Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione

V appello di Statistica per Ingegneria Fisica
21 Febbraio 2019

Cognome, Nome e Numero di matricola:

Problema 1. Nel prossimo laboratorio di Fisica, si studierà il moto di un grave lungo un piano inclinato. L'esperimento in programma è questo: misurando il tempo t (in secondi) impiegato da una sfera d'acciaio per percorrere un piano di lunghezza e inclinazione note, si vuole stimare l'accelerazione di gravità g.

Aldo e Bruno, che faranno l'esperimento insieme, hanno già deciso come ripartirsi il lavoro: Aldo prenderà 10 tempi di rotolamento $X_1, \dots, X_{10}$ e ne registrerà la media campionaria $\bar{X}$. Bruno, invece, prenderà solo altri 5 tempi $Y_1, \dots, Y_5$, ne registrerà anch'egli la media campionaria $\bar{Y}$, e alla fine si occuperà dell'analisi dei dati suoi e di Aldo.

Si può assumere che le misure di Aldo e di Bruno costituiscano due campioni gaussiani indipendenti, con $X_i \sim N(t, \sigma^2)$ e $Y_j \sim N(t, \sigma^2)$, in cui t è il tempo incognito che la sfera impiegherebbe idealmente per percorrere il piano, mentre $\sigma = 0.1\,\text{sec}$ è la precisione nota del cronometro usato da entrambi.

Come stimatore del parametro t, Bruno vorrebbe usare una combinazione lineare del tipo

$$T_a = a\bar{X} + (1-a)\bar{Y}\,,$$

in cui $a \in \mathbb{R}$ è un'opportuna costante reale che però non sa come scegliere. Aiutatelo voi!

(a) Calcolate, in funzione di a, la distorsione e l'errore quadratico medio dello stimatore T_a.

(b) In base a quanto trovato nel punto precedente, qual è il valore di a che rende T_a migliore? Perché?

D'ora in poi, indicate con T lo stimatore migliore che avete trovato al punto precedente. *(Se non ci siete riusciti, continuate con lo stimatore $T = \frac{1}{2}(\bar{X}+\bar{Y})$, che è quello scelto da Bruno fissando la costante a del tutto a caso, e che perciò non è detto sia il migliore.)*

(c) Determinate la densità di probabilità dello stimatore che avete scelto.

(d) Calcolate la probabilità $\mathbb{P}(|T-t| < 0.03\,\text{sec})$, cioè la probabilità che lo stimatore scelto si discosti dal vero valore del parametro t per meno di 0.03 secondi.

(e) Col piano inclinato usato nell'esperimento, l'accelerazione di gravità è legata al tempo t dalla relazione

$$g = \frac{114.7\,\text{m}}{t^2} \qquad (\text{m} = \text{metri è l'unità di misura della costante a numeratore})\,.$$

Proponete uno stimatore almeno approssimativamente non distorto per il parametro incognito g.

(f) Dalle loro misure, Aldo e Bruno hanno rispettivamente ricavato le medie campionarie $\bar{x} = 3.25\,\text{sec}$ e $\bar{y} = 3.79\,\text{sec}$. Con questi dati, fornite una stima di t e di g.

Risultati.

(a) Abbiamo

$$\mathbb{E}[T_a] = \mathbb{E}\left[a\bar{X} + (1-a)\bar{Y}\right] \underset{\substack{\text{linearità}\\ \text{di } \mathbb{E}}}{=} a\mathbb{E}\left[\bar{X}\right] + (1-a)\mathbb{E}\left[\bar{Y}\right] = at + (1-a)t = t$$

$$\text{Var}(T_a) = \text{Var}(a\bar{X} + (1-a)\bar{Y}) \underset{\substack{\text{indipendenza}\\ \text{di } \bar{X},\,\bar{Y}}}{=} \text{Var}(a\bar{X}) + \text{Var}((1-a)\bar{Y})$$

$$\underset{\substack{\text{quadraticità}\\ \text{di Var}}}{=} a^2\text{Var}(\bar{X}) + (1-a)^2\text{Var}(\bar{Y}) = a^2\frac{\sigma^2}{10} + (1-a)^2\frac{\sigma^2}{5} = \frac{\sigma^2}{10}\left[a^2 + 2(1-a)^2\right]$$

e quindi

$$\begin{aligned}\text{bias}(T_a;t) &= \mathbb{E}\left[T_a\right] - t = 0\\ \text{MSE}(T_a;t) &= \text{Var}(T_a) + \text{bias}(T_a;t)^2 = \frac{\sigma^2}{10}\left[a^2 + 2(1-a)^2\right]\,.\end{aligned}$$

(b) Tutti gli stimatori T_a sono non distorti, indipendentemente dal valore di a. Lo stimatore migliore è dunque quello per cui è minimo l'errore quadratico medio. Minimizzando la funzione $f(a) = a^2 + 2(1-a)^2$ rispetto alla variabile a, vediamo subito che $\text{MSE}(T_a;t)$ è minimizzato quando $a = 2/3$. Si noti che per tale valore si ha

$$T_{2/3} = \frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y} = \frac{2}{3}\frac{X_1 + \ldots + X_{10}}{10} + \frac{1}{3}\frac{Y_1 + \ldots + Y_5}{5} = \frac{X_1 + \ldots + X_{10} + Y_1 + \ldots + Y_5}{15}\,,$$

cioè $T_{2/3}$ coincide con la media campionaria delle 15 misure complessive.

(c) Se scegliamo $T = T_{2/3}$, abbiamo già visto che si tratta della media campionaria di un campione di 15 misure gaussiane con media t e varianza σ^2. Di conseguenza,

$$T \sim N\left(t\,,\,\frac{\sigma^2}{15}\right)\,.$$

Se invece scegliamo lo stimatore di Bruno $T = (\bar{X} + \bar{Y})/2 = T_{1/2}$, allora sappiamo che

$$\bar{X} \sim N\left(t, \frac{\sigma^2}{10}\right)\,, \qquad \bar{Y} \sim N\left(t, \frac{\sigma^2}{5}\right)\,, \qquad \bar{X},\,\bar{Y} \text{ indipendenti}\,.$$

Di conseguenza, $T_{1/2}$ è una v.a. gaussiana in quanto combinazione lineare di v.a. gaussiane indipendenti. Più precisamente, la sua densità è

$$T_{1/2} \sim N\left(\mathbb{E}\left[T_{1/2}\right], \text{Var}(T_{1/2})\right) = N\left(t\,,\,\frac{\sigma^2}{10}\left[\left(\frac{1}{2}\right)^2 + 2\left(1-\frac{1}{2}\right)^2\right]\right) = N\left(t\,,\,\frac{3\sigma^2}{40}\right)\,,$$

in cui abbiamo usato $\mathbb{E}\left[T_a\right]$ e $\text{Var}(T_a)$ determinati al punto (a).

(d) Se scegliamo $T = T_{2/3} \sim N(t\,,\,\sigma^2/15)$, allora

$$\begin{aligned}\mathbb{P}\left(|T-t| < 0.03\,\text{sec}\right) &= \mathbb{P}\left(-0.03 < T - t < 0.03\right) = \mathbb{P}\left(\frac{-0.03}{\sqrt{\frac{\sigma^2}{15}}} < \underbrace{\frac{T - \mathbb{E}\left[T\right]}{\sqrt{\text{Var}(T)}}}_{\sim N(0,1)} < \frac{0.03}{\sqrt{\frac{\sigma^2}{15}}}\right)\\ &= \Phi\left(\frac{0.03\sqrt{15}}{\sigma}\right) - \Phi\left(-\frac{0.03\sqrt{15}}{\sigma}\right) = 2\,\Phi(1.162) - 1 \simeq 2\cdot 0.87698 - 1\\ &= 75.396\%\,.\end{aligned}$$

Se invece facciamo la scelta di Bruno, cioè $T = T_{1/2} \sim N(t\,,\,3\sigma^2/40)$, allora, con passaggi simili,

$$\begin{aligned}\mathbb{P}\left(|T-t| < 0.03\,\text{sec}\right) &= \mathbb{P}\left(\frac{-0.03}{\sqrt{\frac{3\sigma^2}{40}}} < \underbrace{\frac{T - \mathbb{E}\left[T\right]}{\sqrt{\text{Var}(T)}}}_{\sim N(0,1)} < \frac{0.03}{\sqrt{\frac{3\sigma^2}{40}}}\right)\\ &= \Phi\left(\frac{0.03\sqrt{40}}{\sqrt{3}\sigma}\right) - \Phi\left(-\frac{0.03\sqrt{40}}{\sqrt{3}\sigma}\right) = 2\,\Phi(1.095) - 1 \simeq 2\cdot\frac{0.86214 + 0.86433}{2} - 1\\ &= 72.647\%\,.\end{aligned}$$

Si noti che la seconda probabilità è minore della prima; in altre parole, usando lo stimatore di Bruno, è meno probabile trovare valori prossimi a t di quanto lo sia usando lo stimatore ottimale trovato al punto (b).

(e) Comunque si scelga a, essendo T_a uno stimatore esattamente non distorto di t, per il metodo delta uno stimatore approssimativamente non distorto di g sarà

$$G_a = \frac{114.7}{T_a^2}\,.$$

Naturalmente, la scelta di a migliore è quella fatta al punto (b), cioè $a = 2/3$. Infatti, minimizzando l'MSE di T_a, tale scelta minimizza anche

$$\begin{aligned}\mathrm{MSE}(G_a; g) &= \mathrm{Var}(G_a) + \mathrm{bias}(G_a; g)^2 \underset{\delta}{\simeq} \mathrm{Var}(G_a) \underset{\delta}{\simeq} \left[\frac{\mathrm{d}g}{\mathrm{d}t}(\mathbb{E}\,[T_a])\right]^2 \mathrm{Var}(T_a) \\ &= \left(-\frac{2 \cdot 114.7}{t^3}\right)^2 \mathrm{MSE}(T_a; t)\,.\end{aligned}$$

(f) Usando lo stimatore $T_{2/3}$ di t e lo stimatore $G_{2/3}$ di g in accordo con quanto trovato ai punti (b)-(e), troviamo le stime

$$\begin{aligned}t_{2/3} &= \frac{2}{3}\bar{x} + \frac{1}{3}\bar{y} = \frac{2}{3}3.25 + \frac{1}{3}3.79 = 3.43\,\mathrm{sec} \\ g_{2/3} &= \frac{114.7}{t_{2/3}^2} = \frac{114.7}{3.43^2} = 9.749\,\mathrm{m/sec}^2\,.\end{aligned}$$

Se invece avessimo scelto lo stimatore di Bruno, avremmo trovato

$$\begin{aligned}t_{1/2} &= \frac{1}{2}\bar{x} + \frac{1}{2}\bar{y} = \frac{1}{2}3.25 + \frac{1}{2}3.79 = 3.52\,\mathrm{sec} \\ g_{1/2} &= \frac{114.7}{t_{1/2}^2} = \frac{114.7}{3.52^2} = 9.257\,\mathrm{m/sec}^2\,.\end{aligned}$$

Problema 2. Un errore tipico di chi studia Statistica è fraintendere (o non studiare proprio) il Teorema del Limite Centrale (TLC). Il prof. T. – che insegna Statistica – lo sa bene. In particolare, conosce dall'esperienza degli anni passati che mediamente il 20% degli studenti è impreparato sull'argomento. Nel tentativo di migliorare la situazione, da quest'anno il professore ha quindi deciso di raddoppiare il tempo dedicato al TLC nel programma del suo corso. Curioso di capire se l'idea abbia funzionato, nel primo appello d'esame dopo la modifica del programma ha aggiunto di proposito una domanda sul teorema. Questo è il risultato: sul totale di 74 compiti consegnati, 25 compiti contengono una risposta errata alla domanda in questione.

(a) In base ai dati ottenuti, si può affermare che il nuovo programma ha effettivamente diminuito la probabilità di trovare studenti impreparati sul TLC? Impostate un opportuno test e traetene una conclusione al livello di significatività del 5%.

(b) La prof.ssa S., che tiene la sezione del corso parallela al prof. T., sostiene che i dati dimostrano invece esattamente il contrario di quanto sperato dal professore. In altre parole, secondo lei dedicare più tempo al TLC in realtà non fa altro che peggiorare sensibilmente la preparazione degli studenti. I dati danno ragione alla prof.ssa S.? Impostate un nuovo test e traetene una conclusione ancora al livello di significatività del 5%.

(c) C'è differenza tra il tipo di conclusione ottenuta nel test del punto (a) e quella del punto (b)? Se sì, quale?

(d) Nella sezione della prof.ssa S., al primo appello d'esame di quest'anno hanno consegnato il compito 82 studenti, e di questi 33 hanno risposto in modo errato alla domanda sul TLC. Verificate con un opportuno test se gli studenti delle due sezioni manifestano un livello di preparazione sul TLC significativamente diverso. Calcolate il p-value del test, e traetene una conclusione.

(e) Ora enunciate voi il TLC in modo corretto.

Risultati.

(a) Bisogna scegliere tra le ipotesi statistiche

$$H_0 : p = 0.20 =: p_0 \qquad \text{contro} \qquad H_1 : p < p_0 \tag{$*$}$$

La regola di un test di livello α è

$$\text{“ rifiuto } H_0 \text{ se } Z_0 := \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} < -z_{1-\alpha}\text{ ”} \tag{$*$}$$

Coi dati, $\bar{x} = \frac{25}{74} = 0.337838$ e quindi

$$z_0 = \frac{0.337838 - 0.20}{\sqrt{0.20 \cdot (1-0.20)}}\sqrt{74} = 2.964\,,$$

mentre al livello $\alpha = 5\%$

$$z_{1-\alpha} = z_{1-0.05} = z_{0.95} = 1.645\,.$$

Dal momento che (palesemente!) $z_0 = 2.964 \not< -z_{0.95} = -1.645$, non si può rifiutare H_0 al 5%. Si poteva arrivare subito a questa conclusione, osservando che già dai dati risulta $\bar{x} > p_0$, e dunque il p-value del test per le ipotesi $(*)$ è senz'altro maggiore del 50%.

(b) Ora le ipotesi statistiche $(*)$ sono scambiate:

$$H_0 : p = 0.20 =: p_0 \qquad \text{contro} \qquad H_1 : p > p_0 \tag{$**$}$$

Regola:

$$\text{“ rifiuto } H_0 \text{ se } Z_0 := \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} > z_{1-\alpha}\text{ ”}$$

La realizzazione della statistica test è la stessa $z_0 = 2.964$ di prima (i dati sono gli stessi), ma questa volta $z_0 > z_{0.95} \Rightarrow$ rifiuto H_0.

(c) La differenza è che la conclusione del test del punto (a) è debole, mentre quella del test del punto (b) è forte.

(d) Si tratta di fare un test per le ipotesi

$$H_0 : p_T = p_S \qquad \text{contro} \qquad H_1 : p_T \neq p_S \,,$$

dove p_T è la probabilità che uno studente a caso del corso del prof. T. risponda erroneamente alla domanda sul TLC, e p_S è la probabilità analoga per uno studente del corso della prof.ssa S.. Bisogna quindi usare un test per la differenza delle frequenze di due campioni bernoulliani numerosi. La regola di un test al livello α è

$$\text{“ rifiuto } H_0 \text{ se } |Z_0| := \left| \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n} + \frac{1}{m}\right)}} \right| > z_{1-\frac{\alpha}{2}} \text{ ”} \qquad (\circ)$$

dove $\bar{X}$ e $\bar{Y}$ sono le frequenze campionarie dei due campioni, m e n le rispettive numerosità, e

$$\hat{P} = \frac{m\bar{X} + n\bar{Y}}{m+n} \,.$$

Coi dati del prof. T. e della prof.ssa S.,

$$\bar{x} = \frac{25}{74} = 0.337838 \qquad \bar{y} = \frac{33}{82} = 0.402439 \qquad \hat{p} = \frac{25+33}{74+82} = 0.371795$$

$$z_0 = \frac{0.337838 - 0.402439}{\sqrt{0.371795 \cdot (1 - 0.371795) \cdot \left(\frac{1}{25} + \frac{1}{82}\right)}} = -0.83368 \,.$$

In base alla regola $(\circ)$, il p-value dei dati si ottiene risolvendo rispetto ad α l'equazione

$$\begin{aligned} |z_0| = z_{1-\frac{\alpha}{2}} \quad &\Leftrightarrow \quad \Phi(|z_0|) = \Phi\Big(z_{1-\frac{\alpha}{2}}\Big) - 1 - \frac{\alpha}{2} \\ &\Leftrightarrow \quad \alpha = 2[1 - \Phi(|z_0|)] = 2[1 - \Phi(0.83368)] = 2(1 - 0.79673) = 0.40654 \,. \end{aligned}$$

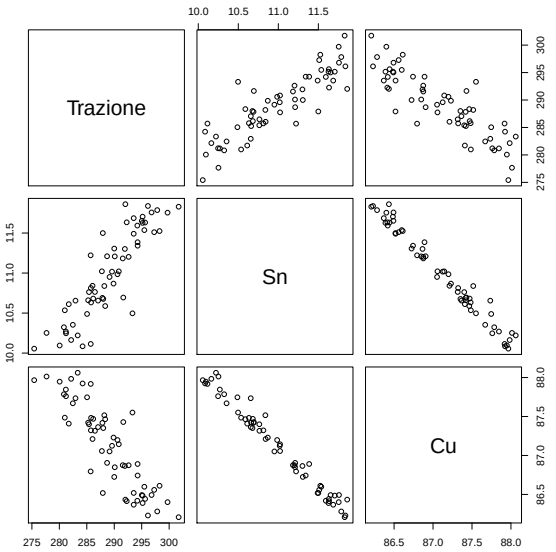
Perciò, p-value = 40.654%. Con un valore cosí alto, non possiamo rifiutare H_0 a nessun livello di significatività ragionevole. Ne concludiamo che non c'è nessuna evidenza che il livello di preparazione sul TLC sia significativamente diverso nelle due sezioni.

(e) Sia $X_1, \ldots, X_n$ un campione aleatorio (= n variabili aleatorie indipendenti e identicamente distribuite), e sia $\bar{X}_n = (X_1 + \ldots + X_n)/n$ la sua media campionaria. Allora, per ogni $z \in \mathbb{R}$, si ha

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} < z \right) = \Phi(z) \,,$$

dove $\mu = \mathbb{E}[X_i]$, $\sigma = \sqrt{\mathrm{Var}(X_i)}$, e Φ è la funzione di ripartizione della normale standard.

Problema 3. Il bronzo è una lega formata principalmente da stagno (Sn) e da rame (Cu). Siamo interessati a studiare la resistenza alla trazione del bronzo, in dipendenza dalla percentuale di stagno e di rame che lo compongono. Nella figura sottostante sono riportati gli scatterplot tra le variabili considerate.



In Figura 1 sono invece riportati gli output di tre diversi modelli empirici lineari gaussiani, che hanno come variabile risposta la resistenza alla trazione, e come regressori la percentuale di stagno e di rame. Sempre in Figura 1 sono riportati gli scatterplot dei residui dei tre modelli. I p-value dello Shapiro-test sui residui sono:

$$p_{\text{completo}} = 0.3603, \quad p_{\text{semplice Sn}} = 0.2648, \quad p_{\text{semplice Cu}} = 0.8248.$$

(a) Scrivere la relazione tra le variabili ipotizzata dai tre modelli.

(b) I tre modelli sono globalmente significativi?

(c) Quale problema presenta il modello completo? A cosa è dovuto?

(d) Quale tra i due modelli ridotti spiega meglio la variabilità della risposta?

(e) Quali sono le ipotesi alla base dei modelli considerati? Per quali modelli sono verificate?

(f) Dopo aver scelto il modello migliore tra i tre proposti, fornire un intervallo di confidenza al 95% per l'intercetta.

Risultati.

(a)
- modello completo: $Y = \beta_0 + \beta_1 \,\text{Sn} + \beta_2 \,\text{Cu} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$
- modello ridotto 1: $Y = \beta_0 + \beta_1 \,\text{Sn} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$
- modello ridotto 2: $Y = \beta_0 + \beta_1 \,\text{Cu} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$

(b) Per rispondere alla domanda occorre fare un test F sui coefficienti dei regressori. In particolare, per quanto riguarda il modello completo, il test ha un p-value minore di $2.2 \cdot 10^{-16}$, di conseguenza il modello è significativo. Per quanto riguarda i modelli ridotti, il test corrisponde al test sul singolo regressore. Per entrambi i modelli si ha un p-value minore di $2.2 \cdot 10^{-16}$ pertanto entrambi sono globalmente significativi.

(c) Il modello completo, benché sia globalmente significativo, presenta entrambi i regressori singolarmente non significativi. La causa del problema è la collinearità dei regressori, come si può vedere dallo scatterplot tra Sn e Cu.

(d) Il modello semplice con regressore Sn ha un R^2 leggermente più alto, per cui è il migliore in termini di di variabilità spiegata. Entrambi i modelli sono comunque piuttosto buoni, avendo un R^2 abbastanza alto.

(e) Le ipotesi alla base del modello sono che i residui siano normali e omoschedastici. L'ipotesi di omoschedasticità è soddisfatta per tutti i modelli, come si può vedere dagli scatterplot dei residui che non presentano particolari pattern. Anche l'ipotesi di normalità è verificata per tutti e tre i modelli dato l'alto valore del p-value dello shapiro-test.

(f) Scegliamo come modello migliore il modello semplice con regressore Sn, dato che ha un R^2 migliore dell'altro modello semplice, e che il modello completo non ha tutti i regressori significativi. L'intervallo richiesto è:

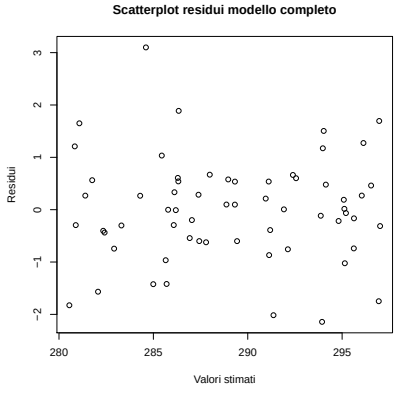
$$\begin{aligned} IC(\beta_0) &= \left(\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\,\mathrm{se}(\hat{\beta}_0)\,,\ \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\,\mathrm{se}(\hat{\beta}_0)\right) \\ &= (187.3012 - 2 \cdot 7.2081, 187.3012 + 2 \cdot 7.2081) = (172.8829, 201.7195) \end{aligned}$$

```
Call:
lm(formula = Trazione ~ Sn + Cu)

Residuals:
   Min     1Q Median     3Q    Max
-6.022 -1.646  0.007  1.517  8.713

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  308.412    390.612   0.790   0.4329
Sn             8.034      4.020   1.999   0.0503 .
Cu            -1.235      3.983  -0.310   0.7576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.81 on 59 degrees of freedom
Multiple R-squared:  0.7696,    Adjusted R-squared:  0.7618
F-statistic: 98.53 on 2 and 59 DF,  p-value: < 2.2e-16
```

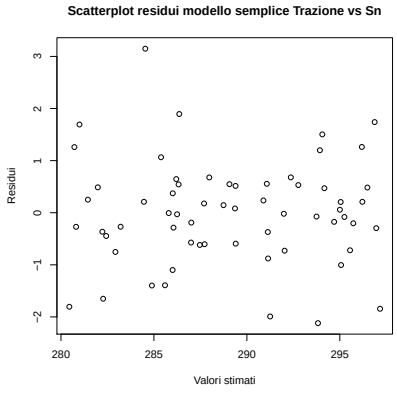


```
Call:
lm(formula = Trazione ~ Sn)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9110 -1.6401 -0.0363  1.4697  8.7814

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 187.3012     7.2081   25.98   <2e-16 ***
Sn            9.2633     0.6551   14.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.789 on 60 degrees of freedom
Multiple R-squared:  0.7692,    Adjusted R-squared:  0.7654
F-statistic:   200 on 1 and 60 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = Trazione ~ Cu)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6211 -1.5843  0.2261  1.7886  8.1678

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1080.7345    58.3782   18.51   <2e-16 ***
Cu            -9.0871     0.6701  -13.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.88 on 60 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.7499
F-statistic: 183.9 on 1 and 60 DF,  p-value: < 2.2e-16
```

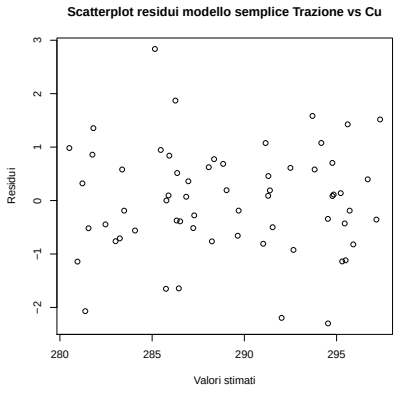


Figura 1: Summary dei modelli