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

IV Appello di Statistica per Ingegneria Energetica
26 settembre 2016

Cognome, Nome e Numero di matricola:

Problema 1. Pierino frequenta la seconda media e tutte le mattine per andare a scuola prende l'autobus. Si sa che, ogni giorno, la probabilità che lo perda è pari al 15%, indipendentemente da quello che è successo gli altri giorni.

(a) Indicare la legge della variabile aleatoria X_n = numero di assenze di Pierino da scuola su n giorni.

(b) Calcolare la probabilità che su 5 giorni Pierino non faccia assenze.

(c) Calcolare la probabilità che su 5 giorni Pierino faccia esattamente un giorno di assenza.

Per essere promosso in terza Pierino non può fare più di 30 giorni di assenze su i 200 giorni di scuola previsti dal calendario scolastico.

(d) Qual è la probabilità che Pierino venga automaticamente bocciato per le sue assenze?

La mamma di Pierino gli promette una bici nuova nel caso faccia almeno 180 giorni di scuola.

(e) Qual è la probabilità che Pierino riceva questo regalo?

Risultati.

(a) $X_n \sim B(n, 0.15)$.

(b) $P(X_5 = 0) = (1-p)^5 = 0.85^5 = 0.4437$.

(c) $P(X_5 = 1) = 5p(1-p)^4 = 5 \cdot 0.15 \cdot 0.85^4 = 0.3915$.

(d) Poichè $np = 200 \cdot 0.15 = 30 > 5$ e $n(1-p) = 170 > 5$, possiamo approssimare X_{200} ad una normale di media $np = 30$ e varianza $np(1-p) = 200 \cdot 0.15 \cdot 0.85 = 25.5$; quindi

$$P(X_{200} > 30) = P(X_{200} > 30.5) = P\left(Z > \frac{30.5 - 30}{\sqrt{25.5}}\right) = P(Z > 0.0990) = 1 - \Phi(0.10) = 0.4602$$

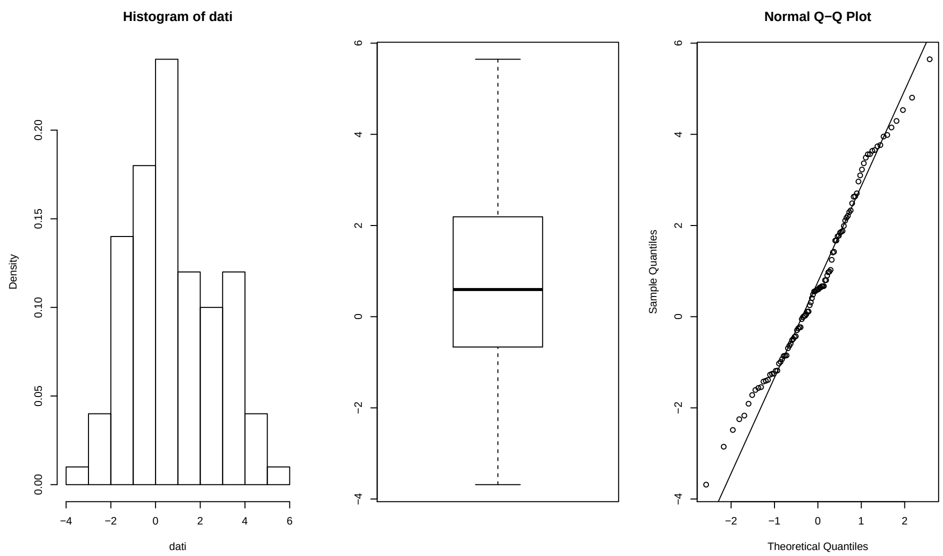
(e) $P(X_{200} \leq 20) = P\left(Z < \frac{20.5-30}{\sqrt{25.5}}\right) = P(Z < -1.88) = P(Z > 1.88) = 1 - P(Z < 1.88) = 0.0301.$

Problema 2. Rambos è alla ricerca di una specie adatta ai suoi esperimenti di potenziamento. Serve trovare una specie i cui individui abbiano un livello di *sotang* abbastanza omogeneo, ovvero, più precisamente, una specie con una varianza del livello di *sotang* inferiore a 3.5. A tal fine ha fatto misurare il livello di *sotang* di 100 waldaster casualmente selezionati. Si riportano i valori di media e deviazione standard campionarie, p-value del test di Shapiro-Wilk e i quartili dei dati raccolti:

$$\bar{x}_{100} = 0.8043 \qquad s_{100} = 1.9502 \qquad SW_{100} = 0.1836$$

$$Q_1 = -0.6524 \qquad Q_2 = 0.5952 \qquad Q_3 = 2.1810$$

Si riportano inoltre l'istogramma dei dati unitamente al boxplot e al Normal Probability Plot.



(a) Fornire una stima puntuale della varianza σ^2 del livello di *sotang* fra i guerrieri waldaster.

(b) Quali condizioni servono per poter inferire su σ^2 anche con intervalli di confidenza e test d'ipotesi? Tali condizioni sono valide per il caso in esame? Perché?

(c) Fornire una stima intervallare al 95% di σ^2.

(d) Si imposti un opportuno test per capire se i dati raccolti forniscono una forte evidenza statistica del fatto che i waldaster sono adatti agli esperimenti di potenziamento. Specificare ipotesi nulla e alternativa, statistica test e regione critica di rifiuto.

(e) Calcolare il p-value del test appena introdotto e decidere di conseguenza.

Risultati.

(a) $\widehat{\sigma}^2 = s_{100}^2 = 3.80$.

(b) L'inferenza su σ^2 tramite gli usuali intervalli di confidenza e test d'ipotesi presuppone (oltre alla causalità del campione) la normalità della popolazione campionata. Questa ipotesi non è confutata dai dati raccolti, come mostrano istogramma, normal probability plot e dal p-value del test di Shapiro-Wilk. In particolare il p-value del test di Shapiro-Wilk non permette di rifiutare l'ipotesi di normalità agli usuali livelli di significatività.

(c) La stima intervallare al 95% sarà data da:

$$(n-1)\frac{s_{100}^2}{\chi^2(0.05/2, n-1)} = 99 \cdot \frac{3.8034}{128.422} = 2.93$$

$$(n-1)\frac{s_{100}^2}{\chi^2(1-0.05/2, n-1)} = 99 \cdot \frac{3.8034}{73.36108} = 5.14$$

(d) Impostiamo un test unilatero sulla varianza σ^2 con soglia di confronto $\sigma_0^2 = 3.5$:

$$H_0 : \sigma^2 \geq \sigma_0^2 \qquad H_1 : \sigma^2 < \sigma_0^2$$

La statistica test è:

$$w_0 = 99 \cdot \frac{s_{100}^2}{\sigma_0^2}$$

La regione critica di livello α è:

$$RC : w_0 < \chi^2(1-\alpha, 99)$$

(e) Il p-value α dei dati per il test introdotto è dato da

$$w_0 = 107.5818 = \chi^2(1-\alpha, 99) \Rightarrow \alpha = 0.74$$

mentre usando le tavole si trova

$$w_0 = 107.5818 = \chi^2(1-\alpha, 99)$$

$$107.5818 \simeq \chi^2(1-\alpha, 100)$$

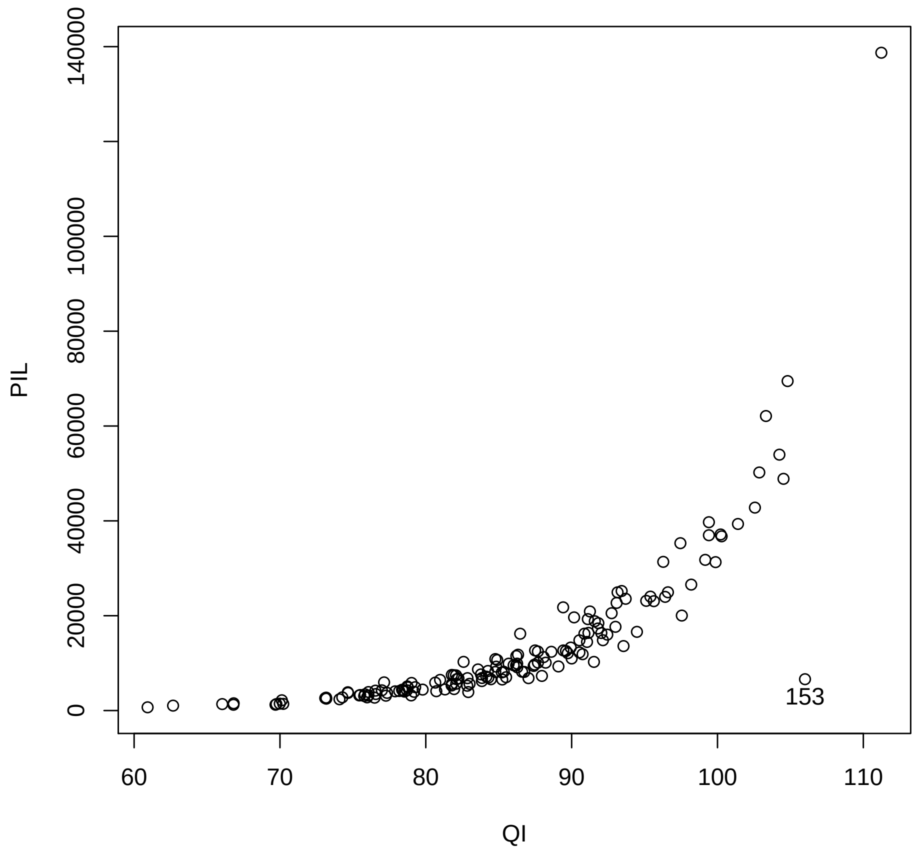
$$\chi^2(0.5, 100) = 99.33 < 107.5818 < 118.50 = \chi^2(0.1, 100)$$

$$0.1 < 1-\alpha < 0.5$$

$$0.5 < \alpha < 0.9$$

Non possiamo quindi rifiutare l'ipotesi nulla: i dati raccolti non forniscono una forte evidenza statistica del fatto che i waldaster siano adatti agli esperimenti di potenziamento.

Problema 3. Uno studio inglese pubblicato nel 2002 afferma che esista una correlazione tra il Prodotto Interno Lordo (PIL) pro capite di uno Stato e il suo Quoziente Intellettivo medio (QI). Per sostenere questa ipotesi e capire meglio la relazione fra le due variabili, sono stati raccolti i dati di 153 Paesi. Qui sotto è riportato lo scatterplot dei dati raccolti:



Con i dati raccolti sono stati elaborati 3 diversi modelli di regressione lineare. Nel primo modello si ipotizza una relazione lineare semplice tra PIL e QI, nel secondo e nel terzo si ipotizza un modello di tipo logaritmico, con la differenza che nell'ultimo è stato rimosso un dato (il caso 153), col sospetto che fosse stato trascritto in modo errato.

Di seguito sono riportati i p-value dei test di Shapiro-Wilk condotti sui relativi residui, gli scatterplot dei residui standardizzati e i valori delle stime e degli indici principali.

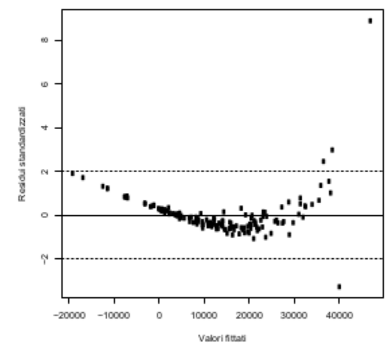
$$sw_1 < 2.2 \times 10^{-16} \quad sw_2 = 6.2 \times 10^{-15} \quad sw_3 = 0.6917$$

```
Call:
lm(formula = PIL ~ QI, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-33679  -4717  -1883   2523  91541

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -99091.41    7707.66  -12.86   <2e-16 ***
QI            1314.96      89.53   14.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10310 on 151 degrees of freedom
Multiple R-squared:  0.5883,    Adjusted R-squared:  0.5855
F-statistic: 215.7 on 1 and 151 DF,  p-value: < 2.2e-16
```

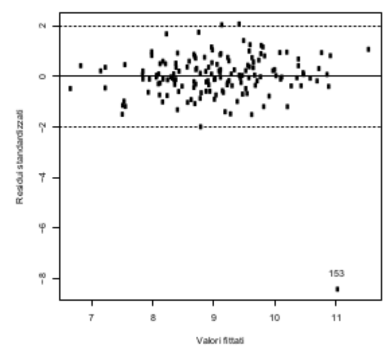



```
Call:
lm(formula = log(PIL) ~ QI, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24545 -0.09750  0.01275  0.13594  0.55712

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74228    0.19969   3.717 0.000283 ***
QI           0.09717    0.00232  41.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.267 on 151 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9203
F-statistic:  1755 on 1 and 151 DF,  p-value: < 2.2e-16
```

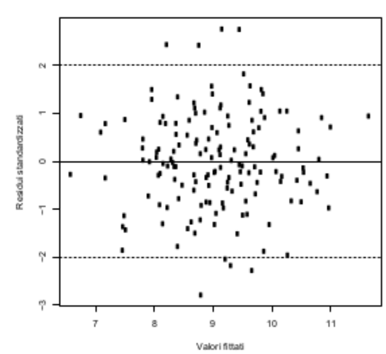



```
Call:
lm(formula = log(PIL) ~ QI, data = data, subset = -153)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53474 -0.11621 -0.00622  0.13601  0.52968

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.449803   0.145601   3.089  0.00239 **
QI          0.100769   0.001694  59.483  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1919 on 150 degrees of freedom
Multiple R-squared:  0.9593,    Adjusted R-squared:  0.9591
F-statistic:  3538 on 1 and 150 DF,  p-value: < 2.2e-16
```



(a) Scrivere la relazione tra PIL e QI ipotizzata dai tre modelli di regressione empirici gaussiani.

(b) Trovare la varianza del PIL degli Stati a QI fissato in funzione di tale valore di QI e dei parametri del modello. Trovare il valore esatto per il primo modello, e il valore approssimato fornito dal metodo delta per gli altri due modelli.

(c) Commentare i grafici dei residui standardizzati per ciascuno dei modelli, specificando quali presentano residui omoschedastici.

(d) Indicare, motivando la risposta, per quali modelli è soddisfatta l'ipotesi gaussiana.

Accertato l'errore del dato 153 vengono scartati primi due modelli e, per i 152 dati rimasti, si ottiene

$$\overline{\mathrm{QI}} = 85.45, \qquad s^2_{\mathrm{QI}} = 84.98.$$

(e) Fornire una previsione puntuale per il PIL di un Paese con QI uguale a 100.

(f) Fornire una previsione intervallare bilatera al 95% per il PIL di un Paese con QI uguale a 100.

Risultati.

(a) Indicando con $\epsilon \sim N(0, \sigma^2)$ l'errore gaussiano presente in ciascun modello, possiamo scrivere:

$$\begin{aligned}
\text{modello 1:} \quad & \mathrm{PIL} = \beta_0 + \beta_1 \mathrm{QI} + \epsilon \\
\text{modello 2-3:} \quad & \log(\mathrm{PIL}) = \beta_0 + \beta_1 \mathrm{QI} + \epsilon \iff \mathrm{PIL} = \exp\left\{\beta_0 + \beta_1 \mathrm{QI} + \epsilon\right\}
\end{aligned}$$

(b) Modello 1: $\mathrm{Var}(\mathrm{PIL}) = \sigma^2$.
Modello 2: $\mathrm{Var}(\mathrm{PIL}) \simeq \mathrm{e}^{2(\beta_0 + \beta_1 \mathrm{QI})}\, \sigma^2$.

(c) Il modello 3 presenta residui omoschedastici. Non si evidenziano pattern particolari, al contrario del modello 1 dove si nota chiaramente un andamento anomalo dei residui e del modello 2 dove abbiamo una nuvola di punti nella fascia tra -2 e 2 deviazioni standard dallo zero, con un punto (il caso 153) estremamente più lontano.

(d) Solo il modello 3 soddisfa l'ipotesi gaussiana sugli errori, come evidenziato al punto (c) e come riportato dai test di Shapiro-Wilk: i p-value dei primi due modelli sono quasi nulli, mentre nel modello 3 abbiamo un valore piuttosto alto (0.6917).

(e)

$$\widehat{\log(\mathrm{PIL})} = \widehat{\beta_0} + \widehat{\beta_1}\mathrm{QI}_0 = 0.4498 + 0.1008 \cdot 100 = 10.5298$$

quindi, ritornando alla variabile originale, otteniamo

$$\widehat{\mathrm{PIL}} = \exp(10.5298) = 37413.99$$

(f) Per calcolare l'intervallo di previsione richiesto, occorre applicare la formula seguente:

$$\widehat{\beta_0} + \widehat{\beta_1} \cdot \mathrm{QI}_0 \pm t_{0.025, n-2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\overline{\mathrm{QI}} - \mathrm{QI}_0)^2}{S_{\mathrm{QIQI}}}}$$

dove $\widehat{\beta_0} = 0.4498$, $\widehat{\beta_1} = 0.1008$, $\mathrm{QI}_0 = 100$, $t_{0.025,150} \approx 1.96$, $\hat{\sigma} = 0.1919$, $n = 152$, $\overline{\mathrm{QI}} = 85.45$, $S_{\mathrm{QIQI}} = s^2_{\mathrm{QI}} \cdot 151 = 84.98 \cdot 151 = 12831.98$. Il risultato per il logaritmo del PIL è:

$$[10.14\,,\ 10.91]$$

quindi, ritornando alla variabile originale, otteniamo

$$[25336.47\,,\ 54720.85]$$