTERVALLI DI CONFIDENZA

- Intervallo di confidenza per la risposta media

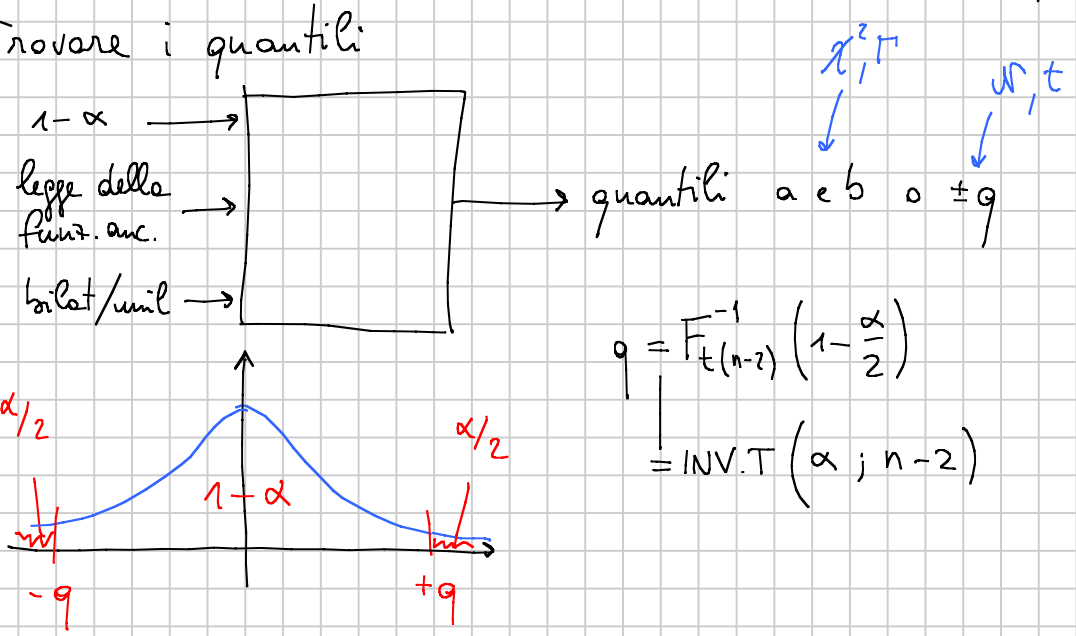
$Y \sim \mathcal{N}(\beta_0 + \beta_1 x\,;\, \sigma^2)$

= intervallo di confidenza per la retta verde

(Lo faccio bilaterale)

1) Fissare il livello di confidenza : $1-\alpha$ $(= 95\%)$

2) Trovare i quantili



3) Scrivere quello che so e ricavare con passaggi algebrici l'intervallo

$$1-\alpha = P\left(-q \le \text{funz. anc.} \le q\right) = P\left(-q \le \frac{B_0-\beta_1\tilde{x}-(\beta_0+\beta_1\tilde{x})}{S_e\sqrt{k_m(\tilde{x})}} \le q\right)$$

$$= P\left(-q\,S_e\sqrt{k_m(\tilde{x})} \le B_0+B_1\tilde{x}-(\beta_0+\beta_1\tilde{x}) \le q\,S_e\sqrt{k_m(\tilde{x})}\right)$$

$$= P\left(B_0+B_1\tilde{x}-q\,S_e\sqrt{k_m(\tilde{x})} \le \beta_0+\beta_1\tilde{x} \le B_0+B_1\tilde{x}+q\,S_e\sqrt{k_m(\tilde{x})}\right)$$

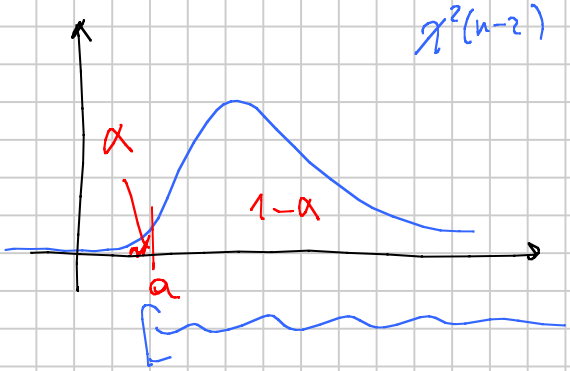
Con livello di confidenza $1-\alpha$, $\beta_0+\beta_1\tilde{x} \in b_0+b_1\tilde{x} \pm q\,s_e\sqrt{k_m(\tilde{x})}$

(l'imbuto arancione di prima) → ?? HW

- Intervallo di confidenza unilaterale sinistro per σ

1) $1-\alpha$ 90%

2)



$a = INV.CHI(1-\alpha\ ;\ n-2)$

3) $1-\alpha = P\left(a \le \frac{S_e^2}{\sigma^2}(n-2)\right) = \dots = P\left(\sigma \le S_e\sqrt{\frac{n-2}{a}}\right)$

$\sigma \le S_e\sqrt{\frac{n-2}{a}}$ con lvl di conf $1-\alpha$

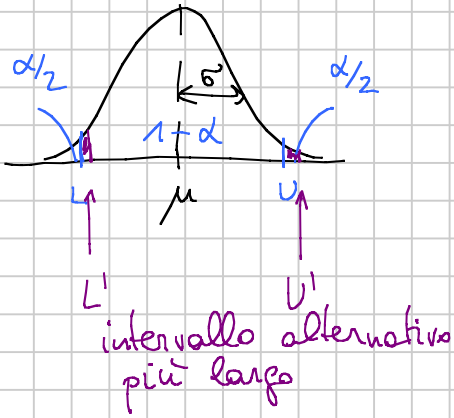
INTERVALLI DI PREDIZIONE

→ sul libro: int. di pred. per la risposta futura nella regressione

- esempio 0

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

voglio prevedere X $\quad X \approx \mu$



trovo L, U t.c. $X \in [L, U]$ con prob. $1-\alpha$ tra le tante scelte possibili, di solito si dà l'intervallo simmetrico: per semplicità e perché è il più stretto.

$L = \text{INV.NORM}(\alpha/2; \mu; \sigma)$

$U = \text{INV.NORM}(1-\alpha/2; \mu; \sigma)$

$1-\alpha = P(L \le X \le U)$

* Si può fare anche unilaterale destro e sinistro

- Esempio 1

$X \sim \mathcal{N}(\mu, \sigma^2)$ μ e σ incognite

$\bar{X}$ S^2 provenienti da un campione

$\underbrace{X_1, X_2, \dots, X_n}_{\text{campione dato}} \sim \mathcal{N}(\mu, \sigma^2)$ ⟶ prevedere $\boxed{X \sim \mathcal{N}(\mu, \sigma^2)}$

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

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

$\frac{X-\bar{X}}{\sigma\sqrt{\frac{1+n}{n}}} \sim \mathcal{N}(0;1)$ ⇨ $\frac{X-\bar{X}}{S\sqrt{\frac{1+n}{n}}} \sim t(n-1)$

indipendenza di S da $X-\bar{X}$

$b = INV.T(\alpha; n-1)$

$a = -INV.T(\alpha; n-1) = -b$

$$1-\alpha = P\left(a \le \frac{X-\bar{X}}{S\sqrt{\frac{1+n}{n}}} \le b\right) = \dots = P\left(\bar{X}+aS\sqrt{\frac{1+n}{n}} \le X \le \bar{X}+bS\sqrt{\frac{1+n}{n}}\right)$$

$$1-\alpha = P\left(X \in \bar{X} \pm bS\sqrt{\frac{1+n}{n}}\right)$$

Esempio 2 : regressione $Y = \beta_0 + \beta_1 x + e$ $\quad e \sim \mathcal{N}(0, \sigma^2)$

risposta futura? $\tilde{x}$ valore ingresso esperimento futuro

$\tilde{Y} = \beta_0 + \beta_1 \tilde{x} + \tilde{e}$ $\quad \tilde{e} \sim \mathcal{N}(0, \sigma^2)$

$$\tilde{Y} \sim \mathcal{N}(\beta_0 + \beta_1 \tilde{x}\ ;\ \sigma^2)$$

$\tilde{x}$ noto

ora ?

$$\beta_0 + \beta_1 \tilde{x} \approx B_0 + B_1 \tilde{x} \sim \mathcal{N}(\beta_0 + \beta_1 \tilde{x}\ ;\ \sigma^2 k_m(\tilde{x}))$$

(x_1, Y_1)
(x_2, Y_2)
$\vdots$
(x_n, Y_n)

campione dato $\longrightarrow$ prevedere $\tilde{Y} \sim \mathcal{N}(\beta_0 + \beta_1 \tilde{x}\ ;\ \sigma^2)$

- B_0
- B_1
- S_e

$B_0 + B_1 \tilde{x} \sim \mathcal{N}(\beta_0 + \beta_1 \tilde{x}\ ;\ \sigma^2 k_m(\tilde{x}))$

$$\tilde{Y} - (B_0 + B_1 \tilde{x}) \sim \mathcal{N}(0\ ;\ \sigma^2 (1 + k_m(\tilde{x})))$$

$$\frac{\tilde{Y} - B_0 - B_1 \tilde{x}}{S_e \sqrt{1 + k_m(\tilde{x})}} \sim t(n-2)$$

$b = INV.T(\alpha; n-2)$

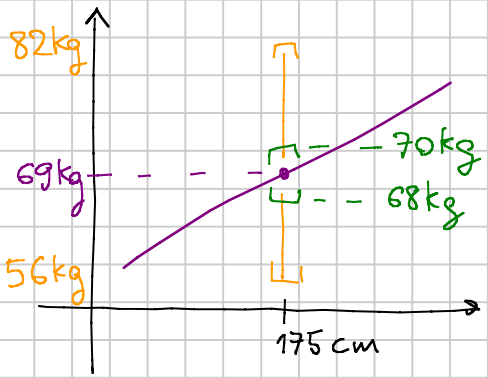
$a = -b$

$$1-\alpha = P\left(\tilde{Y} \in B_0 + B_1 \tilde{x} \pm b\, S_e \sqrt{1 + k_m(\tilde{x})}\right)$$

★ Quasi uguale all'intervallo di confidenza per la risposta media

La differenza è il "1+"

Esempio con statura e peso $\tilde{x} = 175$ cm $\quad B_0 + B_1 \tilde{x} = 69$ kg



una nuova persona alta 175 cm nel 95% dei casi peserà tra i 56 e gli 82 kg (intervallo pred. risp. futura)

il peso medio di tutte le persone alte 175 cm al 95% di conf è compreso tra i 68 kg e i 70 kg (interv. di conf risp. media)

---

ora 9

- Esempio 3 : regressione ; più risposte future

(x_1, Y_1)
(x_2, Y_2)
$\vdots$
(x_n, Y_n)
} campione dato → B_0, B_1, S_e

$(\tilde{x}_1, \tilde{Y}_1)$
$(\tilde{x}_2, \tilde{Y}_2)$
$\vdots$
$(\tilde{x}_m, \tilde{Y}_m)$
} campione futuro

prevedere $\sum_{i=1}^{m} \tilde{Y}_i$

$$\tilde{Y}_1 + \dots + \tilde{Y}_m \sim \mathcal{N}\left(\beta_0 + \beta_1 \tilde{x}_1 + \dots + \beta_0 + \beta_1 \tilde{x}_m \; ; \; \sigma^2 + \dots + \sigma^2\right)$$

$$\sim \mathcal{N}\left(m\beta_0 + \beta_1 \sum_{i=1}^{m} \tilde{x}_i \; ; \; m\sigma^2\right)$$

$$m B_0 + B_1 \sum_{i=1}^{m} \tilde{x}_i \sim \mathcal{N}\left(m\beta_0 + \beta_1 \sum_{i=1}^{m} \tilde{x}_i \; ; \; ?\right)$$

$$\mathrm{Var}\left(m B_0 + B_1 \sum_i \tilde{x}_i\right) = m^2 \mathrm{Var}(B_0) + \left(\sum_i \tilde{x}_i\right)^2 \mathrm{Var}(B_1) + 2m \sum_i \tilde{x}_i \, \mathrm{Cov}(B_0 ; B_1)$$

$$= \sigma^2 \left\{ m^2 k_0 + k_1 \left(\sum_i \tilde{x}_i\right)^2 - 2m \sum_i \tilde{x}_i \, k_{01} \right\} = \sigma^2 k_M$$

vedi ora 6

$$- \sigma^2 \frac{\bar{x}}{n(\overline{x^2} - \bar{x}^2)}$$

$\dots$

$= INV.T(\alpha ; n-2)$ ↑ "Mostro"

$$\sum_{i=1}^{m} \tilde{Y}_i \in m B_0 + B_1 \sum_{i=1}^{m} \tilde{x}_i \pm b \, S_e \sqrt{k_M + m} \quad \text{con prob. } 1-\alpha$$

HW: campione dato $X \sim bin(n, p)$ $\longrightarrow$ prevedere $Y \sim bin(m, p)$

(n: noto, p: incognito, m: noto)

hint: $bin \sim \mathcal{N}$

TEST STATISTICI RELOADED

Ripasso tramite un esempio: inferenza su β_1

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

$\beta_1 = 0 \Rightarrow Y = \beta_0 + e \sim \mathcal{N}(\beta_0; \sigma^2)$ campione normale omogeneo

?

Y non dipende da x

* Y dipendente da x può voler dire che conoscendo o controllando x prevedo meglio Y
* Y non dipendente da x può voler dire che il sistema è robusto (anche se non controllo x, Y non ne risente)

Test: $H_0: \beta_1 = 0$ $\quad H_1: \beta_1 \neq 0$

più in generale

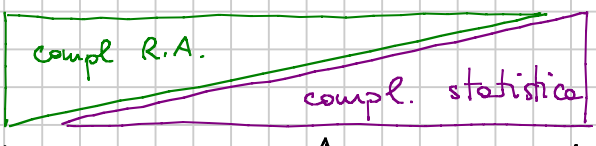
$H_0: \beta_1 = \overline{\beta_1}$	$H_1: \beta_1 \neq \overline{\beta_1}$
$H_0: \beta_1 \leq \overline{\beta_1}$	$H_1: \beta_1 > \overline{\beta_1}$
$H_0: \beta_1 \geq \overline{\beta_1}$	$H_1: \beta_1 < \overline{\beta_1}$

$\overline{\beta_1}$ noto (di solito $= 0$)

i. scelgo il parametro: β_1

ii. scelgo il parametro di confronto, le ipotesi e il livello di significatività: $\overline{\beta_1} = 0$, $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$, $\alpha = 5\%$

iii. scelgo la strategia RA vs statistica



① stimatore

② stat. classica dedotta dalla f. ancillare

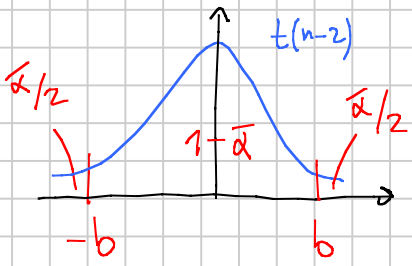
③ p dei dati

iv. ② funzione ancillare per il parametro : $\frac{B_1 - \beta_1}{S_e \sqrt{k_1}} \sim t(n-2)$

stat = f. anc. $\Big|_{\text{param = termine di confr.}}$: $T := \frac{B_1}{S_e \sqrt{k_1}} \overset{H_0}{\sim} t(n-2)$

la statistica ha legge della f. anc. ($t(n-2)$ nell'esempio)
non sempre ma solo se $\beta_1 = 0$ ovvero H_0 vera (al bordo)

v. trovo i quantili per la legge della f. anc. (uso $\bar{\alpha}$ e ipotesi) :



$b = INV.T(\bar{\alpha}\ ;\ n-2)$

vi. fine ② : $RA_T := [-b\ ;\ +b]$ se $T \in RA_T$ dico H_0
se $T \notin RA_T$ dico H_1

vii. ① stimatore B_1 cerco RA_{B_1}

dico $H_0 \Leftrightarrow T \in RA_T \Leftrightarrow -b \le \frac{B_1}{S_e \sqrt{k_1}} \le +b \Leftrightarrow -b\, S_e \sqrt{k_1} \le B_1 \le b\, S_e \sqrt{k_1}$

$\Leftrightarrow B_1 \in \pm b\, S_e \sqrt{k_1}$

deduco $RA_{B_1} := [-b\, S_e \sqrt{k_1}\ ;\ +b\, S_e \sqrt{k_1}]$

viii. ③ per quale valore di $\bar{\alpha}$ cambio decisione?

dico $H_0 \Leftrightarrow T \in RA_T(\bar{\alpha}) \Leftrightarrow |T| \le b = INV.T(\bar{\alpha}\ ;\ n-2) = F^{-1}_{t(n-2)}\left(1 - \frac{\bar{\alpha}}{2}\right)$

ricavo $\bar{\alpha}$

perché $F_{t(n-2)}$ è crescente ($a < b \Rightarrow F(a) \le F(b)$)

$\Leftrightarrow F_{t(n-2)}(|T|) \le 1 - \frac{\bar{\alpha}}{2} \Leftrightarrow \bar{\alpha} \le 2\left(1 - F_{t(n-2)}(|T|)\right)$

definisco $\alpha^* := 2\left(1 - F_{t(n-2)}(|T|)\right)$ p dei dati

$RA_{\alpha^*} := [\bar{\alpha}\ ;\ 1]$ (è la stessa per tutti i test in tutti i casi)

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

↑ test bilaterale

ora 10

Rispetto all'anno scorso, saltiamo all'ora 31

- Esempio 2 : $X_i \sim \mathcal{N}(\mu, \sigma^2)$ σ nota, test su μ $n = 12$ ($\sigma = 2$ ml)

  ↑ riempimento bottiglie di sugo

  i. μ livello medio di riempimento attuale ... incognito

  ii. $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $\mu_0 = 995$ ml target desiderato dall'azienda

  $\alpha = 1\%$ deciso non in base a ragioni statistiche

  target = AQL (acceptable quality level)

  iii. iv. (2) $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \longrightarrow z := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ statistica

  v. vi $RA_z := \pm q$ $q = \text{INV.NORM.ST}\left(1 - \frac{\alpha}{2}\right) = 2{,}58$

  vii. (1) $|z| \leq q \iff \bar{X} \in \mu_0 \pm q\frac{\sigma}{\sqrt{n}}$ $RA_{\bar{X}} = [993{,}5 ; 996{,}5]$

★ Curva OC = P(dire H_0) in funzione del parametro : μ

$h = h(\mu)$

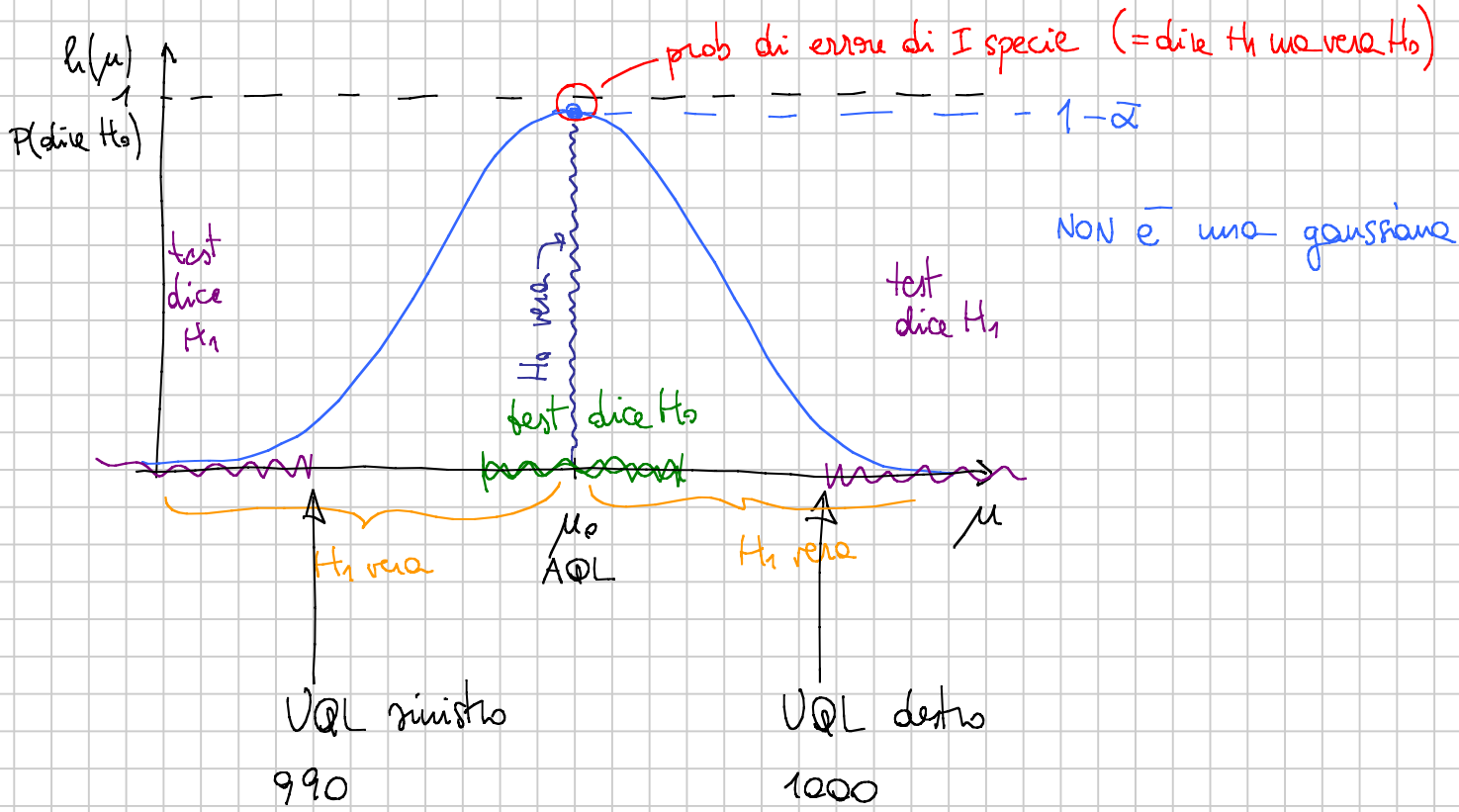
$$h(\mu) = P(\text{dire } H_0; \mu) = P(\bar{X} \in RA_{\bar{X}}; \mu) = F_{\bar{X}}(996{,}5) - F_{\bar{X}}(993{,}5)$$

$\bar{X} \sim ?$ $\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$

$$= \Phi\left(\frac{\mu_0 + q\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\mu_0 - q\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right)$$

$$= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + q\right) - \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - q\right)$$

$$= \text{DISTRIB.NORM}(996{,}5; \mu; \sigma/\text{RADQ}(n)) - \text{DISTRIB.NORM}(993{,}5; \mu; \sigma/\text{RADQ}(n))$$



★ h basso su VQL e alto su AQL

Per costruire un nuovo test con caratteristiche migliori, possiamo agire su $\bar{\alpha}$ e su n

→ se alzo $\bar{\alpha}$ la RA si stringe e h si abbassa ovunque

→ se alzo n spendo più soldi, ma la curva diventa più ripida e quindi il test migliora

★ Def la potenza di un test è la P(dire H_1) quando è vera H_1 $= 1 - h(\mu)$ considerata solo per $\mu \in H_1$, $\mu \neq \mu_0$

★ Def la prob di errore di II specie β è 1 − potenza

$\beta = \beta(\mu) = h(\mu)$ considerata solo per $\mu \in H_1$, $\mu \neq \mu_0$

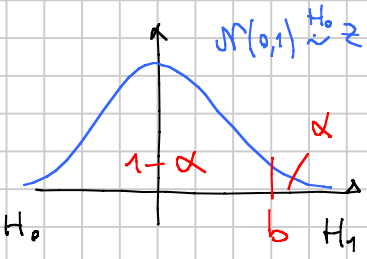
CURVE OC

- Esempio: controlli qualità (in ingresso)
  fornitura (migliaia di pezzi)
  estraggo un campione casuale di $n = 25$ pezzi ($n \approx 30-100$)
  li testo tutti e conto i difettosi: $X \sim bin(n, p)$ ← parametro incognito: difettosità reale della fornitura

ipotesi: $H_0: p \le p_0 = 10\%$ $H_1: p > p_0$
lvl di sign: $\alpha = 5\%$

$\hat{p} := \frac{X}{n} \approx p$ stimatore

$\hat{p} \sim \mathcal{N}\left(p; \frac{p(1-p)}{n}\right)$ $Z := \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} \mathcal{N}(0,1)$



$b = \text{INV.NORM.ST}(1-\alpha) = 1{,}645$
$RA_Z = (-\infty; b]$

dico $H_0 \iff Z \le b \iff \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \le b \iff \hat{p} \le p_0 + b\sqrt{\frac{p_0(1-p_0)}{n}} \cong 19{,}8\%$

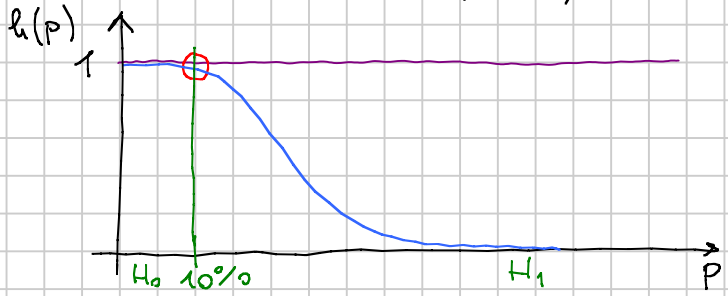
$RA_{\hat{p}} = [0; 19{,}8\%]$

$\frac{X}{n} = \hat{p} \le 19{,}8\% \iff X \le 25 \cdot 19{,}8\% = 4{,}95$

$RA_X = [0; 4]$ arrotondare è un passo critico

★ Adesso parto dalla fine: (n, RA_X) e verifico le caratteristiche del test → $\alpha = ?$ $p_0 = ?$ $h = ?$ ... n è adeguato?

→ Calcolo $h = h(p) = P(\text{dire } H_0; p) = P(X \le 4; p)$
$= P(bin(25; p) \le 4) = \text{DISTRIB.BINOM}(4; 25; p)$



α : $P(\text{dire } H_1 \mid H_0)$ prob errore I specie

β : $P(\text{dire } H_0 \mid H_1)$ prob errore II specie

α è un numero se H_0 è bilaterale $\quad \beta = \beta(p)$ dipende da p

in caso di ipotesi unilaterali

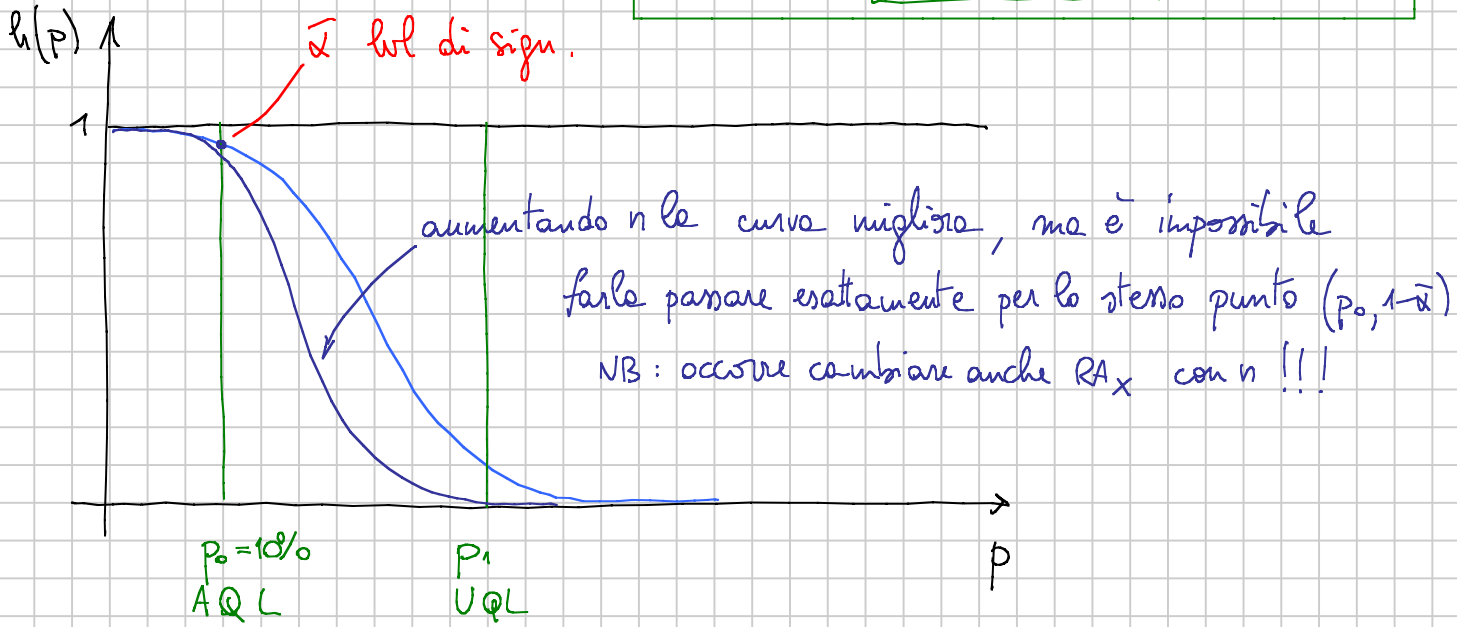
$\alpha = \alpha(p) = 1 - h(p) \qquad p \in H_0$

$\beta = \beta(p) = h(p) \qquad p \in H_1$

$\bar{\alpha}$ il livello di significatività

il test viene costruito in modo tale da controllare $\boxed{\alpha \leq \bar{\alpha}}$

$$\boxed{\bar{\alpha} = 1 - h(p_0)}$$



→ Se il livello di qualità è migliore di AQL, il test è fatto in modo di accettare quasi sempre

→ Se il livello di qualità è peggiore di UQL, il test è fatto in modo di rifiutare quasi sempre

✱ AQL e UQL si "leggono" sul grafico e corrispondono a valori di h alti (90~95%) e bassi (10~20%).

↳ Ad esempio se con il p_0 "dichiarato" si trovasse $h(p_0) = 99{,}97\%$, il vero AQL sarebbe più alto del p_0 dichiarato

Confronto di varianze per campioni normali

$\mathcal{N}(\mu, \sigma^2)$ $\mathcal{N}(\nu, \tau^2)$ due popolazioni

Q_1: $\sigma = \tau$? Q_2: $\sigma = 2\tau$? Q_*: $\sigma = \lambda\tau$ $\lambda > 0$ qualsiasi

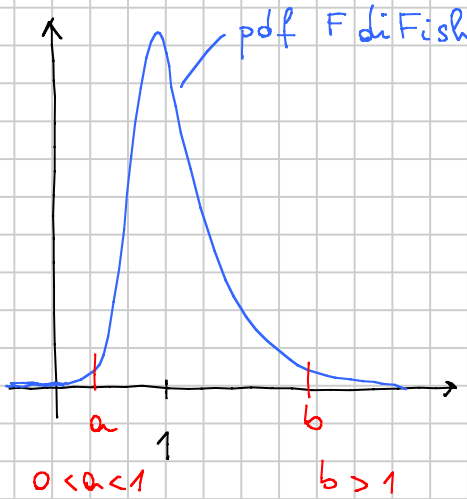
$X_1, X_2, \dots, X_m$ $Y_1, Y_2, \dots, Y_n$

$\bar{X}$, $S_X^2 \approx \sigma^2$ $\bar{Y}$, $S_Y^2 \approx \tau^2$

Q_1 $H_0: \sigma^2 = \tau^2$ $H_1: \sigma^2 \neq \tau^2$ (si può fare anche unilaterale)

funzione ancillare: $$\frac{S_X^2/\sigma^2}{S_Y^2/\tau^2} = \frac{S_X^2}{S_Y^2}\cdot\frac{\tau^2}{\sigma^2} \sim F(m-1; n-1)$$

distribuzione F di Fisher con $m-1$ gdl al numeratore ed $n-1$ gdl al denominatore



DISTRIB.F$(x; \text{gdlnum}; \text{gdlden}) = 1 - F(x)$

INV.F$(p; \dots) = F^{-1}(1-p)$

(HW Check)

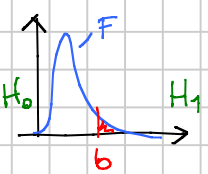
$H_0: \frac{\sigma}{\tau} = 1$ $H_1: \frac{\sigma}{\tau} \neq 1$

$$\frac{S_X^2}{S_Y^2}\Big/\left(\frac{\sigma}{\tau}\right)^2 \sim F(m-1; n-1)$$

→ statistica $F := \frac{S_X^2}{S_Y^2}$ $RA_F = [a; b]$

Curve OC nel caso tipico:

$H_0: \sigma^2 \leq \tau^2$ $H_1: \sigma^2 > \tau^2$ $F = \frac{S_X^2}{S_Y^2}$

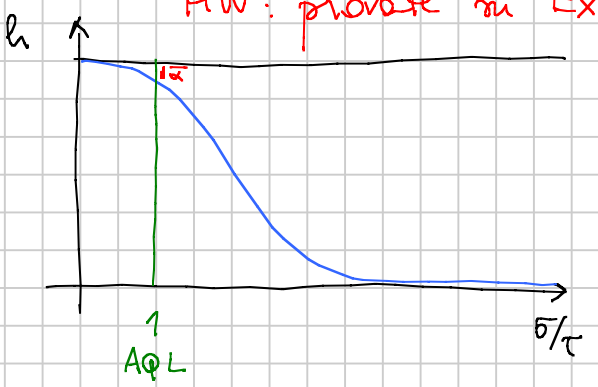


$RA_F: [0; b]$ $b = \text{INV.F}(\alpha; m-1; n-1)$

$$h = h(\sigma/\tau) = P(\text{dire } H_0; \sigma/\tau) = P\left(\frac{S_X^2}{S_Y^2} \leq b; \frac{\sigma}{\tau}\right)$$
$$= P\left(\frac{S_X^2}{S_Y^2}\Big/\frac{\sigma^2}{\tau^2} \leq b\Big/\frac{\sigma^2}{\tau^2}; \frac{\sigma}{\tau}\right) = P\left(F(m-1;n-1) \leq b\Big/\frac{\sigma^2}{\tau^2}\right)$$

$= 1 - \text{DISTRIB. F}\left(b \,/\, \frac{\sigma^2}{\tau^2}\,;\, m-1\,;\, n-1\right)$

HW: provate su Excel $\bar{x}$, m, n



back to ... regressione lineare semplice

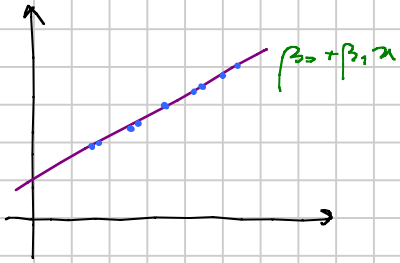
VALUTAZIONE DEL MODELLO

- Coefficiente di determinazione R^2

→ rappresenta la quantità di informazione su Y contenuta in x

→ $R^2 \in [0, 1]$ $\quad R^2 = 0 \iff B_1 = 0$ niente dipendenza

$R^2 = 1 \iff \sigma = 0 = S_e$ dipendenza lineare perfetta



R^2 = errore se $B_1 = 0$ e $S_e = 0$

→ incertezza di Y può essere misurata tramite:

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

varianza campionaria $\qquad$ devianza campionaria

→ incertezza di Y residua se si suppone di conoscere x

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - (B_0 + B_1 x_i)\right)^2 \qquad SS = \sum_{i=1}^{n}\left(Y_i - (B_0 + B_1 x_i)\right)^2$$

errore standard al quadrato $\qquad$ devianza dei residui

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

frazione di devianza spiegata dalle x

$$R^2_{adj} := 1 - \frac{S_e^2}{S_Y^2}$$ "R^2 corretto"

★ Il coefficiente di determinazione R^2 è legato alla correlazione campionaria

Cov. campionaria:

$$\frac{1}{n-1}\left(\sum x_i Y_i - n\bar{x}\bar{Y}\right)$$

$$= \frac{n}{n-1}\left(\overline{xY} - \bar{x}\bar{Y}\right)$$

Corr. campionaria:

$$\frac{\frac{n}{n-1}\left(\overline{xY} - \bar{x}\bar{Y}\right)}{\sqrt{\frac{n}{n-1}\left(\overline{Y^2} - \bar{Y}^2\right)\frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)}}$$

$$= \frac{\overline{xY} - \bar{x}\bar{Y}}{\sqrt{\left(\overline{Y^2} - \bar{Y}^2\right)\left(\overline{x^2} - \bar{x}^2\right)}} = r$$

$$-1 \le r \le 1 \qquad r = \pm 1 \iff Y_i = \beta_1 x_i + \beta_0 \quad \forall i$$

$$r^2 = R^2$$

★ r andrebbe sempre considerato al quadrato.

$X, Y \to Cov(X;Y)$

$$\rho(X;Y) = \frac{Cov(X;Y)}{\sqrt{Var(X)\,Var(Y)}}$$

correlazione lineare

$-1 \le \rho \le 1$

$\rho = \pm 1 \iff Y = \alpha X + \beta$

$$S_Y^2 = \frac{1}{n-1}\left(\sum Y_i^2 - n\bar{Y}^2\right) = \frac{n}{n-1}\left(\overline{Y^2} - \bar{Y}^2\right)$$

MMSD

ora 13

Note Title 25/10/2011

VALUTAZIONE DEL MODELLO

- $H_0: \beta_1 = 0$ $\quad H_1: \beta_1 \neq 0$ dice se "vi è regressione" ovvero se x contribuisca di sicuro a conoscere Y

$$x \xrightarrow{\ ?\ } Y$$

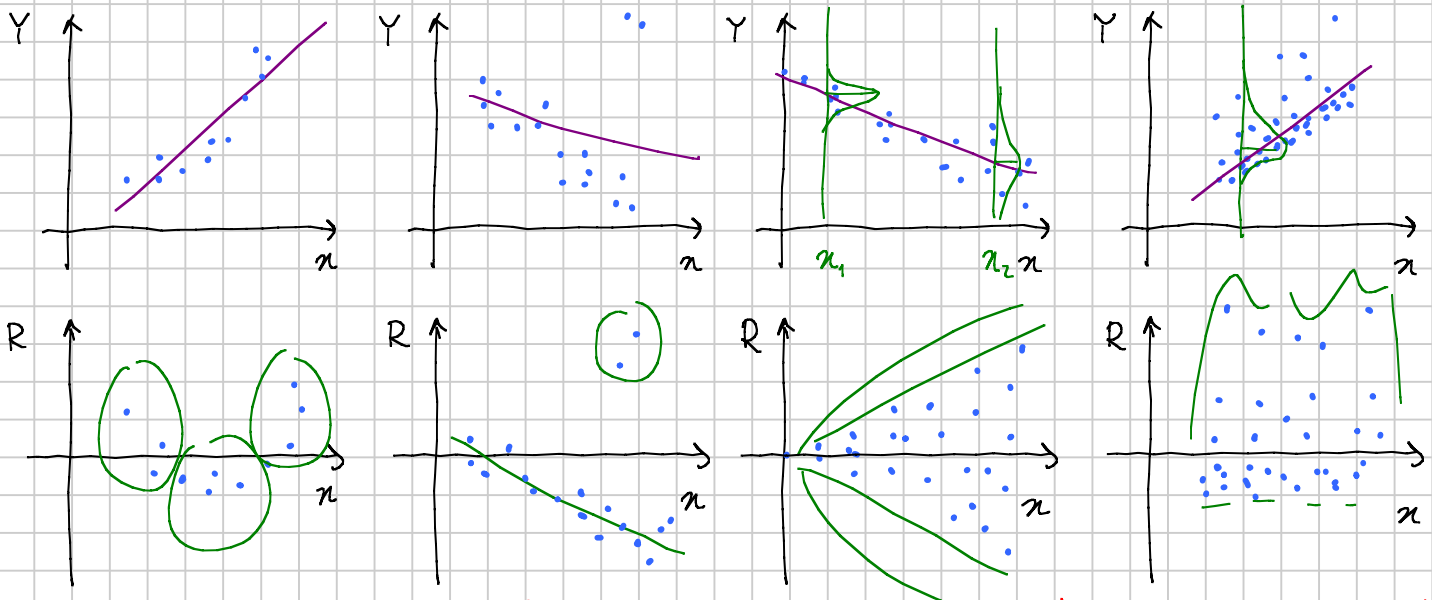
dice se la relazione "esiste"

R^2 dice quanto forte è la relazione

- ANALISI DEI RESIDUI

Dice se il modello è adeguato (altrimenti occorre migliorarlo)

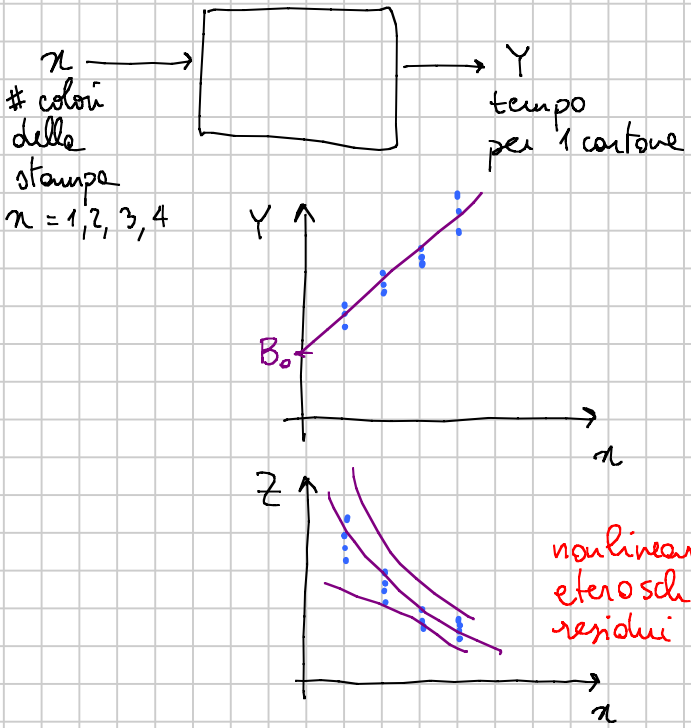
$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2) \text{ indipendenti}$$



nonlinearità	outliers	eteroschedasticità	non gaussianità dei residui
$E(Y) = f(x)$	vanno capiti, spiegati, trattati a parte e tolti	σ^2 non costante	$e \not\sim \mathcal{N}$
II	I	(III)	((?))

- Panacea della regressione: trasformazioni nonlineari

esempio: linea per cartoni da imballaggio



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

Z veramente usata al posto di Y:

$$Z = \frac{1 \text{ ora}}{Y}$$ velocità [cartoni/ora]

nonlineare
eterosch.
residui non $\mathcal{N}$

$$Z = \frac{1}{\beta_0 + \beta_1 x + e} = \gamma_0 + \gamma_1 x + \tilde{e}$$

forzature

★ Soluzione: applicare la trasf. nonlineare $Z \longrightarrow \frac{1}{Z} =: Y$ corregge tutti i difetti solo in questo esempio!

→ $1/Y$ $\log Y$ $\sqrt{Y}$ Y^2 ... potrebbero migliorare il modello

→ anche trasf. nonlineari di x possono aiutare, ma di meno

$$Y = \beta_0 + \beta_1 x + e \qquad z = \frac{1}{x} \qquad Y = \beta_0 + \frac{\beta_1}{z} + e$$

nonlineare omosch. $\mathcal{N}$

★ Le trasformazioni lineari di x e/o Y non cambiano la sostanza della regressione.

→ Cambiano proporzionalmente i coefficienti β_0, β_1 e il valore di σ

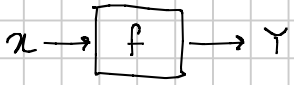
$$Y = 27 + 2{,}4x + \mathcal{N}(0, 3^2) \qquad Y \to \tilde{Y} = 2Y$$

$$\tilde{Y} = 54 + 4{,}8x + \mathcal{N}(0, 6^2) \qquad x \to \tilde{x} = 3x$$

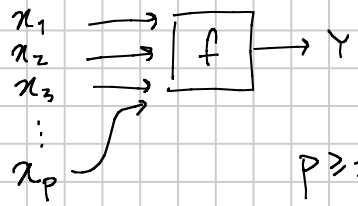
$$\tilde{Y} = 54 + 1{,}6\tilde{x} + \mathcal{N}(0, 6^2)$$

→ Non cambiano R^2, il grafico dei residui, il risultato dei test

REGRESSIONE LINEARE MULTIPLA

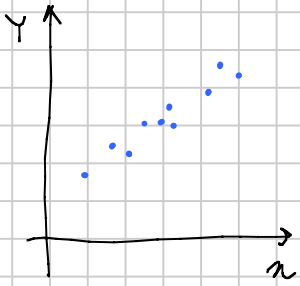


$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, 2, \dots, n$



$p \geq 1$

x	Y
x_1	Y_1
x_2	Y_2
$\vdots$	$\vdots$
x_n	Y_n



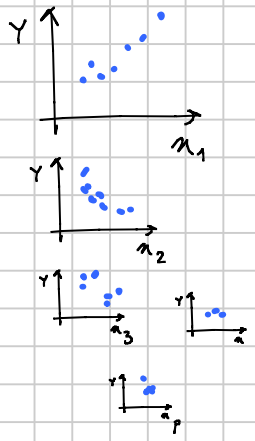
$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + e_i$$

$$Y(i) = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + \dots + \beta_p x_p(i) + e(i)$$

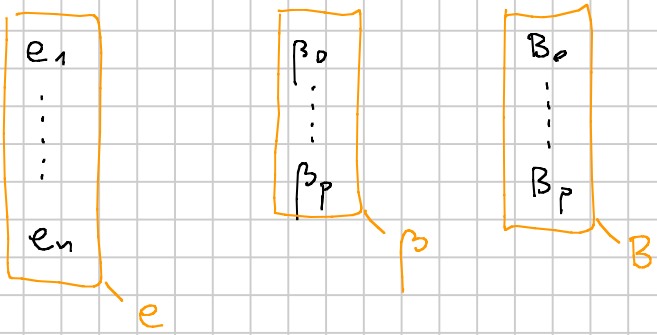
due notazioni

x_0	x_1	x_2	$\dots$	x_p	Y
1	x_{11}	x_{12}		x_{1p}	Y_1
1	x_{21}				
$\vdots$				$\vdots$	$\vdots$
1	x_{n1}	$\dots$	$\dots$	x_{np}	Y_n

X — dummy variable — Y_i



Altri vettori:



$X \in M_{n,p+1}$

$R, e, Y \in \mathbb{R}^n = M_{n,1}$

$\beta, B \in \mathbb{R}^{p+1} = M_{p+1,1}$

$$Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + e_i = \sum_{k=0}^{p} \beta_k X_{i,k} + e_i = [X\beta + e]_i$$

$$\boxed{Y = X\beta + e}$$

$(n,1) = (n,p+1) \cdot (p+1,1) + (n,1)$

$e \sim \mathcal{N}(0, \sigma^2)$ indipendenti

$R = Y - XB$ residui

$SS = \sum_{i=1}^{n} R_i^2 = \|R\|^2$ da minimizzare in funzione di B

$\nabla_B SS = 0 \iff \dots \iff X^T X B - X^T Y = 0$

$$(X^TX)B = X^TY$$

$$\Downarrow$$

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

attenzione alle dimensioni

$$(p+1,n)\cdot(n,p+1)\cdot(p+1,n)\cdot(n,1) = (p+1,1)$$

MATR. TRASPOSTA

MATR. PRODOTTO

MATR. INVERSA

X^TX e $(X^TX)^{-1} \in M_{p+1,p+1}$ e simmetriche

$$(AB)^T = B^TA^T \qquad (X^TX)^T = X^TX$$

$$A^{-1\,T} = A^{T\,-1}$$

★ X^TX è invertibile se $\det(X^TX) \neq 0$

$\det(X^TX) = 0$ solo se c'è una variabile calcolabile linearmente dalle altre: in quel caso il rango di X è minore di $p+1$ e anche quello di X^TX

1	= 0	+ 1	+ 0	
1	= 1	+ 0	+ 0	
1	1	0	0	
1	0	0	1	
1	1	0	0	
1	0	1	0	
1	0	0	1	

★ Relazioni lineari anche approssimative tra le variabili di ingresso:

fatturato $\approx$ 2,4 · tasse pagate + spese − guadagno

si dicono multicollinearità, o si dice che le variabili sono correlate. Ne segue che $\det(X^TX)$ è piccolo e la formula

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

Conseguenza pratica: $k_0, k_1, k_m, \dots, k_{0,1}$ vengono tutti grossi

→ intervalli di conf e predizione larghi

→ test statistici poco potenti

- Distribuzione di B $\qquad Y \sim \mathcal{N}(X\beta, \sigma^2)$

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

- Not: Matrice di covarianza di un vettore aleatorio

X vettore aleatorio $\quad X = (X_1, \dots, X_m)$ $\quad X_j$ variabile aleatoria

$C(X)$ la matrice di covarianza di X

$C(X) \in M_{m,m}$

$[C(X)]_{ij} := \mathrm{Cov}(X_i; X_j) \qquad i,j = 1, 2, \dots, n$

HW: $C(Y) = \sigma^2 I \qquad I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

$N \in M_{n,m} \qquad NY \in M_{n,1}$

$C(NY) = N C(Y) N^T$

MMSD

ora 15

Note Title 08/11/2011

$E(B_i) = \beta_i \quad i = 0, 1, \dots, p \qquad E(B) = \beta$ stimatore corretto

$Var(B_i) = ? \qquad C(B)$ matrice di covarianza di B

★ $C(B) \in M_{p+1,p+1}$

★ $[C(B)]_{ij} := Cov(B_i ; B_j)$ $\qquad Cov(B_0 ; B_1)$

↳ $[C(B)]_{ii} = Var(B_i)$ $\qquad Var(B_0) \quad Var(B_1)$

$B = \underbrace{(X^T X)^{-1} X^T}_{N \in M_{p+1,n}} Y = NY$

Grazie all' HW :

$$C(B) = C(NY) = N C(Y) N^T = \sigma^2 (X^T X)^{-1} X^T I \left[(X^T X)^{-1} X^T\right]^T$$

$$= \sigma^2 (X^T X)^{-1} X^T X \left[(X^T X)^T\right]^{-1} = \sigma^2 (X^T X)^{-1} \cancel{X^T X} \cancel{(X^T X)^{-1}}$$

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

$$\longrightarrow Var(B_i) = \sigma^2 \left[(X^T X)^{-1}\right]_{ii}$$

★ Verifico lo HW

(a) $Y_i \sim \mathcal{N}\left(\sum_{j=0}^{p} x_{ij} \beta_j \; ; \; \sigma^2\right)$ indipendenti

$Var(Y_i) = \sigma^2 = [C(Y)]_{ii}$

$Cov(Y_i ; Y_j) = 0 \qquad i \neq j$

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

(b) $[C(NY)]_{ij} := Cov([NY]_i ; [NY]_j) = Cov\left(\sum_k N_{ik} Y_k \; ; \sum_h N_{jh} Y_h\right)$

$C(NY)$: deterministica

$$= \sum_k \sum_h N_{ik} N_{jh} \operatorname{Cov}(Y_k; Y_h) = \sum_k \sum_h N_{ik} N_{jh} [C(Y)]_{kh}$$

$$\sum_k \sum_h N_{ik} [C(Y)]_{kh} N^T_{hj} = [N C(Y) N^T]_{ij}$$

★ Distribuzione di B

$$\boxed{B \sim \mathcal{N}(\beta; \sigma^2 (X^T X)^{-1})}$$

● Distribuzione di S_e, stimatore di σ

$$R_i = Y_i - \sum_{j=0}^{p} x_{ij} B_j \approx \sigma$$

$$SS = \sum_{i=1}^{n} R_i^2 \qquad \boxed{S_e^2 := \frac{SS}{n-p-1} = \frac{1}{n-p-1} \sum_{i=1}^{n} R_i^2}$$

Perché $n-p-1$? Teorema di Cochran. Vediamo in forma intuitiva

$$Y_i \sim \mathcal{N}\left(\sum_{j=0}^{p} x_{ij}\beta_j ; \sigma^2\right) \quad \text{indipendenti}$$

$$\frac{Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})}{\sigma} \sim \mathcal{N}(0,1) \quad \text{indipendenti}$$

$$\sum_{i=1}^{n} \left[\frac{Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})}{\sigma}\right]^2 \sim \chi^2(n)$$

$$\sum_{i=1}^{n} \left[\frac{Y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \dots + B_p x_{ip})}{\sigma}\right]^2 \sim \chi^2(n-(p+1))$$

thm di Cochran (2)

\# parametri sostituiti

(b) $(B_0, B_1, \dots, B_p)$ indipendente da

$$\boxed{\frac{SS}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} R_i^2 \sim \chi^2(n-p-1)}$$

$$E\left(\frac{SS}{\sigma^2}\right) = n-p-1 \qquad E\left(\underbrace{\frac{SS}{n-p-1}}_{\text{stimatore corretto di } \sigma^2}\right) = \sigma^2$$

$$\frac{SS}{\sigma^2} \sim \Gamma\left(\frac{n-p-1}{2}; 2\right)$$

$$SS \sim \Gamma\left(\frac{n-p-1}{2}; 2\sigma^2\right)$$

$$S_e^2 \sim \frac{SS}{n-p-1} \sim \Gamma\left(\frac{n-p-1}{2}; \frac{2}{n-p-1}\sigma^2\right)$$

$\Gamma(\alpha,\beta)$ $\begin{cases} E(\Gamma(\alpha,\beta)) = \alpha\beta \\ Var(\Gamma(\alpha,\beta)) = \alpha\beta^2 \end{cases}$

$\chi^2(m) \sim \Gamma\left(\frac{m}{2}, 2\right)$

$W \sim \Gamma(\alpha,\beta) \quad cW \sim \Gamma(\alpha, \beta c)$

$$E(S_e^2) = \sigma^2$$

$$Var(S_e^2) = \frac{2}{n-p-1}\sigma^4$$

siccome $\searrow 0$ al crescere di n S_e^2 è anche consistente

ora 16

THM DI COCHRAN

Ipotesi:

i. $Y_1, Y_2, \dots, Y_n$ $\quad Y_i \sim \mathcal{N}(\mu_i; \sigma^2)$ indipendenti

ii. M, N due matrici $\quad M \in M_{n,k} \quad N \in M_{k,n}$ di rango k entrambe $k < n$

iii. $E(Y - MNY) = 0$ $\quad$ ovvero $\quad E([MNY]_i) = \mu_i$

Tesi

a) $\dfrac{\|Y-MNY\|^2}{\sigma^2} \sim \chi^2(n-k)$

b) $\|Y-MNY\|^2$ è indipendente da NY

★ Chiamo $B = NY \qquad B = (B_1, B_2, \dots, B_k)$

$B_i = \sum_{j=1}^{n} N_{ij} Y_j$ $\quad$ B_i sono combinazioni lineari delle Y_j

N di rango $k \iff$ le k righe di N sono linearmente indip.

↳ Nel caso della regressione $\beta = \underbrace{(X^TX)^{-1}X^T}_{N} Y$ se X ha rango $p+1$, N ha rango $p+1$ altrimenti X^TX non è invertibile

★ $R_i = Y_i - \sum_{j=0}^{p} x_{ij} B_j$ $\qquad R = Y - XB = Y - MNY$

$\uparrow M$

i. ii. verificati $\qquad$ iii. $E(R) = 0$ vero

a) $\|Y - MNY\|^2 = \|R\|^2 = \sum_{i=1}^{n} R_i^2 = SS \qquad \frac{SS}{\sigma^2} \sim \chi^2(n-p-1)$

b) SS indipendente da B.

HW: Applicate il thm di Cochran per verificare

i. $S^2 \sim \chi^2(n-1)$ indipendente da $\bar{X}$

ii. $S_p^2 \sim \chi^2(n+m-2)$ indipendente da $(\bar{X}, \bar{Y})$

■ Funzioni ancillari e inferenza

β_i $\qquad \frac{B_i - \beta_i}{\sigma \sqrt{[(X^TX)^{-1}]_{ii}}} \sim N(0,1) \qquad \boxed{\frac{B_i - \beta_i}{S_e \sqrt{[(X^TX)^{-1}]_{ii}}} \sim t(n-p-1)}$

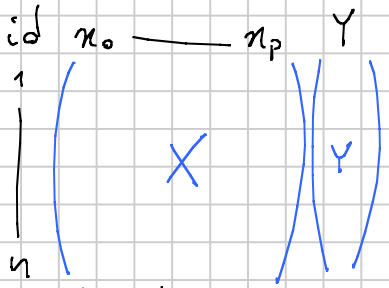
σ $\qquad \boxed{\frac{S_e^2}{\sigma^2}(n-p-1) \sim \chi^2(n-p-1)}$

HW: trovare la funzione ancillare per $\beta_i - \beta_j$

→ intervalli di confidenza, intervalli di predizione e test statistici su β_i, σ, $\beta_i - \beta_j$, ...

Quello che segue è spiegato nell'ora 19 (non dall'inizio)

● Inferenza su risposte future



dati del passato
aka gruppo di addestramento

↳ B ↳ $(X^TX)^{-1}$ ↳ S_e ↳ R_x^2 ...

$$\text{weight} \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

una o più persone nuove $1, 2, \dots, m$

conosciamo le x $\longrightarrow$ prevedere le Y

$$\tilde{x}_1 = (\tilde{x}_{11}, \tilde{x}_{12}, \dots, \tilde{x}_{1p})$$
$$\tilde{x}_2 = \vdots$$
$$\vdots$$
$$\tilde{x}_m = (\tilde{x}_{m1} \dots \tilde{x}_{mp})$$

non ho le Y, le voglio prevedere $\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_3$

$$\tilde{Y}_i \approx \sum_{j=0}^{p} \beta_j \tilde{x}_{ij}$$ stima puntuale

$\longrightarrow$ Che distribuzione ha $\sum_{j=0}^{p} \beta_j \tilde{x}_{ij} = \beta \cdot \tilde{x}_i$?

$Z \sim \mathcal{N}(\mu, \Sigma)$, $N \in M_{c,d}$, $NZ \in \mathbb{R}^c$ aleatorio

(Z: $\mathbb{R}^d$ aleatorio; μ: $\mathbb{R}^d$; Σ: $M_{d,d}$)

$$NZ \sim \mathcal{N}(N\mu; N\Sigma N^T)$$

$$\tilde{x}_i \cdot \beta = (\tilde{x}_{i0}, \tilde{x}_{i1}, \tilde{x}_{i2}, \dots \tilde{x}_{ip}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = N\beta$$

$N \in M_{1,p+1}$; $N\beta$: var aleat $\mathbb{R}^1$

$$N\beta \sim \mathcal{N}(N\beta; NC(\beta)N^T) \sim \mathcal{N}(\tilde{x}_i \cdot \beta; \sigma^2 \tilde{x}_i \cdot (X^T X)^{-1} \tilde{x}_i)$$

$N\beta = \beta \cdot \tilde{x}_i$; $C(\beta) = \sigma^2 (X^T X)^{-1}$

$\longrightarrow$ funzione ancillare di $\sum_{j=0}^{p} \beta_j \tilde{x}_{ij}$

$$\frac{\sum_{j=0}^{p} \beta_j \tilde{x}_{ij} - \left(\sum_{j=0}^{p} \beta_j \tilde{x}_{ij}\right)}{Se\sqrt{\tilde{x}_i^T (X^T X)^{-1} \tilde{x}_i}} \sim t(n-p-1)$$

HW: Ricavare intervallo di confidenza per $\sum_{j=0}^{p} \beta_j \tilde{x}_{ij} = E(\tilde{Y}_i)$ la risposta media per tutti gli esperimenti con var di ingresso $\tilde{x}_i$

Ricavare intervalli di predizione per

a. $\tilde{Y}_i$

b. $\sum_{i=1}^{m} \tilde{Y}_i$ $\quad$ (d. $\sum_{i=1}^{m} \alpha_i \tilde{Y}_i$)

c. $\frac{1}{m} \sum_{i=1}^{m} \tilde{Y}_i$

Da qui si ritorna con l'audio all'ora 16

SELEZIONE DELLE VARIABILI

non c'è, ovvero $\beta_4 = 0$

$$CS = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

- CS: resistenza calcestruzzo
- x_1: cemento
- x_2: sabbia fine
- x_3: sabbia grossa
- x_4: detriti
- x_5: acqua

In pratica $B_4 \approx 0$ e il test: $H_0: \beta_4 = 0$ $\quad$ $H_1: \beta_4 \neq 0$ dirà H_0

Allora elimino la variabile x_4 e rifaccio la regressione

Metodo stepwise backward: prevede di fare regressioni una dopo l'altra, partendo con tutte le variabili e togliendo via via quelle per cui il test dica H_0

1) Correzione di Bonferroni

200 geni possibili

esiste un test statistico: H_0: gene non corr H_1: gene corr con morbo

$\alpha = 5\%$

purtroppo il gene giusto non è tra quei 200

su 200 test, ciascuno dei quali ha il 5% di prob di avere un falso positivo, il numero di test positivi bin(200; 5%)

ovvero in media 10 test saranno positivi (4 – 16)

Paradosso dei falsi positivi nei test multipli

→ Correzione di Bonferroni: se devo fare n test e voglio lvl di significatività globale α, i singoli test vanno fatti con lvl di sign $\frac{\alpha}{n}$
i falsi positivi: $\text{bin}(n; \frac{\alpha}{n})$ in media sono $\alpha \ll 1$

→ Nella selezione stepwise, un test $\beta_i = 0$ che dice H_1 con lvl di sign $\frac{\alpha}{p}$ è veramente "sicuro" di H_1, altrimenti è solo un indizio

2) Multicollinearità

Dico che le variabili di ingresso sono linearmente dipendenti se esiste una relazione lineare esatta tra le colonne (tra le variabili)
Dico che vi è multicollinearità tra le var di ingresso se vi sono relazioni approssimate (ad esempio se una variabile è prevedibile tramite regressione sulle altre)

★ Tranne che nel D.o.E, diamo sempre per buono che ci sia.

---

ora 17

★ Effetto principale: X^TX con determinante piccolo
$(X^TX)^{-1}$ e $\sigma^2(X^TX)^{-1}$ avranno coefficienti grossi
quindi $\hat{B}$ avrà varianze (e covarianze) elevate e l'inferenza sarà cattiva: intervalli larghi e test poco potenti

★ Effetto secondario: tra una regressione con tutte le variabili e una con una variabile in meno, i coefficienti B_i e i p dei dati relativi ai test $\beta_i = 0$ possono cambiare moltissimo.

$Y = \sum \beta_i x_i$ $\quad x_1 \approx x_2$ $\quad Y = B_0 + 2x_1 + 7x_2 + \sum_3^p B_i x_i$

$H_0 : \beta_1 = 0$ $\quad H_1 : \beta_1 \neq 0$ $\quad \alpha_1^* = 89\%$ $\quad \approx B_0 + 9x_1 + \sum_3^p B_i x_i$

$H_0 : \beta_2 = 0$ $\quad H_1 : \beta_2 \neq 0$ $\quad \alpha_2^* = 94\%$ $\quad \approx B_0 + 9x_2 + \sum_3^p B_i x_i$

$\approx \dots + 27x_1 - 18x_2 + \dots$

$Y = \beta_0 + \beta_1 x_1 + \sum_3^p \beta_i x_i$ $\quad Y = B_0 + 9x_1 + \sum_3^p B_i x_i$

$H_0 : \beta_1 = 0$ $\quad H_1 : \beta_1 \neq 0$ $\quad \alpha_1^* = 10^{-7}$

★ Mai togliere più di una variabile alla volta

★ Normalmente le var con pdd alto si tolgono in ordine di pdd, in realtà non è per forza giusto: andrebbero fatti vari tentativi

3) Quando fermarsi?

→ quando tutti i pdd dei test $\beta_i = 0$ delle variabili selezionate sono per me significativi, ovvero al di sotto del valore $\bar{\alpha}$ che sto considerando

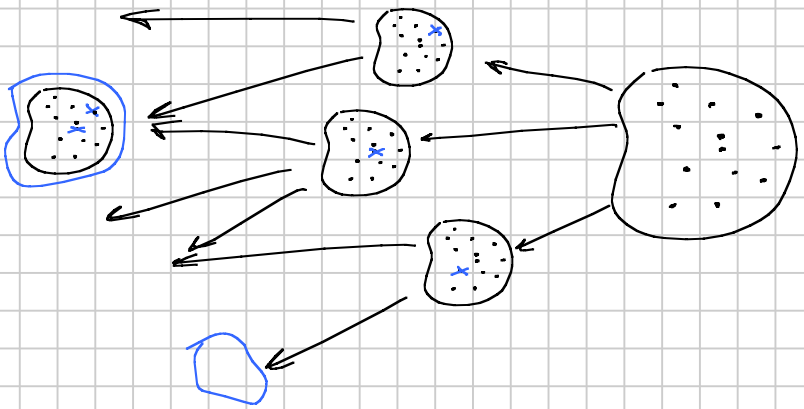
★ Cor di Bonferroni : $\bar{\alpha} \approx 1‰$

★ per modello predittivo : $\bar{\alpha} \approx 25-30\%$

Vendite quotidiane : giorno settimana ; periodo dell'anno ; inserto ; dimensioni font ; .... ; .... ; ....

→ produzione : prevedere le vendite → $\bar{\alpha} = 30\%$

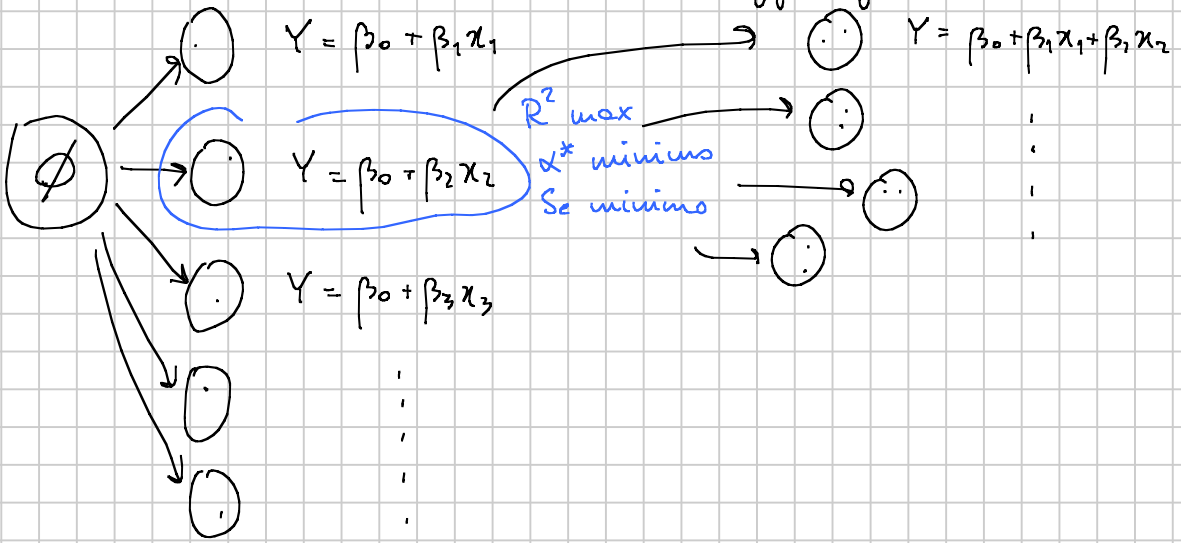
→ marketing : scelte comerciale → $\bar{\alpha} = 1‰$



A seconda della strada ci si ferma in posti eventualmente diversi

Metodo stepwise forward

Si parte con 0 variabili e si aggiungono una alla volta



* Ci si ferma quando nel modello compaiono variabili non significative (un passo prima)

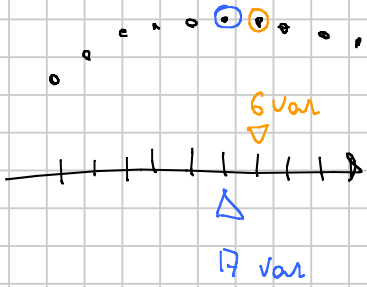
* La stepwise forward è non pratica su Excel, anche se i software statistici dedicati la implementano in modo automatico

* Su Excel si può fare il I passo con REGR. LIN questo aiuta ad orientare la scelta della backward

* Esistono anche la stepwise generale (implementa sia forward sia backward) e metodi di ottimizzazione globale (ci vuole uno "score" calcolabile e poi si fanno $2^p - 1$ regressioni e si sceglie il set di variabili ottimo)

$2^p - 1$ ↑ numero di sottoinsiemi non vuoti

R^2_* AIC



VALIDAZIONE DEL MODELLO

1) Selezione delle variabili
2) Correzione problemi (outliers, nonlinearità, eteroschedasticità) — da chiarire
3) Controllo delle statistiche globali: S_e, R^2, R^2_*, "p_{finale}" — tutto da fare

- Coefficiente di determinazione (corretto o meno)

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

$$SS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad s_e^2 = \frac{SS}{n-p-1}$$

previsto: $\hat{Y}_i = \sum_{j=0}^{p} B_j x_{ij}$

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$ var. campionaria

$\bar{Y}$: previsto migliore in assenza di variabili di ingresso

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

R^2: frazione di devianza spiegata dal modello

$$R^2_{adj} = R^2_* := 1 - \frac{s_e^2}{S_Y^2}$$

R^2_*: frazione di varianza spiegata dal modello

- Come cambiano R^2, S_e, R^2_* durante la selezione delle var.

→ se tolgo una variabile, la variabile x_p

SS' B'_j SS'' B''_j

$$SS' = \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} B'_j x_{ij}\right)^2$$

B'_j minimizzano questa espressione

$$SS'' = \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p-1} B''_j x_{ij} - B''_p x_{ij}\right)^2$$

$B''_p = 0$

siccome è la stessa scrittura: $SS' < SS''$

→ $SS' < SS''$ se tolgo una var. SS cresce sempre
se aggiungo una var. SS diminuisce sempre

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

se tolgo una var. R^2 decresce sempre
se aggiungo una var. R^2 cresce sempre

→ siccome succede anche se la var. in questione è del tutto assurda (tipo il numero di telefono) questo è un difetto di R^2

→ Invece R^2_* si comporta meglio

$$R^2_* = 1 - \frac{S_e^2}{S_Y^2} \qquad S_e^2 = \frac{SS}{n-p-1}$$

S_e^2 cala solo se aggiungo una var. abbastanza utile perché il denominatore in quel caso cala meno del numeratore

→ Cercare il modello che massimizza R^2_* equivale a cercare quello che minimizza S_e.

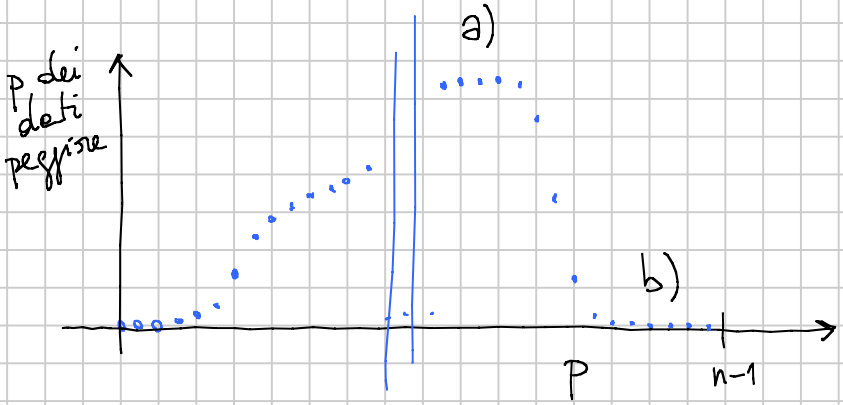
- Valore di p rispetto ad n → OVERFITTING (vedi foglio Excel)

Se nella regressione (ad un certo punto) il valore di p è troppo vicino a quello di n (p+1 coefficienti per fittare n punti) si è in overfitting

→ i test $H_0: \beta_j = 0$ diventano inaffidabili
a) da un lato $(X^TX)^{-1}$ ha coefficienti molto grandi
→ B_j poco precisi // molto variabili
→ test $H_0: \beta_j = 0$ poco potente dice spesso H_0
p dei dati molto alti
→ si tolgono variabili e si esce un po' alla volta

dall'overfitting; purtroppo si tolgono var. "a caso"

b) quando $\frac{p}{n}$ poco meno di 1 il modello è sempre più perfetto per i dati vecchi e i p-dei-dati delle variabili crollano al crescere delle variabili



ora 19

★ Ad esempio: file body.xlsx

id	x_0	x_1	x_2	...	x_{24}	Y (weight)
1	1	~	~		~	~
2	1	~	~			
.	.	~				
.	.					
.	.					
.	.	1			~	~
507	1	~			~	~

n

ci sono 507 punti di $\mathbb{R}^{25}$

$p = 24$

$p+1 = 25$ $\quad n = 507$

$\frac{n}{p} = \frac{507}{24} \approx 21,1$

lontani dall'overfitting

$\frac{n}{p} \gtrsim 5 \qquad \#B_j = p+1$

$\frac{2n}{p^2} \gtrsim 1-2 \qquad \#(X^TX)_{ij} = \frac{(p+1)(p+2)}{2} \approx \frac{p^2}{2}$

★ In caso di possibile overfitting, meglio la stepwise forward.

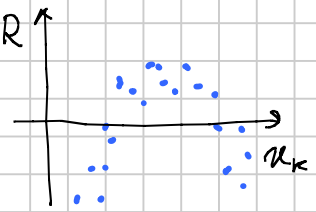
ora 20

CURARE LE NONLINEARITÀ

(approccio standard, senza trasformazioni nonlineari creative)

→ si aggiungono dummy variables che sono termini di grado superiore delle variabili del database

$Y_i = \sum_{j=0}^{p} B_j x_{ij}$ ⟶ analisi dei residui



residui non lineari

aggiungo $x_{p+1}(i) := x_k^2(i)$

$Y_i = \sum_{j=0}^{p} B'_j x_{ij} + B''_{p+1} x_{ik}^2$ se $p=1$ $Y = B_0 + B_1 x + B_1 x^2$

test : $H_0 : \beta_{p+1} = 0$ → H_0 torno indietro di un passo; → H_1 ok, procedo

$H_0 : \beta_k = 0$ → H_0 / → H_1 } comunque, se tengo x_k^2 non posso togliere x_k "regola gerarchica"

analisi dei residui (non guardo i residui di x_k^2)

i residui di x_k mostrano nonlinearità? (di grado ≥ 3)
→ sì : aggiungo x_k^3
→ no : guardo le altre var

★ E' bene controllare anche il grafico dei residui vs i previsti (anche se va curata indirettamente)

★ Conviene aggiungere dummy variables una alla volta

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ . . . . . $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1^3 + \beta_5 x_2^2$

★ Fenomeno : se su una singola variabile ho fatto crescere troppo il grado, spesso i coeff di $x_k, x_k^2, \ldots, x_k^d$ risultano avere p dei dati né alti né bassi e tutti simili.

- Ci sono anche i termini di interazione:

  $x_1 x_2$, $x_1^2 x_2$, $x_1 x_2 x_3^2$, ...

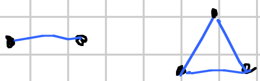
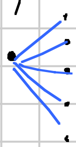
  ↑ i più interessanti di solito

→ interpretazione dei coefficienti dei termini di interazione
  vedi DoE

→ non ci sono sintomi della necessità / utilità dei term. di int.
  i residui non aiutano

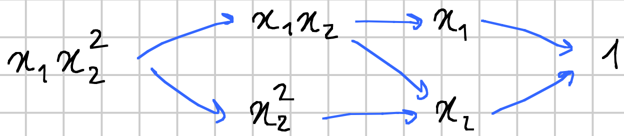
  ↳ sono $\binom{p}{2} = \frac{p(p-1)}{2} \approx \frac{p^2}{2}$ $\qquad p = 24 \quad \binom{24}{2} = 276$

  troppi per provarli tutti

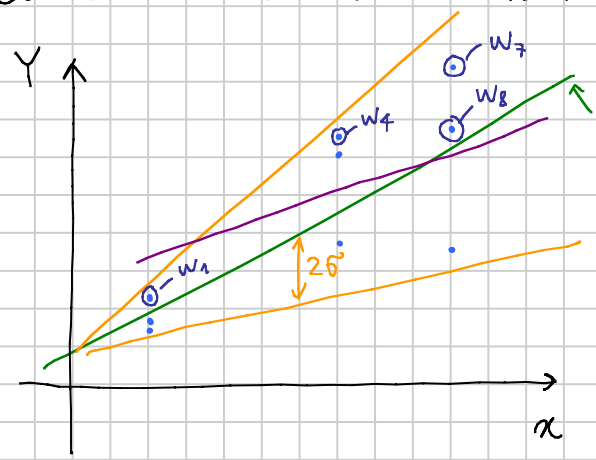
  ↳ se $n/p \gg 1$ e ci sono 2-3 variabili molto più importanti delle altre,  almeno quelle poche vanno provate 

☆ REGOLA GERARCHICA GENERALE

se tenete una variabile, dovete tenere tutte quelle che la dividono



CURARE LA ETEROSCHEDASTICITÀ



$\sigma = \sigma(x)$

in questo caso $\sigma = \lambda x$

Per curare questo fenomeno bisogna inventare una regressione che pesi i punti in maniera diversa

→ associamo un peso ad ogni punto

$\sigma^2 = \sigma^2(i)$ → $\sigma^2(x_{ik})$ dipendente da una variabile

→ $\sigma^2(\hat{Y}_i)$ " dal previsto

→ $\sigma^2(i)$ nota / stimabile

→ $\boxed{w(i) = \frac{1}{\sigma^2(i)}}$ andrebbe bene anche $w(i) = \frac{100}{\sigma^2(i)}$

↳ non importano i fattori di scala: se ad esempio raddoppio tutti i pesi, la regressione non cambia

→ ad esempio: $\sigma^2(i) \propto x_{ik}$ ↑ proporzionale $\sigma^2(i) = \lambda x_{ik}$ $\quad w(i) \propto \frac{1}{\sigma^2(i)}$ $\quad \boxed{w(i) = \frac{1}{x_{ik}}}$

$\sigma(i) \propto \hat{Y}_i \quad w(i) \propto \frac{1}{\sigma(i)^2} \quad \boxed{w(i) = \frac{1}{\hat{Y}_i^2}}$

$\sigma(i) \propto x_{ik} \quad \boxed{w(i) = \frac{1}{x_{ik}^2}}$ ...

se ho $x_1 \dots x_7$ uguali $\sigma^2(i) \approx \text{VarCamp}(Y_1, \dots, Y_7) \quad i = 1, \dots, 7$

se ho $Y_i \sim \text{bin}(n_i, p_i)$ e $p_i = \beta_0 + \beta_1 x_{i1} + e_i$

$\text{Var}(Y_i) = \sigma^2(i) = n_i p_i (1-p_i) \approx n_i \hat{p}_i (1-\hat{p}_i)$

REGRESSIONE PESATA

Occorre pesare i residui

$$R_i = Y_i - \sum_{j=0}^{p} B_j x_{ij} \approx Y_i - \sum_{j=0}^{p} \beta_j x_{ij} = Y_i - E(Y_i) \sim \mathcal{N}(0, \sigma_i^2)$$

$$SS = \sum_{i=1}^{n} R_i^2 \qquad SS_w = SS_w(B) := \sum_{i=1}^{n} R_i^2 w_i$$

(w_i: pesi)

$$R_i\sqrt{w_i} \approx \sqrt{w_i}\left(Y_i - E(Y_i)\right) \sim \mathcal{N}\left(0, \underbrace{\sigma_i^2 w_i}_{1 \text{ o costante}}\right) \qquad w_i = \frac{1}{\sigma_i^2} \qquad w_i \propto \frac{1}{\sigma_i^2}$$

I coefficienti della regressione pesata sono B^w, quelli che minimizzano SS_w

Non c'è l'equivalente di Se (HW: come stimo σ_i^2 ?)

★ Tecnica pratica per affrontare la regressione pesata:

x_{ij} $\quad i = 1, \dots, n \quad j = 0, \dots, p \quad x_{i0} \equiv 1 \ \forall i$

$$Y_i \sim \mathcal{N}\left(\sum_{j=0}^{p} \beta_j x_{ij}, \sigma_i^2\right)$$

$$w_i \propto \frac{1}{\sigma_i^2} \qquad w_i = \frac{\tau^2}{\sigma_i^2} \iff w_i \sigma_i^2 = \tau^2$$

Costruisco un nuovo database:

$$x'_{ij} := x_{ij}\sqrt{w_i} \qquad x'_{i0} = \sqrt{w_i} \ \forall i \qquad \text{non c'è dummy var} \equiv 1$$

$$Y'_i := Y_i\sqrt{w_i} \sim \mathcal{N}\left(\sum_{j=0}^{p} \beta_j \underbrace{x_{ij}\sqrt{w_i}}_{x'_{ij}}, \underbrace{w_i \sigma_i^2}_{\tau^2}\right) \sim \mathcal{N}\left(\sum_{j=0}^{p} \beta_j x'_{ij}, \tau^2\right)$$

$$\boxed{Y' = \sum_{j=0}^{p} \beta_j x'_{ij} + e'_i}$$

è un modello di regressione omoschedastica

Tutta l'analisi standard omoschedastica può essere condotta su questo modello, salvo poi ritradurre all'indietro i risultati nel modello di partenza (tipicamente dividendo per $\sqrt{w_i}$ opportuna)

GESTIONE VARIABILI CATEGORICHE

→ var numeriche : altezza, peso, velocità, ...

→ var categoriche :

a) nominali (tipo di valvola : KC1, HV, rossa)

b) dicotomiche (due soli valori)

c) ordinali (titolo di studio, giudizio scolastico, fascia d'età)

★ Come <u>variabili di risposta</u> sono quasi inaccettabili

↳ solo le dicotomiche vengono a volte accettate (analisi discriminante) (oppure regressione logistica)

★ Come variabili di ingresso :

b) dicotomiche = numeriche   non c'è problema

↳ qualunque codifica numerica è equivalente

a) nominali → vanno "esplose in dicotomiche"

c) ordinali → = nominali   dimentico l'ordine

↳ codificare le categorie in numeri

tante scelte possibili → decisione delicata

★ Codifica ordinali

→ titolo di studio

titolo di studio	codifica		diverse
	0	−1	0
V elem	1	1	5
III media	2	3	8
diploma professionale	3	5	13
maturità	4	7	13
laurea triennale	5	9	16
laurea mag + spec + 4 + 5	6	11	18
dottorato	7	13	21

(prime due codifiche: ≡, trasf lineari; differenze codifica diversa: 5, 3, 5, 0, 3, 2, 3)

→ giudizio

giudizio				
completamente negativo	0	−1	−2	0
abbastanza negativo	1	0	−1	2
abbastanza positivo	2	1	1	5
completamente positivo	3	2	2	6

(1ª ≡ 2ª ≠ 3ª ≠ 4ª)

- Esplodere in dicotomiche

k categorie ⟶ k−1 dicotomiche

	pri	est	aut	
primavera	1	0	0	codifica
estate	0	1	0	univoca
autunno	0	0	1	
inverno	0	0	0	

default

★ Sale (molto) p e peggiora (a volte troppo) il rapporto n/p

★ Come si interpretano i coefficienti B corrispondenti

→ Dicotomica: sesso (maschi = 0, femmine = 1) del conducente

$$Y = B_0 + B_1 x_1 + \dots + B_k x_k + \dots + e$$

Y: consumi

$B_k = -0,7$

$Y(\text{maschi}) = 16,4$

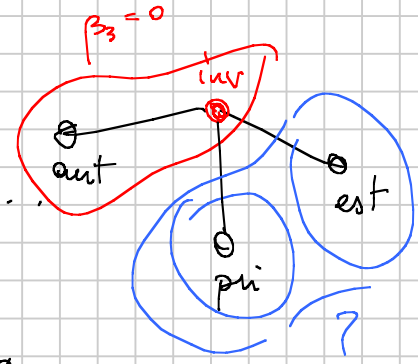
$Y(\text{femmine}) = 15,7$

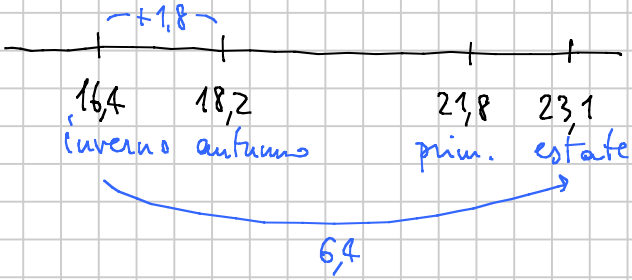
→ Quattro stagioni, tre coefficienti

$$Y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + \dots$$

Y: consumi; x_1: pri; x_2: est; x_3: aut

$B_1 = 5,4 \quad B_2 = 6,7 \quad B_3 = 1,8$





$H_0: \beta_3 = 0$ vuol dire che inverno ≈ autunno

$H_0: \beta_1 = \beta_2$ se ≈ pri ≈ est

★ Se la stepwise backward elimina una di queste dicotomiche vuol dire che non la distingue da quella di default.
L'effetto è automaticamente che le due categorie si fondono
→ se autunno non è "significativa", ovvero il test dice $H_0: \beta_3 = 0$

	PRI	EST
primavera	1	0
estate	0	1
autunno	0	0
inverno	0	0

★ Se si vole tentare di fondere due categorie che non sono di default occorre testare $H_0: \beta_i = \beta_j$

ora 22

★ Se le var di ingresso sono:
a) una sola categorica → ANOVA a 1 via
b) due sole categoriche → ANOVA a 2 vie
la tecnica ottimale non è la regressione, ma la analisi della varianza

● Trasformazioni lineari
Non hanno (quasi) alcun effetto

$$Y_i = \sum \beta_j x_{ij}$$

$$Y_i' = m + q Y_i \qquad x_{ij}' = m_j + q_j x_{ij} \qquad j = 1, \dots, p$$

$$B_j' = \frac{1}{q_j} B_j \qquad B_0' \text{ cambia in modo caotico}$$

$$S_e' = q S_e$$

non cambiano:
- $\alpha^*_{\beta_j}$ p dei dati
- R^2, R^2_*
- grafici dei residui

DESIGN OF EXPERIMENT

(Sleeper capitolo 10)

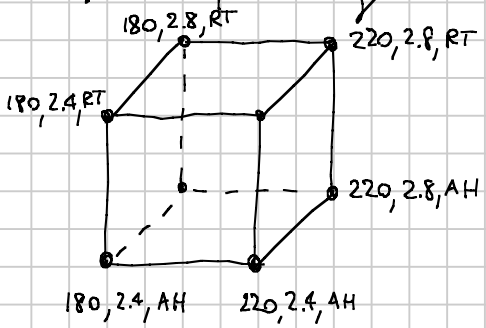
Montgomery

- Pianifico gli esperimenti → imposto i valori delle x_i
- Tutte le x_i sono forzate ad essere dicotomiche
  - ↳ numeriche di cui scelgo due valori (T = 180 , 220)
  - ↳ categoriche con due scelte (due valvole)
- Fissati i valori di ingresso si fanno alcune prove, mai una sola
- La versione "base" del DoE prevederebbe di provare tutte le combinazioni delle var di ingresso

x_1 : temperatura (180 ; 220)

x_2 : pressione (2,4 ; 2,8) ⇝ Y : umidità residua

x_3 : tipo di ugello (AH o RT)



3 var → $2^3 = 8$

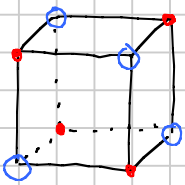
8 RUN ovvero configurazioni diverse delle var di ingresso

potrei fare 5 REPLICHE per ogni run → 40 esperimenti

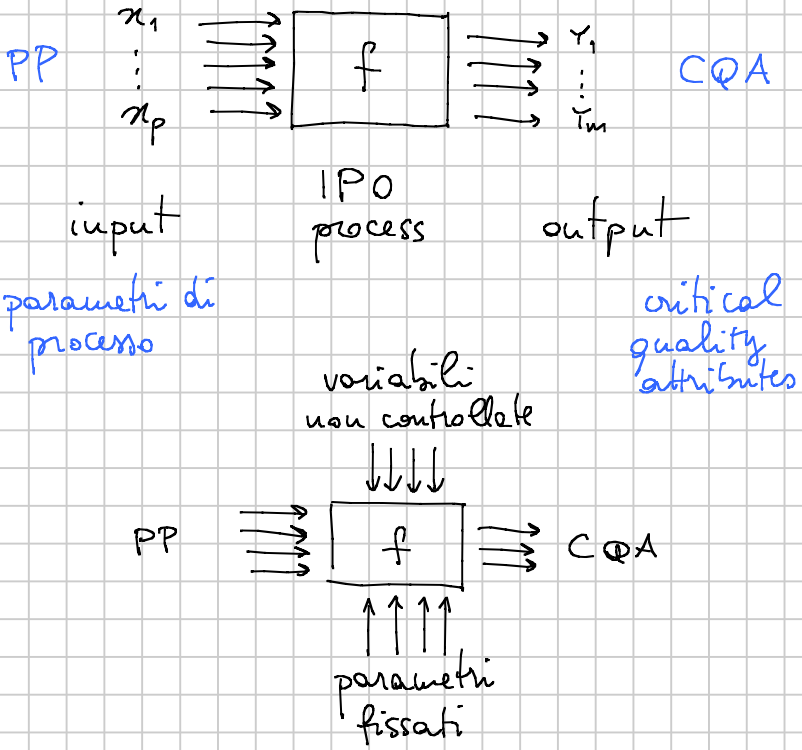
→ 40 valori misurati per Y

✱ run = 2P cresce troppo in fretta ... quando mai in pratica potrò fare più di 32 o 64 run

→ per fortuna, oltre ai design fattoriali completi esistono anche design fattoriali frazionari



Scelta delle variabili



* Occorre studiare accuratamente il processo e identificare tutte le variabili che potranno avere un effetto. Quelle dimenticate finiscono tra le non controllate e come tali diventano fonte di rumore
* E' preferibile tenere basso il numero delle variabili di output e di input, sempre tenendo conto delle finalità dell'analisi
  ↳ le var di ingresso dovrebbero essere l'unione di quelle che si pensano impattare su Y_1, su Y_2, ... fino a Y_m.

---

ora 23

* Quando ci sono tante var di ingresso è possibile impostare un esperimento preliminare che con poche prove sia in grado di screenare drasticamente l'insieme delle variabili
  = esperimento di SCREENING

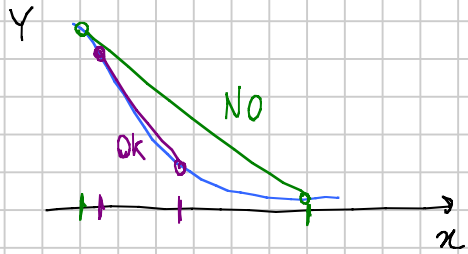
  Di contro l'esperimento che serve a capire bene l'effetto delle PP sulle CQA deve avere poche variabili ≤ 7
  = esperimento di MODELING

- Come scegliere i due valori delle var di ingresso numeriche

Criteri

a) linearità

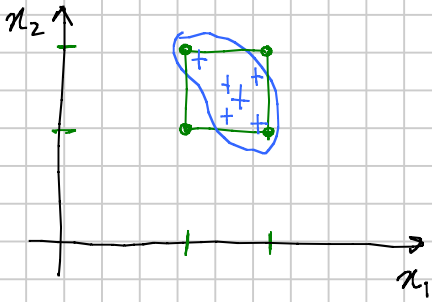
Se la risposta delle Y è non lineare, solo tenendo vicini i due valori il modello lineare sarà adeguato

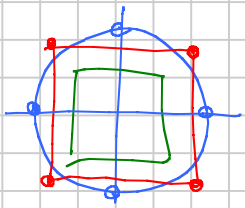


b) rapporto segnale/rumore

I due valori vanno presi abbastanza distanti che l'effetto sulle Y sia misurabile

c) funzionamento ai bordi

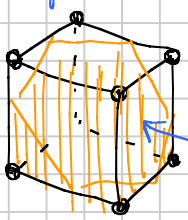




Bisogna essere sicuri che il processo funzioni e permetta di misurare valori ragionevoli di Y in tutte le combinazioni, che spero possono essere veramente estreme

d) utilità del design space

= regione di spazio delimitata dai punti delle run



$$Y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + e$$

dentro è valido, fuori chissà?

Ogni ottimizzazione o scelta dei PP una volta noto il modello dovrà essere all'interno del design space: è importante che non sia troppo costrittivo

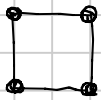
- La var di risposta non può essere dicotomica, e in generale dovrebbe essere numerica, informativa, precisa e con errore gaussiano

Scelta del design

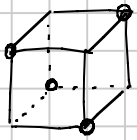
- Cos'è un design? Esempi

  Serie di Taguchi: L4, L8, L16, L32 ...

  L4 : 2 o 3 fattori = variabili di ingresso ; 4 run



full factorial
a 2 fattori



fattoriale $\frac{1}{2}$-frazionario
a 3 fattori

run	A	B
1	−1	−1
2	−1	+1
3	+1	−1
4	+1	+1

run	A	B	C
1	−1	−1	+1
2	−1	+1	−1
3	+1	−1	−1
4	+1	+1	+1

La matrice X della regressione si ottiene semplicemente aggiungendo la "dummy" colonna di 1

1	−1	−1	1
1	−1	1	−1
1	1	−1	−1
1	1	1	1
v_0	v_A	v_B	v_C

è un design ORTOGONALE nel senso che i vettori colonna sono a due a due ortogonali

$$v_0 \cdot v_A = 0 = v_0 \cdot v_B = v_0 \cdot v_C = v_A \cdot v_B = v_B \cdot v_C = v_A \cdot v_C$$

★ l'ortogonalità è fondamentale, tutti i design che troviamo sui libri sono ortogonali

↳ se non tutti gli esperimenti funzionano e restano buchi nella tabella (righe mancanti) il design perde la ortogonalità

MMSD

ora 24

Note Title 29/11/2011

DOE

- Design = matrice che prescrive quali RUN si eseguono

L8 di Taguchi numero di fattori : almeno 3

I	A	B	C	A^2	AB	BC	AC	ABC	Y
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	-1	-1	-1	1	1	1	1	-1	~

usando un design di questo tipo si riescono a indagare anche le interazioni

$Y = b_0 + b_1 A + b_2 B + b_3 C + b_4 AB + \dots$

↑ diversa da tutte e ortogonale con tutte

se il design è a 2 livelli non c'è modo di vedere le nonlinearità tipo A^2

- 8 colonne ortogonali hanno uno SPAN che è tutto $\mathbb{R}^8$, quindi qualunque altra colonna non potrebbe essere né ortogonale né linearmente indipendente con le precedenti

ABCD+	BCD+	ACD+	ABD+	ABC+	CD+	AD+	BD+	
I	A	B	C	D	AB	BC	AC	AD
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	-1	-1	-1	-1	1	1	1	1

uguali

ABD = ~~AB~~ ~~ABC~~ ABCD = ~~ABC~~ ~~ABC~~

I + (ABCD) → termine noto
A + (BCD)
B + (ACD)
C + (ABD)
D + (ABC)
AB + CD
AC + BD
AD + BC

effetti dei fattori principali

ALIAS structure del design

★ Se un design ha 8 run, può ammettere al massimo 8 colonne ovvero 8 variabili (comprese le "dummy") e 8 coefficienti stimati

★ Se non è fattoriale completo in totale le variabili con tutte le interazioni sono di più, quindi per ogni run / colonna / coefficiente stimato ci saranno 2, 4, 8, ... effetti in alias

★ Gli effetti delle interazioni a 3 o più fattori nel mondo industriale si trascurano e questo migliora l'utilità del design, perché la struttura di alias è più leggera

Fractional Factorial Design

L16 a 6 fattori

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

→ minimo = 4

I + ~~ABCE + ADEF + BCDF~~	0 + 4 = 4
A + ~~BCE + DEF + ABCDF~~	1 + 3 = 4
B + ACE + CDF + ABDEF	.
C + ABE + BDF + ACDEF	.
D + AEF + BCF + ABCDE	.
E + ABC + ADF + BCDEF	.
F + ADE + BCD + ABCEF	4
AB + CE + ~~ACDF + BDEF~~	2+2 = 4
AC + BE + ~~ABDF + CDEF~~	.
AD + EF + ~~ABCF + BCDE~~	.
AE + BC + DF + ~~ABCDEF~~	4
AF + DE + ~~ABCD + BCEF~~	.
BD + CF + ~~ABEF + ACDE~~	.
BF + CD + ~~ABDE + ACEF~~	4
~~ABD + ACF + BEF + CDE~~	3+3 = 6
~~ABF + ACD + BDE + CEF~~	6

output di Minitab

★ La RISOLUZIONE è il minimo sulle righe della somma dei gradi dei due termini di grado minore

(Anche L8 a 4 fattori era di IV)

L4 a 3 fattori:

A	B	C
1	1	1
1	-1	-1
-1	1	-1
-1	-1	1

$I + (ABC)$ — 0+3 = 3
$A + BC$ — 1+2
$B + AC$ — 1+2
$C + AB$ — 1+2 3

III risoluzione

→ III ris. gli effetti principali sono in alias con le int. a due

⇒ non posso essere sicuro degli effetti principali

⇒ non posso in nessun caso stimare le interazioni

★ Vanno bene per esperimenti di screening: ad esempio bastano 32 run per screenare 20 fattori. Una volta scesi a numeri più ragionevoli, si sceglierà un design di risoluzione maggiore per fare un esperimento di modeling.

ora 25

★ Per gli esperimenti di screening si usano anche i design di Plackett - Burman che presentano il confounding al posto dell'aliasing

$A + CF \longrightarrow A + \frac{1}{3}CF + \frac{1}{3}BD + \frac{1}{3}EG$

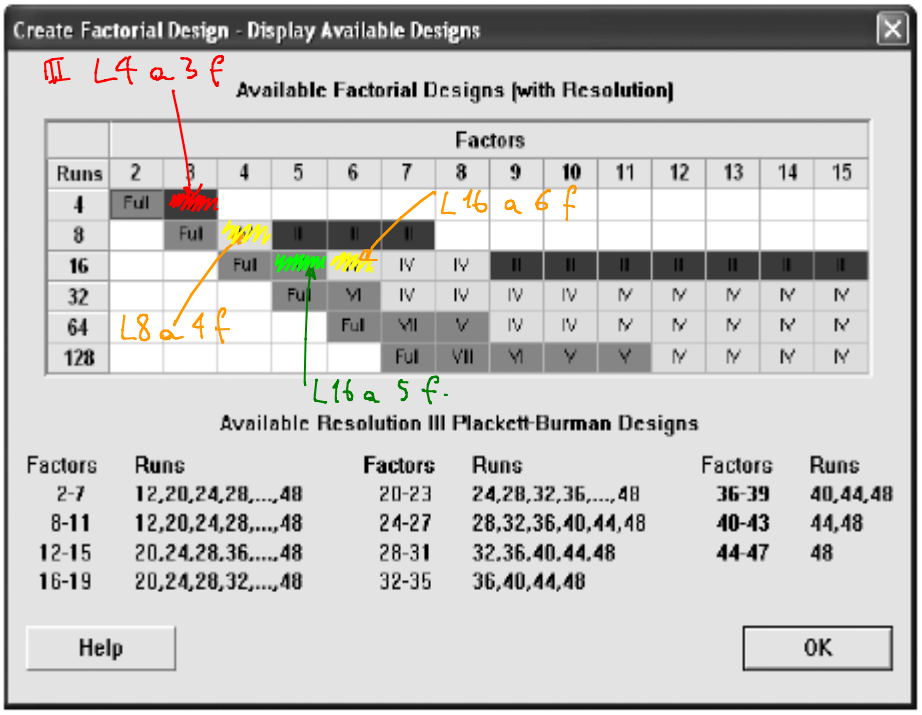
→ V ris. effetti principali e interazioni a due posso sempre essere stimati efficacemente, perché in alias solo con termini di grado ≥ 3

⇒ perfetti per esperimenti di modeling

Alias structure: Taguchi L16 with 5 factors

I+~~ABCDE~~
A+~~BCDE~~
B+~~ACDE~~
C+~~ABDE~~
D+~~ABCE~~
E+~~ABCD~~
AB+~~CDE~~
AC+~~BDE~~
AD+~~BCE~~
AE+~~BCD~~
BC+~~ADE~~
BD+~~ACE~~
BE+~~ACD~~
CD+~~ABE~~
CE+~~ABD~~
DE+~~ABC~~

→ IV ris. effetti principali stimabili, interazioni a due talvolta in alias tra di loro



Runs	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	Full													
8		Full												
16			Full											
32														
64														
128														

Factors	Runs	Factors	Runs	Factors	Runs
2-7	12,20,24,28,...,48	20-23	24,28,32,36,...,48	36-39	40,44,48
8-11	12,20,24,28,...,48	24-27	28,32,36,40,44,48	40-43	44,48
12-15	20,24,28,36,...,48	28-31	32,36,40,44,48	44-47	48
16-19	20,24,28,32,...,48	32-35	36,40,44,48		

◎ EFFETTI DI INTERESSE ED EFFICIENZA

★ EO1 "universali" sono gli effetti principali e tutte le interazioni a due

mtst 10 01. 4 $(X1, \dots, X7) = (A, B, \dots, G)$

tra tutte le interazioni a due posso escludere quelle $\{A,B,C\} \times \{D,E,F\}$

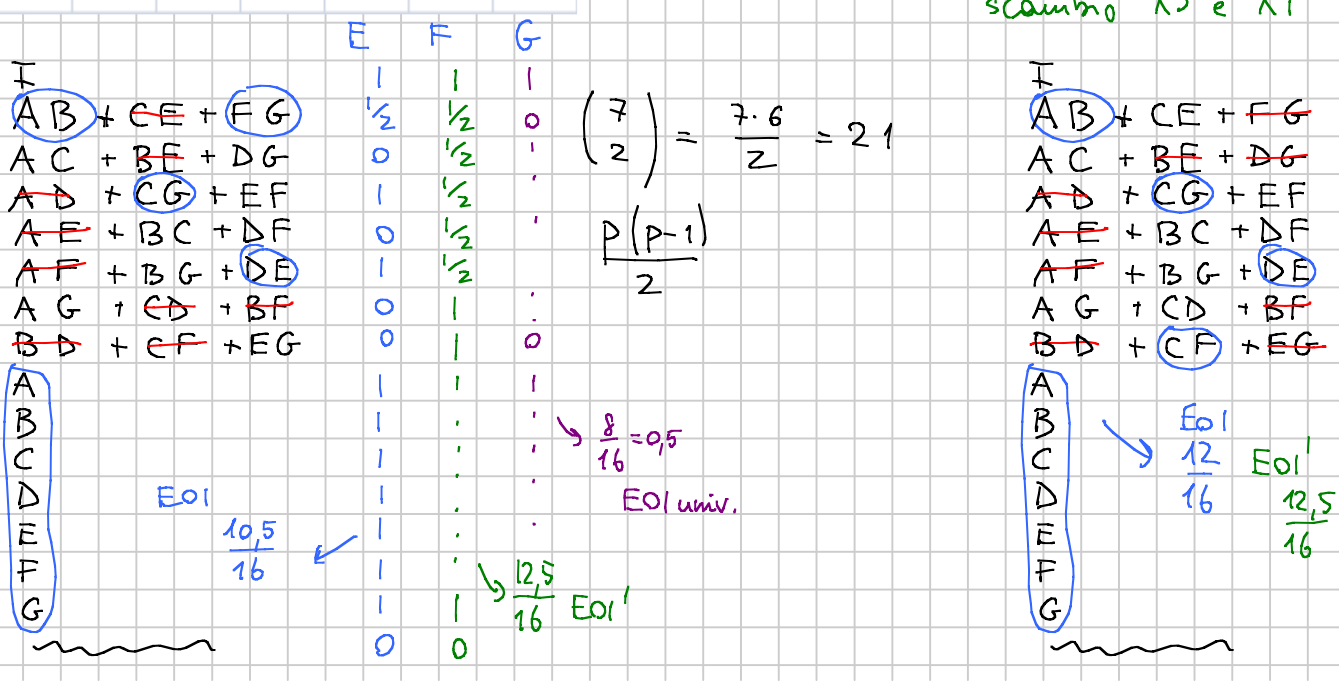
mentre GC e GF sono importanti, come AB e DE

EO1: $\{A, B, C, D, E, F, G, CG, FG, AB, DE\}$

EO1': $\{A, B, C, D, E, F, G, CG, FG, AB, DE, AC, DF, BC, EF, G \text{ con tutto}\}$

A	B	C	D	ABC	BCD	ACD

scambio X3 e X7



$$\binom{7}{2} = \frac{7 \cdot 6}{2} = 21$$

$$\frac{p(p-1)}{2}$$

★ EFFICIENZA :

→ L'efficienza di una riga è :

- 1 se c'è un sob EOI
- 1/2 se ci sono due EOI
- 0 se ci sono 0, 3, 4, ... EOI

→ L'efficienza di un design è la media di quella delle righe

★ Tabella delle efficienze per EOI universali

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

ora 26

● REPLICHE

→ Più numerose sono, meglio viene l'analisi (che è fatta sulle loro medie)

→ Meglio che siano comunque almeno 2 (per individuare outlier o problemi)

→ Quante ?

Il numero può essere determinato se si conoscono :

1) σ la dev. standard del processo
2) UQL + potenza relativa
   = la grandezza degli effetti che voglio poter rilevare e con che probabilità

★ Esempio : essiccatore Y = umidità residua polvere ≈ 2%

$$Y = b_0 + b_1 A + b_2 B + b_3 C$$

codificate ± 1

se $b_1 = 1\%$ l'effetto di A su Y è 2% → molto grande

se $b_2 = 0{,}02\%$ " " B " " è 0,04% → non importante

⇝ 0,3% → effetto 0,6% → deve essere rilevabile ... 80%

VQL potenza = 80%

→ Quante, nel serio?

"rule of thumb" add 32

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

↑ arrotondo per eccesso

Sembra valida nel mondo industriale

L4 9 repliche × run = 36 esperimenti

L8 5 repliche × run = 40 esperimenti

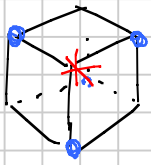
L16 3 repliche × run = 48 esperimenti

L32 2 repliche × run = 64 esperimenti

...2 da qui in poi...

} per la serie di Taguchi

- Center point

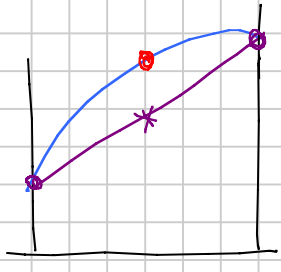


I A B C ...

center point: 1 0 0 0 ...

Non si può fare se ci sono variabili categoriche

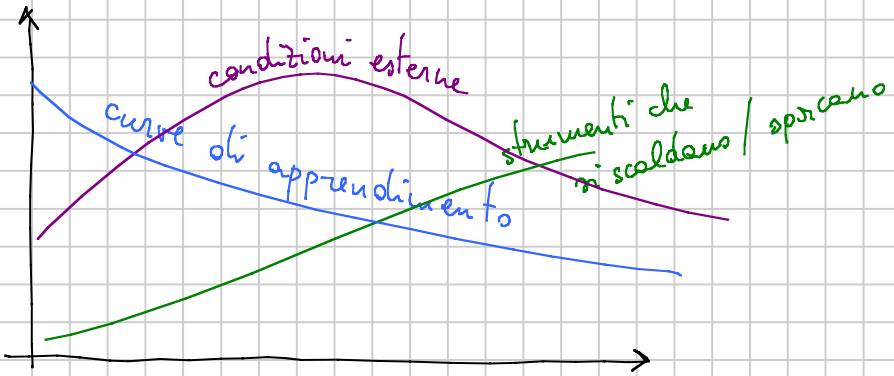
Si fa alla fine una prova nel center point, replicata tante volte quanto le altre



Si confronta con il previsto per lo stesso punto dal modello, o verificando se cade nell'interv. di predizione, o anche a occhio

* Se non corrisponde vuol dire che sono in gioco nonlinearità

- Randomizzare le repliche



A	+1	+1	+1	+1	-1	-1	-1	-1	tempo
B	+1	+1	-1	-1	+1	+1	-1	-1	
C	+1	-1	+1	-1	+1	-1	+1	-1	

in questo caso un effetto legato al tempo può essere rilevato come associato ad A (o anche a B, pur in peso minore)

★ Randomizzare le repliche trasforma questo problema in un aumento di σ^2.

FACCIO L'ESPERIMENTO

- Regressione quasi standard

→ $X^TX = r \cdot I$ r: # repliche $I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

$(X^TX)^{-1} = r^{-1} \cdot I$

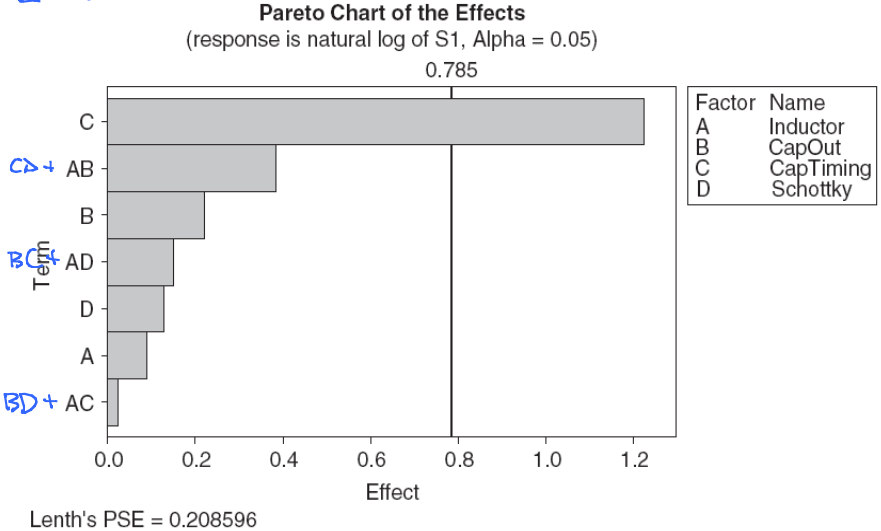
→ tutte le covarianze dei coefficienti B_i, B_j sono nulle

→ B_i non cambia se aggiungo o tolgo variabili

↳ non ha senso fare la selezione delle variabili classica

$$Y = 3{,}4 + 1{,}2A - 0{,}7B + \cancel{0{,}04C} - 0{,}9AB + \cancel{0{,}7BC} + \cancel{0{,}2AC} - \cancel{0{,}01ABC}$$

L8 a 4 fattori



MMSD

ora 27

Note Title 06/12/2011

Selezione variabili nel DoE

→ è qualitativa : (p dei dati non del tutto affidabile se $\frac{n}{p}$ basso)

→ si usa il grafico di Pareto degli effetti

↳ quali effetti sono rilevanti dal punto di vista ingegneristico

↳ quali effetti hanno senso dal punto di vista ingegneristico

↳ regola gerarchica $AB \rightarrow A$ e B

→ si possono togliere tante variabili assieme (design ortogonale)

- Analisi dei residui va fatta sul modello ridotto e a livello di singole repliche

- Outliers : sarebbe giusto rifare quegli esperimenti, in modo che le repliche per ogni run siano le stesse

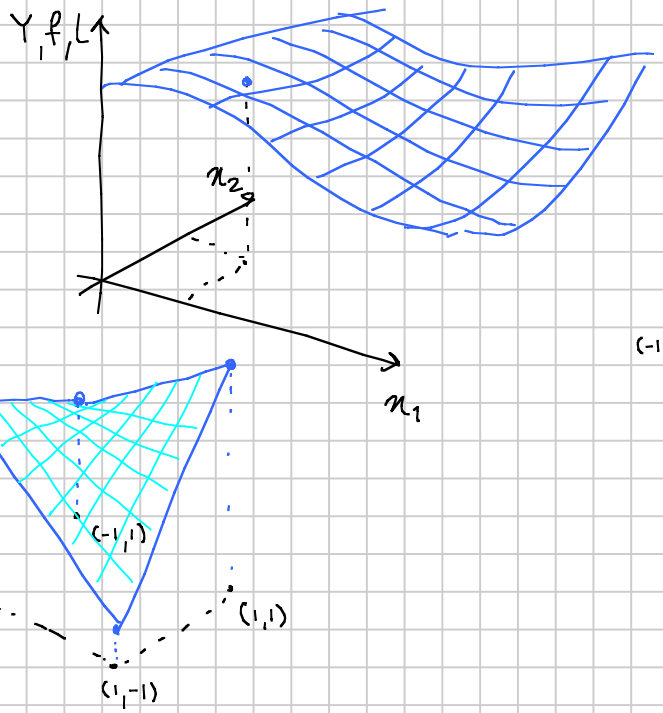
TERMINI DI INTERAZIONE (DoE E REGRESSIONE)

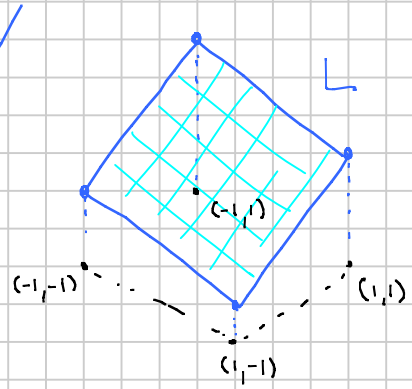
$$Y = f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$$L(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

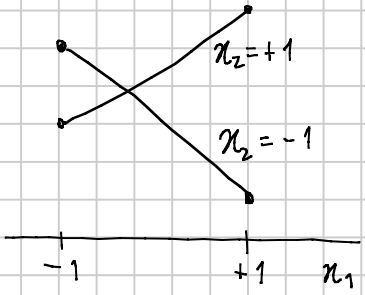
lineare da $\mathbb{R}^2 \rightarrow \mathbb{R}$

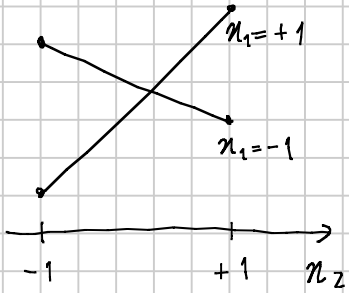
(il grafico è un piano)

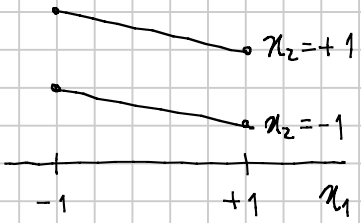


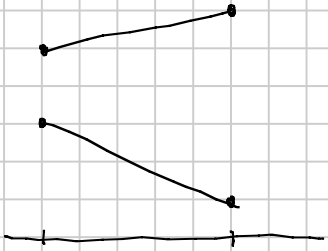


★ Gli effetti in presenza di interazioni

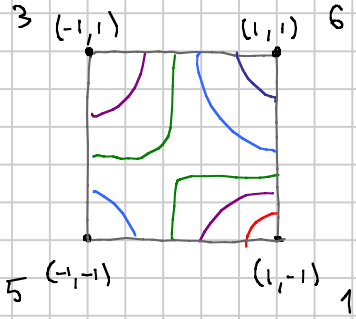








★ Curve di livello, contour plot



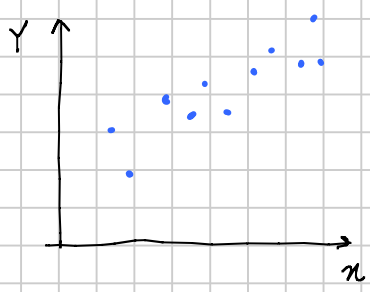
ANALISI DELLA VARIANZA (Cap. 10 Ross)

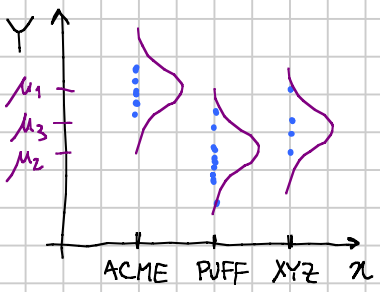
var numeriche → [?] → $Y \sim \mathcal{N}(?; \sigma^2)$ REGRESSIONE

1 var categorica → [?] → $Y \sim \mathcal{N}(?; \sigma^2)$ ANOVA A 1 VIA

2 var categoriche → [?] → $Y \sim \mathcal{N}(?; \sigma^2)$ ANOVA A 2 VIE

ANOVA A UNA VIA





- I dati sono un certo numero di campioni

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

m: # categorie per ciascuna ho un campione

categ. 1: $Y_{1,1}, Y_{1,2}, \dots, Y_{1,n_1} \sim \mathcal{N}(\mu_1; \sigma^2)$ iid

categ. 2: $Y_{2,1}, Y_{2,2}, \dots, Y_{2,n_2} \sim \mathcal{N}(\mu_2; \sigma^2)$ iid

. . . . . .

categ. m: $Y_{m,1}, Y_{m,2}, \dots, Y_{m,n_m} \sim \mathcal{N}(\mu_m; \sigma^2)$ iid

- Test principale:

$H_0: \mu_1 = \mu_2 = \dots = \mu_m$ — Y non dipende da x

H_1: non tutte uguali — Y dipende da X

- Stimatore di σ^2 ottimo: varianza "within" (= varianza "pooled")

$S_w^2 = S_p^2$ = media pesata di $S_1^2, S_2^2, \dots, S_m^2$ con pesi proporzionali ai gdl

$$Y_{i,*} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}$$ media campionaria categoria i

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(Y_{i,j} - Y_{i,*}\right)^2$$ varianza campionaria categoria i

$$S_W^2 = \sum_{i=1}^{m}\frac{n_i - 1}{N-m}S_i^2$$ varianza within

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

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

$$S_W^2 \sim \text{Gamma}\left(\frac{N-m}{2}; \frac{2\sigma^2}{N-m}\right)$$

★ Queste si usano per fare inferenza su σ^2.

★ Coincide con S_p^2 perché questa situazione è proprio la generalizzazione di quelle in cui si introduceva lo stimatore pooled

★ Qualunque media pesata delle S_i^2 rappresenta uno stimatore corretto di σ^2. I coefficienti $\frac{n_i - 1}{N - m}$ garantiscono che sia quello più preciso (= di varianza minima).

- Varianza "between"

$Y_{1,*}, Y_{2,*}, \dots, Y_{m,*} \longrightarrow$ NO $\quad \frac{1}{m-1} \sum_{i=1}^{m} \left(Y_{i,*} - \frac{1}{m} \sum_j Y_{j,*} \right)^2$

$$Y_{*,*} = \frac{1}{N} \sum_{i,j} Y_{i,j} = \sum_{i=1}^{m} \frac{n_i}{N} Y_{i,*}$$

$$S_B^2 = \sum_{i=1}^{m} \frac{n_i}{m-1} \left(Y_{i,*} - Y_{*,*} \right)^2$$

Essendo una stima di var. campionaria delle medie, dovrebbe essere grande in H_1 e piccola in H_0

★ Quando è vera H_0, S_B^2 è uno stimatore di σ^2 indipendente da S_W^2

solo se $\mu_1 = \mu_2 = \dots = \mu_m$ $\quad \frac{S_B^2}{\sigma^2} (m-1) \sim \chi^2(m-1)$

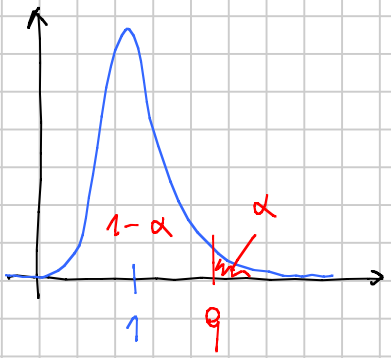
Altrimenti è tendenzialmente maggiore

- Come si fa il test:

F di Fisher, vedi ora 12

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

Siccome in H_1 $S_B^2 \geq S_W^2$ posso fare il test unilaterale



$q = \text{INV.F}(\alpha\ ;\ m-1\ ;\ N-m)$

$RA_{stat} : [0\ ;\ q]$

$\alpha^* = \text{DISTRIB.F}(stat\ ;\ m-1\ ;\ N-m)$

✱ Se $m=2$ il test basato su $\dfrac{\bar{X}-\bar{Y}}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$ è uguale a questo (i p dei dati coincidono)

◉ Identità utili

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

✱ Identità delle devianze:

$$\boxed{SS_Y = SS_W + SS_B}$$

$$SS_* = S_*^2 \cdot gdl_*$$

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

dove $S_Y^2 = \dfrac{1}{N-1}\sum_{i,j}\left(Y_{i,j}-Y_{*,*}\right)^2$

◉ Altra inferenza

$$\mu_i \approx Y_{i,*} \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right)$$

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

$$\mu_i-\mu_j \approx Y_{i,*}-Y_{j,*} \sim \mathcal{N}\left(\mu_i-\mu_j, \sigma^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\right)$$

$$\boxed{\frac{Y_{i,*}-Y_{j,*}-(\mu_i-\mu_j)}{S_W\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}} \sim t(N-m)}$$

◉ Opportuno fare l'analisi dei residui per validare il modello (outliers ; omoschedasticità ; residui normali)

★ A volte la eteroschedasticità può essere dovuta a:

a) dati lognormali ⟶ a volte basta farne il log

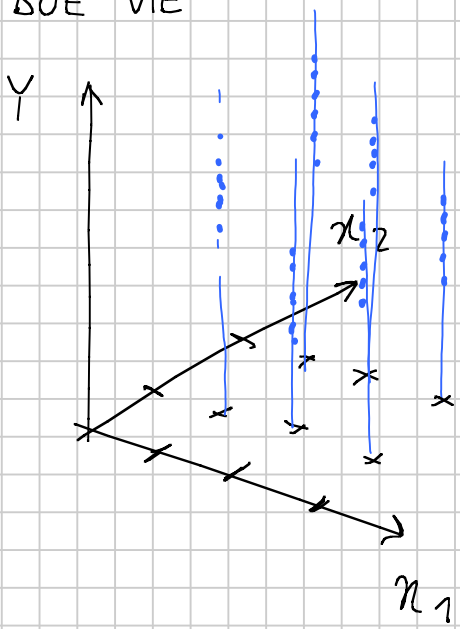
b) dati di Poisson $Pois(\mu)$ media μ e var μ

⟶ a volte basta farne la $\sqrt{\ }$

$X \sim Pois(\mu) \Rightarrow \sqrt{X}$ ha media che dipende da μ e varianza $\approx \frac{1}{4}$

c) in generale: provare trasformazioni nonlineari

ANOVA A DUE VIE



- Struttura dei dati

x_1 ha m categorie $\{1, 2, \dots, m\}$ (3 nell'esempio)

x_2 ha n categorie $\{1, 2, \dots, n\}$ (2 nell'esempio)

★ Rigidità dell'ANOVA a 2 VIE: ogni campione deve avere la stessa numerosità l.

↳ Se le numerosità l_{ij} non sono tutte uguali, si può a volte decidere di buttare via un po' di dati dalle caselle più numerose, riconducendosi alla numerosità minima.

$Y_{i,j,k}$ dato k-esimo del campione corrispondente a $x_1 = i$, $x_2 = j$

$Y_{1,1,1}, \dots Y_{1,1,\ell} \sim \mathcal{N}(\mu_{1,1}; \sigma^2)$ iid

$Y_{1,2,1}, \dots Y_{1,2,\ell} \sim \mathcal{N}(\mu_{1,2}; \sigma^2)$ iid

$\dots\dots\dots$

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

tutte indipendenti

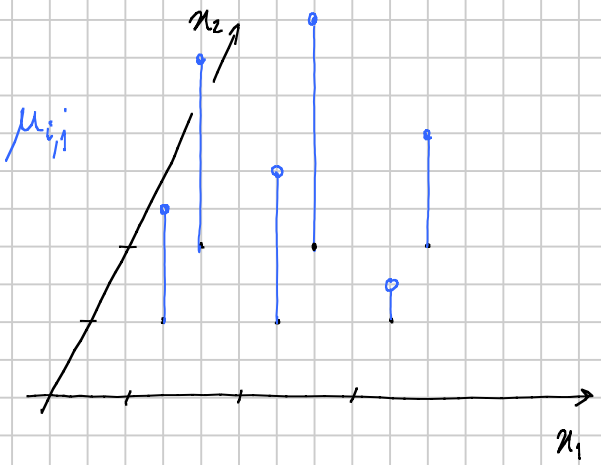
★ Se $\ell > 1$ la tecnica ANOVA a 2 VIE con repliche e posso ottenere informazioni più dettagliate

Tuttavia posso sempre ricondurmi al caso $\ell = 1$ facendo la media su ogni "casella"

$$Y_{i,j,1}, Y_{i,j,2}, \dots, Y_{i,j,\ell} \longrightarrow Y_{i,j,*} := \frac{1}{\ell} \sum_{k=1}^{\ell} Y_{i,j,k}$$

ora 29

SENZA REPLICHE ($\ell = 1$)



ipotesi fondamentale di linearità

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

media globale; effetto di x_1; effetto di x_2

→ ipotesi tecnica (non restrittiva)

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

$$Y_{i,j} \sim \mathcal{N}(\mu + \alpha_i + \beta_j; \sigma^2)$$ indipendenti

- Test principali

a) Effetto riga

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$ — H_1: non tutti uguali

Y non dipende da x_1 — Y dipende da x_1

b) Effetto colonna

$H_0: \beta_1 = \beta_2 = \dots = \beta_n = 0$ — H_1: non tutti uguali

Y non dipende da x_2 — Y dipende da x_2

- Stimatori elementari

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

(HWC oppure ora 25, 2010)

$$Y_{i,*} = \frac{1}{n}\sum_{j=1}^{n} Y_{i,j} \sim \mathcal{N}\left(\frac{1}{n}\sum_{j=1}^{n}\mu_{i,j};\ \frac{1}{n^2} n\sigma^2\right) \sim \mathcal{N}\left(\mu+\alpha_i;\ \frac{\sigma^2}{n}\right)$$

$$\sum_{j=1}^{n}\mu_{i,j} = \sum_{j}(\mu+\alpha_i+\beta_j) = n\mu + n\alpha_i + \sum_{j=1}^{n}\beta_j = n(\mu+\alpha_i)$$

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

$$\mu+\beta_j \approx Y_{*,j} \sim \mathcal{N}\left(\mu+\beta_j;\ \frac{\sigma^2}{m}\right)$$

$$\Rightarrow \quad \alpha_i \approx Y_{i,*} - Y_{*,*}, \qquad \beta_j \approx Y_{*,j} - Y_{*,*}$$

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

→ Che legge hanno questi stimatori?

È complicato perché ad esempio $Y_{i,*}$ e $Y_{*,*}$ non sono indip.

$$\alpha_i \approx Y_{i,*} - Y_{*,*} \sim \mathcal{N}\left(\alpha_i;\ \frac{m-1}{mn}\sigma^2\right)$$

$$\beta_j \approx Y_{*,j} - Y_{*,*} \sim \mathcal{N}\left(\beta_j;\ \frac{n-1}{mn}\sigma^2\right)$$

$$\mu_{i,j} \approx \underbrace{Y_{i,*} + Y_{*,j} - Y_{*,*}}_{\text{previsto nella casella } i,j} \sim \mathcal{N}\left(\mu_{i,j};\ \frac{m+n-1}{mn}\sigma^2\right)$$

- Stimatore di σ^2 ottimo

→ Servono i residui

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

$$S_e^2 = \frac{1}{(m-1)(n-1)} \sum_{i,j} R_{ij}^2$$

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

applicazione test di Cochran

- Stimatori di σ^2 veri solo sotto H_0

a) Varianza per righe:

$$S_R^2 := \frac{1}{m-1} \sum_{i=1}^{m} n \left(Y_{i,*} - Y_{*,*} \right)^2$$

vero solo in H_0 a)

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

b) Varianza per colonne:

$$S_C^2 := \frac{1}{n-1} \sum_{j=1}^{n} m \left(Y_{*,j} - Y_{*,*} \right)^2$$

vero solo in H_0 b)

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

- Come si fanno i test:

a) Se è vera H_0 (non vi è effetto riga)

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

Se è vera H_1 F_R è tendenzialmente più grande

$$\alpha_R^* = \text{DISTRIB.F}\left(F_R ; m-1 ; (m-1)(n-1) \right)$$

b) Se è vera H_0 (non vi è effetto colonna)

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

Se è vera H_1 F_C è tendenzialmente più grande

$$\alpha_C^* = \text{DISTRIB.F}\left(F_C ; n-1 ; (m-1)(n-1) \right)$$

- Identità delle devianze

$$SS_Y = SS_R + SS_C + SS_e$$

sono indipendenti

$mn-1$ $\quad$ $m-1$ $\quad$ $n-1$ $\quad$ $mn-m-n+1$ $\quad$ i g.d.l.

- Altre inferenze

→ $\alpha_i \approx Y_{i,*} - Y_{**}$

$$\frac{Y_{i,*} - Y_{**} - \alpha_i}{S_e\sqrt{\frac{m-1}{mn}}} \sim t((m-1)(n-1))$$

→ e analoghi per β_j, $\mu_{i,j}$, μ

→ previsto nella casella i,j

$Y'_{i,j} \sim N(\mu_{i,j}, \sigma^2)$ indipendente dai dati

$$Y'_{i,j} - Y_{i,*} - Y_{*,j} + Y_{*,*} \sim N\left(0 ; \sigma^2 \frac{mn+m+n-1}{mn}\right)$$

σ^2 $\quad$ $\sigma^2 \frac{m+n-1}{mn}$

$t((m-1)(n-1))$

$$Y'_{i,j} \in \underbrace{Y_{i,*} + Y_{*,j} - Y_{*,*}}_{\text{previsto}} \pm q\, S_e\sqrt{\frac{mn+m+n-1}{mn}}$$

MMSD

ora 30

Note Title 13/12/2011

Legami tra ANOVA e i test per il confronto di due campioni normali

$X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma^2)$

$Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma^2)$

$H_0: \mu_X = \mu_Y \quad H_1: \mu_X \neq \mu_Y$

$$T := \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \overset{H_0}{\sim} t(n+m-2)$$

t - paired

$X_1, X_2, \dots, X_n$

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

$Y_1, Y_2, \dots, Y_n$

$D_i := X_i - Y_i \sim \mathcal{N}(\mu, \sigma^2)$

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

$$\frac{\bar{D}}{S_D/\sqrt{n}} \overset{H_0}{\sim} t(n-1)$$

ANOVA a 1 via con $m = 2$

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

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

$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$

$$\frac{S_B^2}{S_W^2} \overset{H_0}{\sim} F(1; \overbrace{n_1+n_2-2}^{N-m})$$

ANOVA a 2 vie con $m = 2$ (o $n = 2$) (senza repliche)

$Y_{11} \dots Y_{1,n} \quad H_0: \alpha_1 = \alpha_2 = 0 \quad H_1: \alpha_1 \neq \alpha_2$

$Y_{2,1} \dots Y_{2,m}$

$$\frac{S_R^2}{S_e^2} \overset{H_0}{\sim} F(1; n-1)$$

EQUIVALENTI

★ Vengono p dei dati identici

EQUIVALENTI

★ Vengono p dei dati identici

ANOVA a 2 VIE CON REPLICHE / CON INTERAZIONI / NONLINEARE

$X_{i,j,k} \sim \mathcal{N}(\mu_{i,j}; \sigma^2) \quad i = 1,2,\dots,m \quad j = 1,2,\dots,n \quad k = 1,2,\dots,l \quad l \geq 2$

se $l = 1$ $\quad \mu_{i,j} = \mu + \alpha_i + \beta_j$ ipotesi lineare

se $l > 2$ $\quad \mu_{i,j}$ arbitrario → è comunque sempre possibile scriverlo come:

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

↑ interazioni

con $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$

- Test di linearità

$H_0: \gamma_{ij} = 0 \ \forall i,j$ $\quad$ H_1: non tutti nulli

È di nuovo un test basato sulla F di Fisher

* Vengono effettuati anche i test per righe e per colonne

$$F_{IN} = \frac{S^2_{IN}}{S^2_E} \overset{H_0^{LIN}}{\sim} F\left((m-1)(n-1);\ mn(\ell-1)\right)$$

$$F_R = \frac{S^2_R}{S^2_E} \overset{H_0^{R}}{\sim} F\left(m-1;\ mn(\ell-1)\right)$$

$$F_C = \frac{S^2_C}{S^2_E} \overset{H_0^{C}}{\sim} F\left(n-1;\ mn(\ell-1)\right)$$

$$S^2_E = S^2_P = \frac{1}{mn}\sum_{i,j}\frac{1}{\ell-1}\sum_k \left(Y_{ijk} - Y_{ij*}\right)^2$$

stimatore pooled

$$\frac{S^2_E}{\sigma^2} mn(\ell-1) \sim \chi^2\left(mn(\ell-1)\right)$$

$S_E \approx \sigma$ stimatore ottimo di σ

$$mn(\ell-1) + (m-1)(n-1) + m - 1 + n - 1 = \underbrace{mn(\ell-1) + mn}_{mn - m - n + 1} - 1 = mn\ell - 1 = N - 1$$

$$SS_{IN} + SS_R + SS_C + SS_E = SS_Y$$

tra loro indipendenti

- Altra inferenza

$$\mu_{ij} \approx Y_{ij*} \sim \mathcal{N}\left(\mu_{ij};\ \frac{\sigma^2}{\ell}\right)$$

(un esempio)

$$\frac{Y_{ij*} - \mu_{ij}}{S_E/\sqrt{\ell}} \sim t\left(mn(\ell-1)\right)$$

- Vedi laboratorio 9 per l'analisi dei residui

Relazione tra i tre tipi di ANOVA

a) Se si può fare l'ANOVA a 2 vie con repliche
allora si può fare l'ANOVA a 2 vie lineare (si mediano le repliche)
se si può fare quest'ultima
allora si può fare l'ANOVA a 1 via (basta ignorare una delle due var. di ingresso)

b) È legittimo solo se i test ce lo consentono
se H_0: modello lineare, medio le repliche
se H_0: niente effetto riga (colonna) passo a una via

c) Qualunque risultato H_1 va considerato legittimo

d) Se si può fare un'ANOVA "maggiore" e non si fa
è un errore: tale errore è grave se il test dice H_0

★ (Anche nella regressione) se una variabile o un effetto sono "impattanti", non vanno esclusi dal modello nemmeno quando ci interessano solo altre variabili. Il non conoscere tale variabile si trasforma in rumore.

Rapporti con la regressione

- Test globale di regressione $\quad Y_i \sim \mathcal{N}\left(\underbrace{\sum_{j=0}^{p}\beta_j x_{ij}}_{\mu_i};\ \sigma^2\right)$

$H_0: \mu_1 = \dots = \mu_n \qquad H_1: \mu_i$ dipende da $x_{i1} \dots x_{ip}$

$n-1 \quad n-p-1 \quad p$

$$SS_Y = SS_R + SS_D$$

(nuovo oggetto)

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

$$S_D^2 = \frac{SS_D}{p} \qquad S_e^2 = \frac{SS_R}{n-p-1}$$

$$F_G := \frac{S_D^2}{S_e^2} \overset{H_0}{\sim} F(p;\ n-p-1)$$

- Variabili numeriche e categoriche insieme

Due strade possibili, ognuna con difetti

a) esplodo le categoriche in dicotomiche e uso la regressione
   - → n/p peggiora
   - → test multipli e Bonferroni
   - → confronti sempre relativi a una categoria

b) faccio una fra regressione e ANOVA, inserendo solo le variabili relative; alle fine prendo i residui e faccio l'altra tecnica reinserendo le var accantonate prima e usando i residui come dati

x_1	x_2	x_3	C_4	C_5	Y	R_{ANOVA} = Y_{REG}
~	~	~	~	~	~	~
~	~	~	~	~	~	~

- → non è una analisi quantitativamente robusta
- → la prima tecnica risente delle var. accantonate come rumore.

TEST DI ADATTAMENTO / DEL CHI-QUADRO

TEST DEL CHI-QUADRO ELEMENTARE

esempio: ho un dado da gioco e voglio verificare se è onesto

→ lo tiro n volte (n grande)

$X_1, X_2, \dots, X_n$ gli esiti $O_1, O_2, \dots, O_6$ il numero di volte che sono usciti i numeri da 1 a 6

$$O_1 + O_2 + \dots + O_6 = n$$

Se il dado è onesto, mi aspetto che $O_i \approx \frac{n}{6} = A_i$

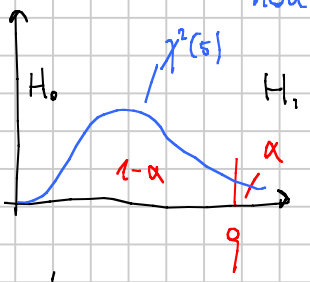
(O_i: osservati, A_i: attesi)

$$W := \sum_{i=1}^{6} \underbrace{\frac{(O_i - A_i)^2}{A_i}}_{\text{scarto } i}$$ statistica del test

Se W è grande, (H_1) il dado è disonesto, se W è piccola, (H_0) è onesto

$$W \overset{H_0}{\sim} \chi^2(6-1)$$

(non conta n, ma solo quante alternative c'erano)



$W > q$ dico H_1

$W \leq q$ dico H_0

★ Formalmente

i. è dato un campione $X_1, X_2, \dots, X_n$ (con n grande) di variabili i.i.d. di legge incognita

ii. si ipotizza per questo campione una particolare legge discreta φ : $\varphi(j) := P(X = j)$ che ammette un piccolo numero k di valori possibili con lo scopo di testare se questa legge sia plausibile per il campione

iii. $H_0 : \varphi_{X_i} = \varphi$ ovvero $P(X_i = j) = \varphi(j)$ per ogni j

$H_1 : \varphi_{X_i} \neq \varphi$

— altro esempio: dati Istat —

$\varphi \longrightarrow$ laurea: 4%, maturità: 75%, III media: 7%, meno: 14%

prendo $n = 200$ residenti a Pisa e conto le categorie

$O_i \longrightarrow$	15	164	12	9
$A_i \longrightarrow$	8	150	14	28

iv. O_i : numero di $X_1, X_2, \dots, X_n$ nella categoria i (li conto)

A_i : atteso della categoria i

$$A_i = n\varphi(i)$$

statistica del test : $$W := \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i} \overset{H_0}{\sim} \chi^2(k-1)$$

v. La RA_W va sempre presa del tipo $[0, q]$ unilaterale sinistra

$\alpha^* = \text{DISTRIB.CHI}(\text{stat} ; \text{gdl})$

★ L'approssimazione è buona se n è grande rispetto a k, o meglio se i numeri $A_i = n\varphi(i)$ sono grandi

→ "rule of thumb"

i. tutti gli $A_i \geq 5$ tranne al più uno
ii. il min $A_i \geq 1$ comunque

→ si può evitare la approssimazione con una simulazione MC

Si usa la simulazione per capire la legge di W sotto H_0

$H_0 : \varphi_{X_i} = \varphi$ → genero N un campione di lunghezza n di var. con legge φ ; per ciascun campione calcolo W ; ottengo un campione di num. N di var con legge di W sotto H_0

↳ $$\alpha^* = \frac{\#\{W_j \text{ tali che } W_j \geq W_{\text{dati veri}}\}}{N}$$

(da vedere in lab)

ora 32

★ Non sempre gli A_i sono interi e non occorre che lo siano

★ Il test funziona meglio se gli A_i sono più o meno simili e tutti grandi

GENERALIZZAZIONE A LEGGI φ CONTINUE O CON TANTI VALORI

→ in questi casi bisogna definire dei **bin** ovvero degli intervalli di valori da usare al posto delle categorie

↳ k è il numero dei bin

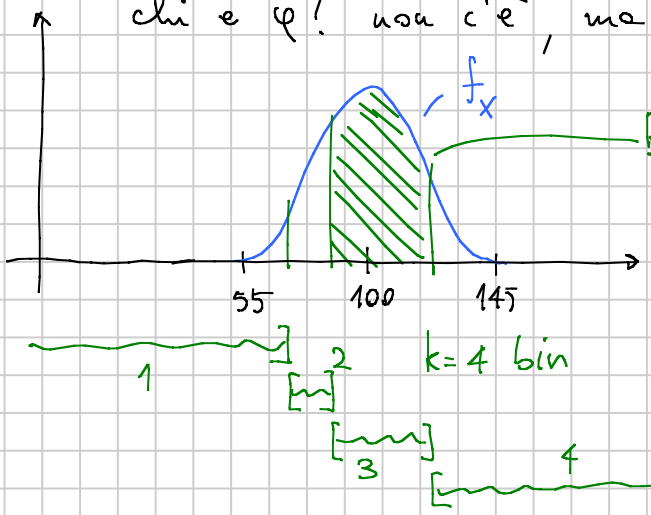
↳ O_i il numero di dati che cadono nel bin i

→ per il resto è tutto uguale

★ Esempio QI ha legge $\mathcal{N}(100, 15^2)$

potrei domandarmi se il QI degli studenti universitari segue la stessa distribuzione

chi è φ? non c'è, ma c'è f

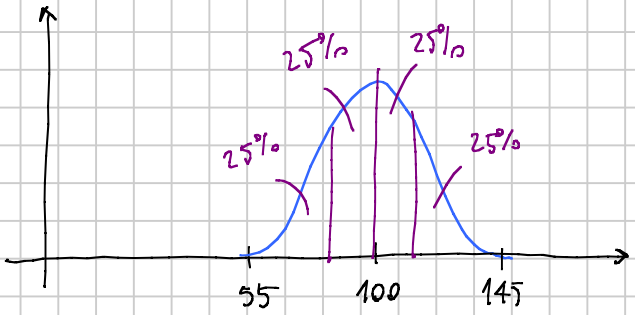


$$= P(\text{bin } 3)^n = P(X \in \text{bin } 3) = \int_{\text{bin } 3} f(t)\,dt$$

$$=: \varphi(3)$$

$$\varphi(i) := \int_{\text{bin } i} f(t)\,dt$$

$$A_i = n\,\varphi(i) = n \int_{\text{bin } i} f(t)\,dt$$



★ Conviene, per avere gli A_i uguali tagliare i bin sui quantili/percentili

→ Se ho fissato k

bin 1 : $(-\infty ; a_1]$

bin 2 : $[a_1 ; a_2]$

. . . . . .

$$a_1 = \text{INV.NORM}\left(\frac{1}{k} ; 100 ; 15\right)$$

$$a_2 = \text{INV.NORM}\left(\frac{2}{k} ; 100 ; 15\right)$$

$$a_{k-1} = \text{INV.NORM}\left(\frac{k-1}{k} ; 100 ; 15\right)$$

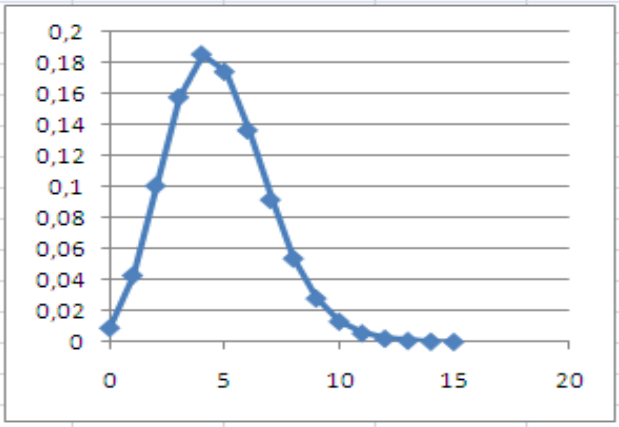
$$\varphi(i) \equiv \frac{1}{k} \quad \forall i$$

$$A_i \equiv \frac{n}{k} \quad \forall i$$

★ Esempio $\varphi : \text{Pois}(4,7)$

Poisson di media 4,7

4,7				
0	0,009095277			
1	0,042747802			
2	0,100457336	0,1523		0,15
3	0,157383159		0,31	0,16
4	0,184925212	0,3423	0,18	0,18
5	0,173829699	0,1738		0,17
6	0,136166598			0,14
7	0,091426144			0,2
8	0,05371286			
9	0,028050049			con k=6
10	0,013183523			
11	0,00563296			
12	0,002206243			
13	0,000797642			
14	0,00026778			
15	8,39043E-05	0,3316		
		1		



In questo caso i bin non sono proprio equiprobabili, ma va bene lo stesso.

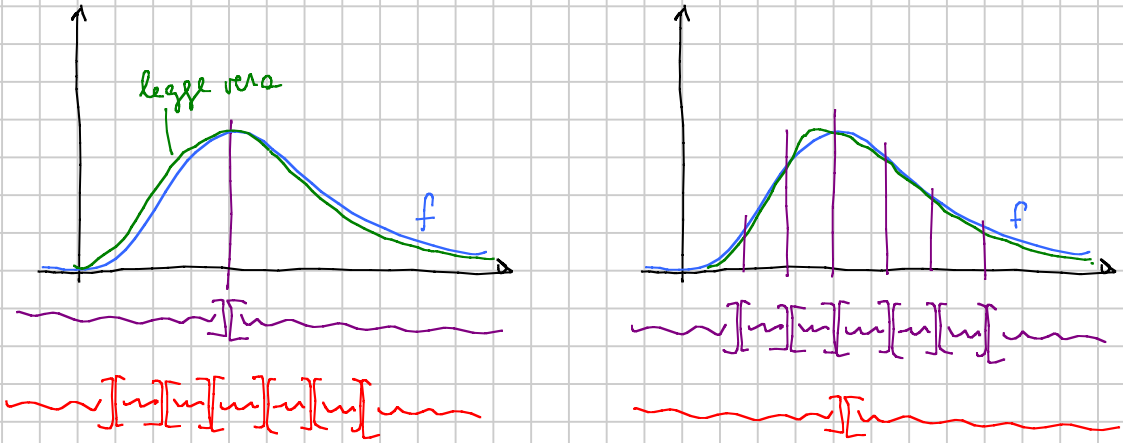
★ Quanti bin?

→ mai troppi rispetto a n, perché gli attesi devono essere grandi

A parte ciò è una scelta soggettiva

i. più k è piccolo più il test è potente
(nel senso che si accorge di piccole differenze di probabilità)

ii più k è grande più il test è sensibile
(si accorge di deviazioni locali)



GENERALIZZAZIONE A CLASSI DI V.V.A.A. SPECIFICATE A MENO DI PARAMETRI

Esempio: questo campione è normale?

i. Si stimano puntualmente i parametri

$$\mu \approx \bar{X} \qquad \sigma \approx S$$

ii. Si testa la legge $\mathcal{N}(\bar{x}; s^2)$ con i parametri stimati

→ avendo cura di calare i gdl della $\approx \chi^2$ di un numero (1 o 2) corrispondente ai parametri stimati

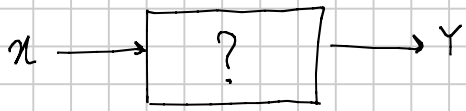
$$gdl = k - 1 - \text{parametri stimati}$$

MMSD

ora 33

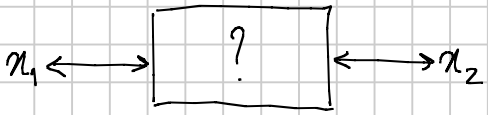
Note Title 19/12/2011

TABELLE DI CONTINGENZA



→ regressione (non) lineare semplice

→ ANOVA a 1 via se x è categorica



→ tabella di contingenza

→ studia la relazione tra due variabili

→ non impone restrizioni su di esse: possono avere qualunque distribuzione

→ il prezzo da pagare è che il test è poco potente

→ la generalità si ottiene facendo diventare le variabili categoriche con poche categorie (si divide in bin, possibilmente equiprobabili)

→ simmetria: si cercano relazioni tra le due variabili

- Struttura dei dati ed esempio

Esempio X: ospedale Y: tipo di parto dato: singolo parto

$X_1, X_2, \dots, X_n$ categorica $\{DH, GA, ER\}$ m_1 categorie

$Y_1, Y_2, \dots, Y_n$ categorica $\{N, I, C\}$ m_2 categorie

uguale

id	X	Y
1	GA	N
2	ER	I
⋮	⋮	⋮
n		

	DH	GA	ER
N	10	25	15
I	15	40	30
C	20	30	20

si riempie con i conteggi

c'è una relazione

	DH	GA	ER	
N	5	10	30	45 (30%)
I	35	30	5	70 (50%)
C	5	20	5	30 (20%)
	45	60	40	145

non c'è relazione

	DH	GA	ER
N	5 (20%; 12.5%)	20 (20%; 50%)	15 (20%; 37.5%)
I	10 (40%; 12.5%)	40 (40%; 50%)	30 (40%; 37.5%)
C	10 (40%; 12.5%)	40 (40%; 50%)	30 (40%; 37.5%)

perfetta indipendenza tra le variabili

	DH	GA	ER	
N	13,5 (30%)	18 (30%)	12 (30%)	45
I	22,5 (50%)	30 (50%)	20 (50%)	70
C	9 (20%)	12 (20%)	8 (20%)	30
	45	60	40	145

"osservati"

"attesi" in caso di indipendenza

cosa mi sarei aspettato di vedere se ci fosse indipendenza e i numeri aggregati/marginali fossero quelli della prima tabella

test del χ^2

$$W := \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(O_{i,j} - A_{i,j})^2}{A_{i,j}} \overset{H_0}{\sim} \chi^2\left((m_1-1)(m_2-1)\right)$$

$$A_{i,j} = \frac{\sum_k O_{i,k} \cdot \sum_h O_{h,j}}{\sum_{k,h} O_{k,h}}$$

$\sum_k O_{i,k}$: totale riga i; $\sum_h O_{h,j}$: totale colonna j; $\sum_{k,h} O_{k,h}$: n numero di osservazioni

- Ipotesi formali del test

H_0: le due var x_1, x_2 sono indipendenti H_1: no ....

il test è unilaterale come i precedenti test del χ^2

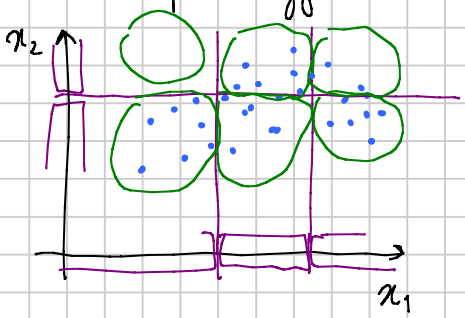
$RA_W : [0 ; q]$ $q = \text{INV.CHI}\left(\alpha ; (m_1-1)(m_2-1)\right)$

★ Possibili letture alternative del test

H_0: la percentuale dei parti N/I/C è la stessa nei tre ospedali H_1: no.

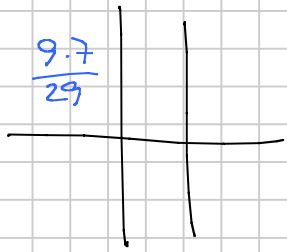
H_0: no... H_1: la var ospedale ha un impatto su quella tipo di parto

- Sul passaggio da numeriche a categoriche



0	6	3	9
7	7	6	20
7	13	9	29

→ attesi



il numero complessivo di bin diventa $m_1 \cdot m_2$ e non deve essere troppo grande (dell'ordine di una decina)

- Sull'approssimazione

$W \overset{H_0}{\sim} \chi^2$ l'approssimazione è buona se gli attesi sono grandi → rule of thumb

tutti ≥ 5 tranne al più uno ≥ 1

* Se l'approssimazione è cattiva, non si risolve con una simulazione MC
* Solo nel caso 2×2 esiste però anche un test esatto "Test di Fisher-Irwin" (cfr ore 30 anno scorso)

7	2	9
4	9	13
11	11	22

4,5	4,5	9
6,5	6,5	13
11	11	22

N		9
		13
11	11	22

N sotto H_0 ha legge

IPERGEOMETRICA$\left(9 ; \overset{11}{11} ; \overset{9}{22}\right)$