

Tema di Statistica Matematica

Simulazione

1) Sia (X_1, X_2, \dots, X_n) un campione casuale da una popolazione distribuita secondo $N(\theta, 1)$. Stabilire se lo stimatore $\sum_{i=1}^n X_i$ è uno stimatore *UMVU* di $n\theta$. È anche efficiente?

2) Sia (X_1, X_2, \dots, X_n) un campione casuale semplice da una distribuzione Beta di parametri (α, β) con $\alpha = \theta > 0$, $\beta = \theta^2$ e funzione di densità

$$f(y; \theta) = \frac{\Gamma(\theta + \theta^2)}{\Gamma(\theta) \cdot \Gamma(\theta^2)} x^{\theta-1} (1-x)^{\theta^2-1} \mathbb{I}_{(0,1)}(x).$$

a) Fornire una statistica sufficiente minimale per l'inferenza su θ .

b) Sia

$$\hat{\theta} = (1/\bar{X}) - 1$$

uno stimatore di θ . Stabilire se $\hat{\theta}$ è stimatore non distorto e consistente per θ .

c) Fornire un'approssimazione asintotica della distribuzione di $\hat{\theta}$.

3) Il responsabile di un certo corso di laurea ritiene che la percentuale di studenti iscritti al corso interessati al servizio di tutoraggio sia pari al 45%. Per programmare meglio il servizio, decide di condurre un'indagine campionaria in cui vengono selezionati casualmente e intervistati 500 studenti tra quelli iscritti al corso in questione; di questi, 220 si dichiarano interessati al tutoraggio.

a) Costruire un intervallo di confidenza al 95% per la proporzione degli studenti del corso di laurea interessati al programma di tutoraggio.

b) Se 300 dei 500 studenti intervistati avessero dichiarato che non intendono avvalersi del tutoraggio, ci sarebbero validi motivi per rivedere l'opinione iniziale del responsabile del percorso di studio relativa alla percentuale di studenti interessati al programma di tutoraggio?

4) Sia (X_1, X_2, \dots, X_n) un campione casuale da distribuzione Uniforme su $[\theta, \theta + 1]$. Dimostrare che la statistica sufficiente minimale per θ non è completa.

Suggerimento: nel caso in questione, $R \sim Beta(n-1, 2)$, dove R è il range di X .

5) Siano $Y_1 < Y_2 < Y_3 < Y_4$ le statistiche ordinate di un campione casuale di ampiezza $n = 4$ proveniente da una distribuzione uniforme sull'intervallo $(0, \theta)$ con $\theta > 0$. Volendo sottoporre a verifica il seguente sistema di ipotesi

$$H_0 : \theta = 1 \quad vs. \quad H_1 : \theta > 1$$

si decide di rifiutare H_0 a favore di H_1 ognqualvolta $Y_4 \geq c$.

- a) Trovare la costante c tale per cui il livello di significativitá α del test sia uguale a 0.05.
- b) Determinare la funzione di potenza del test e disegnarne il grafico.

6) Nel corso di uno studio sulle proprietá di alcune leghe metalliche si é ipotizzata una relazione lineare tra la concentrazione di carbonio (x) e la tensione di snervamento (Y) delle stesse; sulla base di questa assunzione si é costruito un modello di regressione lineare $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ che é poi stato stimato sulla base di un campione di 9 osservazioni $\{(x_i, Y_i), i = 1, 2, \dots, 9\}$, ottenendo

$$\hat{Y}_i = 3.858 + 1.3925 x_i, \quad i = 1, 2, \dots, 9. \quad (1)$$

- a) Sapendo che $SSTO = 846.00$ e $SSE = 70.38$, calcolare il coefficiente di correlazione tra x e Y .
- b) Valutare, ricorrendo a un test opportuno di livello $\alpha = 0.05$, la bontá di adattamento ai dati del modello di regressione ipotizzato e stimato in (1).
- b) Sapendo che la varianza di b_1 , stimatore a minimi quadrati del coefficiente di regressione β_1 , é risultata pari a 0.025, costruire un intervallo di confidenza di livello 0.99 per β_1 .
- c) Di quanto varia la tensione di snervamento per un incremento unitario nella concentrazione di carbonio?
- d) Quale prevedete sia la tensione di snervamento in presenza di una concentrazione di carbonio nella lega metallica pari a 50?

DX.1)

$$(X_1, \dots, X_n) \text{ da } N(\theta, 1) \neq f_{X_i}(x_i; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (x_i - \theta)^2} \prod_{i=1}^n$$

a) $T_n(X_1, \dots, X_n) = \sum_{i=1}^n X_i \quad \text{con } \theta \in \mathbb{R}$

$$\cdot E(T_n) = \sum_{i=1}^n E(X_i) = n\theta, \quad \forall \theta \in \mathbb{R}$$

b) $T_n = \sum_{i=1}^n X_i$ è STFT. MFF per θ sicché

lo stimatore T_n soli $n\theta$ solta da essere NON
DISTORSO e anche FUNZIONE soli STFT. MFF (minimale)
per θ sicché in virtù del Corollario sulle
THM soli RGO-BLACKWELL è STRATTORE UNIV.

c) $\text{Var}_\theta(T_n) = \text{Var}_\theta\left(\sum_{i=1}^n X_i\right) = n(\text{Var}(X_i)) = n \cdot 1 = n$

e

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2} \ell(\theta; \underline{x})\right]$$

$$= -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2} \left(-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right)\right]$$

$$= -\mathbb{E}_\theta\left[\frac{d}{d\theta}\left(-\frac{1}{2} \cdot (-2) \sum_{i=1}^n (x_i - \theta)\right)\right]$$

$$= -\mathbb{E}[-n] = n$$

sicché
 $n = \text{Var}_\theta(T_n) \geq \frac{n^2}{I_n(\theta)} = \frac{n^2}{n} = n$ con T_n stim. u. Non distorto
di n.B.

Q.2)

a) La f.d. è data da

$$L(\theta, \theta^2; z) = \left[\frac{\Gamma(\theta + \theta^2)}{\Gamma(\theta) \Gamma(\theta^2)} \right]^n \left(\prod_{i=1}^n z_i \right)^{\theta-1} \left(\prod_{i=1}^n (1-z_i) \right)^{\theta^2-1}$$

$\prod_{i=1}^n [0,1]^{z_i}$

Sicché in vista del Thm di PREFERENZA di NEYMAN

$$g(\theta, \theta^2; t'_n, t_n) = \left[\frac{\Gamma(\theta + \theta^2)}{\Gamma(\theta) \Gamma(\theta^2)} \right] \left(\prod_{i=1}^n z_i \right)^{\theta-1} \left(\prod_{i=1}^n (1-z_i) \right)^{\theta^2}$$

$$h(z) = \prod_{i=1}^n [0,1]^{z_i}$$

e $\left(\prod_{i=1}^n X_i, \prod_{i=1}^n (1-X_i) \right)$ è STAT. MFF per (θ, θ^2) .
(congiuntamente)

Tenuto conto del fatto che

$$\frac{L(\theta, \theta^2; z)}{L(\theta, \theta^2; y)}$$

NON DIPENDE in θ a patto che $\prod_{i=1}^n X_i = \prod_{i=1}^n Y_i$ e che
 $\prod_{i=1}^n (1-X_i) = \prod_{i=1}^n (1-Y_i)$, la STAT. MFF può anche molti dati
e' anche per il Thm di LEHRHORN-SCHAFER, MINIMALE.

b) Elenchiamo $X \sim \text{Beta}(\theta, \theta^2)$ si avrà

$$E(X) = \frac{\theta}{\theta + \theta^2}$$

e

$$\text{Var}(X) = \frac{\theta^3}{(\theta + \theta^2)^2 (\theta + \theta^2 + 1)}$$

Ricordiamo la divergenza di JENSEN

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}(X))$$

Con g PIAZZONE CONVESA, e posto $g(x) = \frac{1}{x}$ si ha

$$\mathbb{E}\left(\frac{1}{\bar{X}}\right) \geq \frac{1}{\mathbb{E}(\bar{X})} = \frac{1}{\frac{\theta + \theta^2}{\theta}} = \frac{\theta + \theta^2}{\theta} = \frac{\cancel{\theta}(1+\theta)}{\cancel{\theta}} = \theta + 1$$

essendo \bar{X} uno stimatore NON DISTORTO della media $\mathbb{E}(X)$ della popolazione. Ora

$$\mathbb{E}\left(\frac{1}{\bar{X}} - 1\right) \geq \frac{1}{\mathbb{E}(\bar{X})} - 1$$

$$\mathbb{E}\left(\frac{1}{\bar{X}} - 1\right) \geq \frac{\theta + \theta^2}{\theta} - 1 = \theta$$

ricci $\hat{\theta} = \frac{1}{\bar{X}} - 1$ è STIMATORE DISTORTO di θ

D'altra parte, dato che la media campionaria è STIMATORE CONSISTENTE della media della popolazione θ .

$$\frac{1}{n} \sum_{i=1}^n x_i \xrightarrow{P} \frac{\theta}{\theta + \theta^2}$$

ed essendo $g(z) = \frac{1}{z} - 1$ una PIAZZONE CONTINUA, per il Thm dei CONTINUI MAPPING

$$g(\bar{X}) = \frac{1}{\bar{X}} - 1 \xrightarrow{P} \frac{1}{\frac{\theta + \theta^2}{\theta}} - 1 = \frac{\theta^2 + \theta - \theta}{\theta} = \theta = g(\theta)$$

ricci $\hat{\theta} = \frac{1}{\bar{X}} - 1$ è STIMATORE DEMOLENTEMENTE CONSISTENTE per θ .

c) Per il TLC, \bar{X} ha dist. statisticamente normale con
medio $E(\bar{X})$ e varianza $\frac{Var(\bar{X})}{n}$. Nel nostro caso

$$Var(\bar{X}) = \frac{\theta^3}{(\theta + \theta^2)(\theta + \theta^2 + 1)} = \frac{\theta}{(1+\theta)^4 - \theta(1+\theta)^2}$$

D'altra parte la funzione $g(z) = \frac{1}{z} - 1$ ha derivate

$$g'(z) = -\frac{1}{z^2} \Rightarrow (g'(z))^2 = \frac{1}{z^4}$$

e applicando il METODO DELLA si ha che

$$\hat{\theta} \underset{n}{\sim} N\left(g(E(\bar{X})), \underbrace{\left(\frac{1}{E(\bar{X})}\right)^4 \frac{Var(\bar{X})}{n}}_{[g'(E(\bar{X}))]^2}\right)$$

con

$$g(E(\bar{X})) = \frac{1}{E(\bar{X})} - 1 = \frac{1}{\frac{\theta}{\theta + \theta^2}} - 1 = \theta$$

$$[g'(E(\bar{X}))]^2 = \left[-\frac{1}{(E(\bar{X}))^2}\right]^2 = \left[\frac{1}{E(\bar{X})}\right]^4 = \left[\frac{1}{\frac{\theta}{\theta + \theta^2}}\right]^4 = \frac{(\theta + \theta^2)^4}{\theta^4}$$

ricordi

$$\left[\frac{1}{E(\bar{X})}\right]^4 \cdot \frac{Var(\bar{X})}{n} = \frac{(\theta + \theta^2)^4}{\theta^4} \cdot \frac{\theta^3}{n(\theta + \theta^2)^2(\theta + \theta^2 + 1)} \\ = \frac{(\theta + \theta^2)^2}{n\theta(\theta + \theta^2 + 1)} = \frac{\theta^2(1+\theta)^2}{n\theta(\theta + \theta^2 + 1)}$$

$$= \frac{\theta(1+\theta)^2}{n(\theta + \theta^2 + 1)} = \frac{\theta(1+\theta)^2}{n[(\theta+1)^2 - \theta]}$$

Esercizio 3)

Dai dati a disposizione

$$p = p_0 = 0,45$$

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n x_i = \frac{220}{500} = 0,44 \quad \text{con } x_i \sim b(1, p)$$

a) Ricordando che qui il TLC

$$\hat{p}_n \xrightarrow{n} N(p, \frac{p(1-p)}{n}) \Rightarrow \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{n} N(0,1) \rightarrow \frac{\hat{p}_n(1-\hat{p}_n)}{n} \xrightarrow{n} \frac{p}{1-p}$$

si ha

$$IC_p(1-\alpha): \left[\hat{p}_n - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right]$$

$$= \left[0,44 - 1,96 \sqrt{\frac{0,2464}{500}}, \right.$$

$$= \left[0,44 - \frac{0,0435}{1,96 \cdot 0,0222}, \right]$$

$$= [0,396, 0,484]$$

b) si tratta di verifica

$$\begin{cases} H_0: p = 0,45 = p_0 \\ H_1: p \neq 0,45 \end{cases} \equiv IC_p(0,95): \left[\frac{200}{500} + 1,96 \sqrt{\frac{\frac{2}{5}(1-\frac{2}{5})}{500}}, \frac{2}{5} + \dots \right] = [0,357, 0,443]$$

Dal momento che $p_0 = 0,45$ non appartiene a $IC_p(0,95)$ si può REFIUTARE H_0 .

Esercizio 4)

Dimostrare che $T_{\text{r}}(X_1, \dots, X_n)$ non è STATISTICA COMPLETA
implica trovare $g(T_{\text{r}}(\underline{X}))$ tale che $E(g(T_{\text{r}}(\underline{X}))) = 0$
per ogni $\theta \in \Theta$ ma $g(T_{\text{r}}(\underline{X})) \neq 0$.

Un cumulo naturale è

$$R = X_{(n)} - X_{(1)}$$

dal momento che si può dimostrare che la sua distribuzione
NON DIPENDE da θ ; in questo contesto di q.s. esercizio

$$R \sim \text{Beta}(m-1, 2)$$

e ricordando che per $X \sim \text{Beta}(\alpha, \beta)$ si ha

$$E(X) = \frac{\alpha}{\alpha+\beta}$$

ricchi

$$E(R) = \frac{m-1}{(m-1)+2} = \frac{m-1}{m+1}$$

che NON dipende da R e $E(R - E(R)) = 0, \forall \theta \in \Theta$.

Perciò

$$\begin{aligned} g(T_{\text{r}}) &= g(X_1, \dots, X_n) = (X_{(n)} - X_{(1)}) - \frac{m-1}{m+1} \\ &= R - \frac{m-1}{m+1} = R - E(R) \end{aligned}$$

è una funzione NON nulla il cui valore atteso è
SEMPRE NULLO. Da ciò segue immediatamente che
la STATISTICA CONGIUNTAMENTE INSUFFICIENTE $(X_{(1)}, X_{(n)})$
per $H_0(\theta_0, \theta_1)$ NON È COMPLETA.

Ex. 5)

Com'è ricavare la DISTRIBUZIONE della m -ma STATISTICA d'ORDINE
 $Y_{(m)}$ che si chiama anche DENSITÀ?

$$f_{Y_{(m)}}(y; \theta) = m \binom{n}{m} f_Y(y; \theta) \left[F_Y(y; \theta) \right]^{m-1} \left[1 - F_Y(y; \theta) \right]^{n-m}$$
$$= m \binom{n}{m} \frac{1}{\theta} \left(\frac{y}{\theta} \right)^{m-1} \left(1 - \frac{y}{\theta} \right)^{n-m}$$

Allora, essendo $n=4$, avremo

$$f_{Y_{(4)}}(y; \theta) = 4 \binom{4}{4} \frac{1}{\theta} \left(\frac{y}{\theta} \right)^{4-1} = 4 \frac{y^3}{\theta^4}$$

a) Sulla base di quest'ultimo risultato e ricordando che d'altrò non è che la PROBABILITÀ di commettere un ERRORE di 1° TIPO (vale a dire RIFUTARE H_0 quando è VERA) proviamo scrivere essendo fissato $\alpha = 0,05$

$$0,05 = P(Y_{(4)} \geq c \mid \theta = 1)$$

$$= \int_c^{\theta} \frac{4y^3}{\theta^4} dy \Big|_{\theta=1} = \int_0^1 4y^3 dy = y^4 \Big|_0^1 = 1 - c^4$$

dai cui

$$1 - c^4 = 0,05$$

e quindi

$$c = (0,95)^{1/4} = 0,987$$

b) Convien ricordare che

$$\eta_{C_\alpha}(\theta) = P(Y_{(4)} > 0.987 \mid \theta > 1)$$

$$= \int_{0.987}^{\theta} \frac{4y^3}{\theta^4} dy = 1 - \left(\frac{0.987}{\theta}\right)^4$$

e allungare

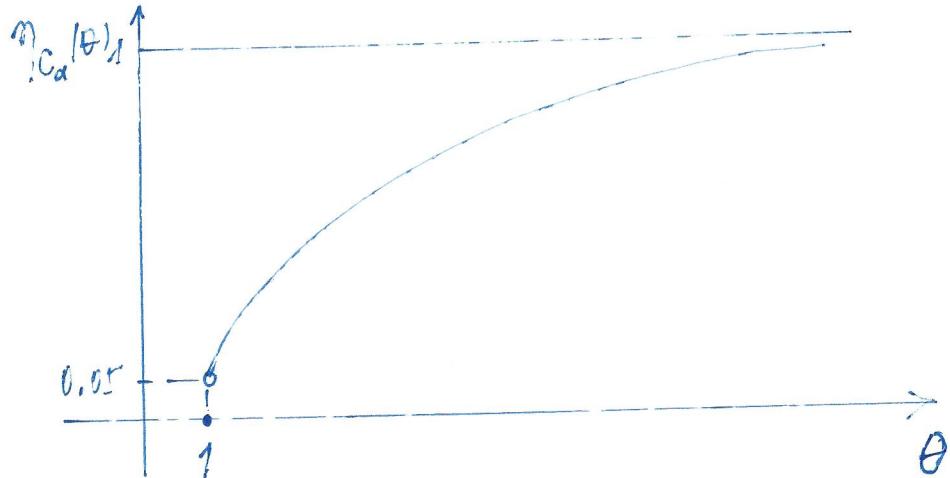
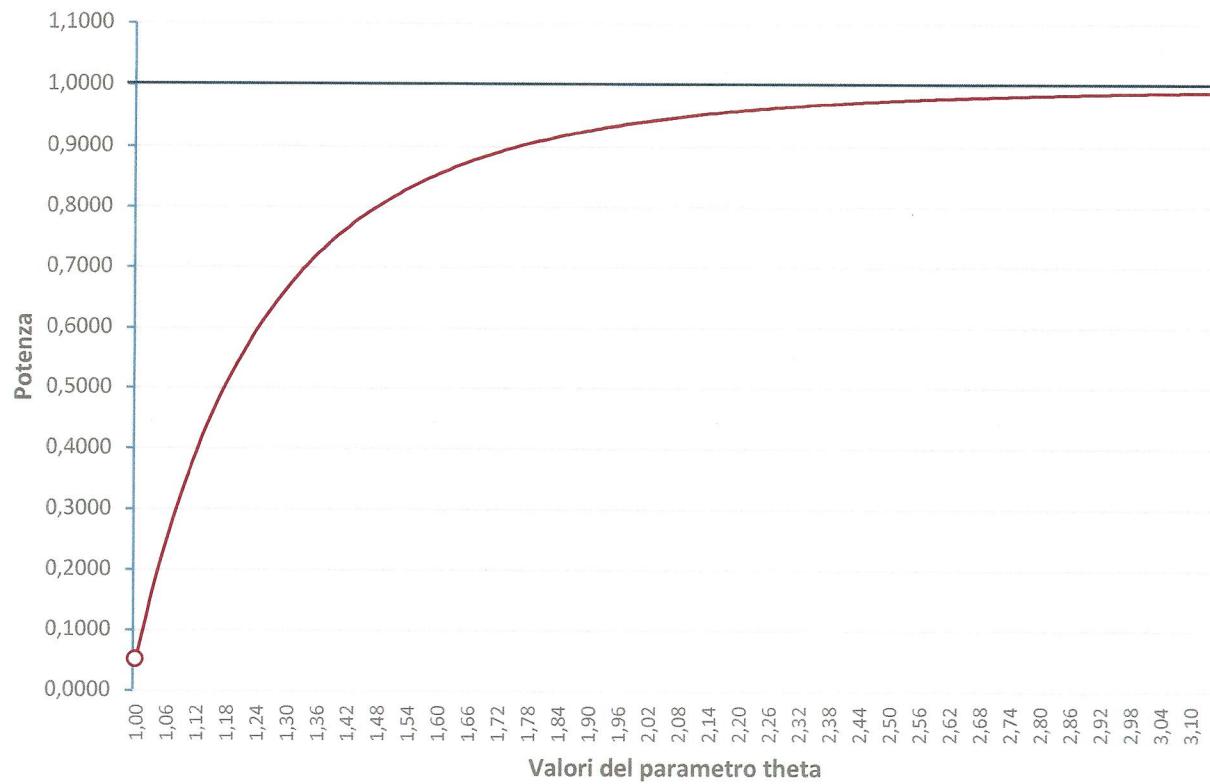


Grafico della funzione di potenza



EX 6)

a) È immediato calcolare il COEFFICIENTE di DETERMINAZIONE

$$R^2 = \frac{SJR}{SSTO} = \frac{SSTO - SSE}{SSTO} = \frac{846 - 70,38}{846} = 0,917$$

e ricordando che

$$\text{Cor}(x, Y)^2 = R^2$$

$$\text{si ha che } \text{Cor}(x, Y) = \sqrt{0,917} = 0,957$$

b) Come abbiamo avuto modo di vedere nel Corso delle lezioni, proseguiamo costruire un test per la BONTÀ di ADATTAMENTO del modello ai dati riconciliando alla STATISTICA

$$F = \frac{SJR/p-1}{SSE/n-p} = \frac{MSR}{MSE} \sim F_{p-1, n-p}$$

che può essere impiegata nella COSTRUZIONE di un test per la verifica sulla SIGNIFICATIVITÀ STATISTICA delle COEFFICIENTI di DETERMINAZIONE R^2 , che misura la BONTÀ di ADATTAMENTO DEL MODELLO AI DATI

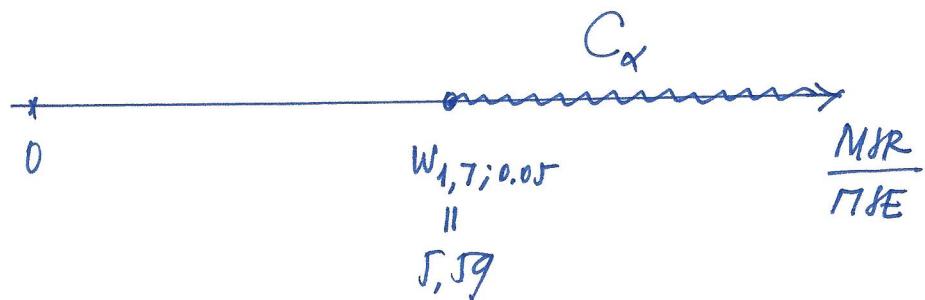
In altre parole si testa l'ipotesi

$$\begin{cases} H_0: R^2 = 0 & (\text{NON HA SIGNIFICATIVITÀ STATISTICA DEL MODELLO}) \\ \text{vs.} \\ H_1: R^2 > 0 \end{cases}$$

rimane associata la regine CRITICA

$$C_\alpha = \left\{ x \in \mathcal{X} : \frac{MSR}{MSE} > w_{1,7;\alpha} \right\}$$

Con $W_{1,7;\alpha}$ a rappresentare il limite superiore della decisione di REJUTTO sull' $H_0: R^2=0$ costruito sulla base di una distribuzione di FISHER/PNEDCOCK con $p-1 = 2-1=1$ e $n-p = 9-2=7$ gradi di libertà ossia, fissato $\alpha = 0.05$,



Nel nostro caso

$$\frac{MSE}{SSE} = \frac{HR/p-1}{SSE/n-p} = \frac{773,62/2-1}{70,38/9-2} = 77,15$$

che cade a destra del valore critico 7,59 a sottolineare la necessità di REJUTTO dell'ipotesi nulla $H_0: R^2=0$ a favore della SIGNIFICATIVITÀ statistica sulle misure di regressione proposto.

c) Come abbiamo già avuto modo di vedere nell'altra soluzione, la STATISTICA PIVOT attorno alla quale costruire l'INTERVALLO di CONFIDENZA per il parametro β_1 è data da

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}_{\hat{\beta}_1}^2}} \sim t_{n-2}$$

Con qualche passaggio a perline da

$$P \left[-t_{n-2; \alpha/2} \leq \frac{b_1 - \beta_1}{\sigma_{b_1}} \leq t_{n-2; \alpha/2} \right] = 1 - \alpha$$

si arriva a 1 soluz β_1 ossia

$$P \left[b_1 - t_{n-2; \alpha/2} \sigma_{b_1} \leq \beta_1 \leq b_1 + t_{n-2; \alpha/2} \sigma_{b_1} \right] = 1 - \alpha$$

ossia ottenere l'INTERVALLO di CONFIANZA cercato

$$IC_{\beta_1}(1-\alpha) = [b_1 - t_{n-2; \alpha/2} \sigma_{b_1}, b_1 + t_{n-2; \alpha/2} \sigma_{b_1}]$$

Che, stabilità circa $(1-\alpha) = 0.99$, $\sigma_{b_1}^2 = 0.025$ (dati a disposizione), $b_1 = 1.3925$ e $t_{7; 0.005} = 3.499$, risulta uguale a

$$\begin{aligned} IC_{\beta_1}(0.99) &= [1.3925 - 3.499 \sqrt{0.025}, 1.3925 + 3.499 \sqrt{0.025}] \\ &= [0.8392, 1.9457] \end{aligned}$$

a) Per ogni INCREMENTO UNITARIO nella CONCENTRAZIONE di CROMO si ha UNA MIGLIORE PREVISIONE di Y aumentando di 1.3925, quindi un incremento di 100% alla STIMA del COEFFICIENTE DI REGRESSIONE β_1 .

e) La PREVISIONE del valore di Y, tensione di funzionamento, in corrispondenza di una CONCENTRAZIONE di CROMO pari a 50 è

$$\hat{Y} = 3.858 + 1.3925 \cdot 50 = 73.483$$