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

Seconda prova in itinere di Statistica per Ingegneria Energetica
8 luglio 2014

Cognome, Nome e Numero di matricola:

Problema 1. Per una generica partita del campionato mondiale di calcio 2014, consideriamo il numero di persone X che acquista in anticipo il biglietto di ingresso, ma che per un imprevisto dell'ultimo minuto non può recarsi allo stadio. Supponiamo che la distribuzione di X sia ben approssimabile con una Poisson, di media λ incognita, e che su n partite i rispettivi numeri $X_1, \dots, X_n$ siano indipendenti.
Vogliamo innanzitutto stimare $\lambda = \mathbb{E}[X]$ con lo stimatore puntuale

$$\widehat{\lambda} = \overline{X}_n.$$

1. Calcolare distorsione, errore quadratico medio, errore standard e distribuzione asintotica di $\widehat{\lambda}$.

Vogliamo ora stimare $\sigma = \sqrt{\mathrm{Var}(X)}$ con lo stimatore puntuale

$$\widehat{\sigma} = \sqrt{\widehat{\lambda}} = \sqrt{\overline{X}_n}.$$

2. Utilizzando il metodo delta, calcolare in modo approssimato distorsione, errore quadratico medio ed errore standard di $\widehat{\sigma}$.

Vogliamo infine una stima intervallare di λ, basata su $\widehat{\lambda}$. Scopriamo che, per n sufficientemente grande,

$$\frac{\widehat{\lambda} - \lambda}{\sqrt{\frac{\widehat{\lambda}}{n}}} \simeq N(0,1).$$

3. Proporre un intervallo di confidenza di livello γ per λ.

Per le prime 10 partite si registrano i seguenti valori di X:

$$23,\ 18,\ 20,\ 26,\ 20,\ 23,\ 12,\ 20,\ 16,\ 35.$$

4. Fornire una stima puntuale di λ e di σ.
5. Stimare i relativi errori standard.
6. Fornire una stima intervallare di λ ad un livello di confidenza $\gamma = 0.95$.

Risultati.

1. $\mathbb{E}[\widehat{\lambda}] - \lambda = 0, \qquad \text{MSE}(\widehat{\lambda}) = \text{Var}(\widehat{\lambda}) = \lambda/n, \qquad \text{se}(\widehat{\lambda}) = \sqrt{\text{Var}(\widehat{\lambda})} = \sqrt{\lambda/n}, \qquad \widehat{\lambda} \simeq N\Big(\lambda, \frac{\lambda}{n}\Big).$
2. $\mathbb{E}[\widehat{\sigma}] - \sigma = \mathbb{E}\Big[\sqrt{\widehat{\lambda}}\Big] - \lambda \simeq \sqrt{\mathbb{E}[\widehat{\lambda}]} - \lambda = 0, \qquad \text{MSE}(\widehat{\sigma}) \simeq \text{Var}(\widehat{\sigma}) \simeq \frac{\lambda}{n}\left(\frac{1}{2\sqrt{\lambda}}\right)^2 = \frac{1}{4n},$
   $\text{se}(\widehat{\sigma}) = \sqrt{\text{Var}(\widehat{\sigma})} \simeq \sqrt{\frac{1}{4n}}.$
3. $\widehat{\lambda} \pm \sqrt{\frac{\widehat{\lambda}}{n}}\, z_{\frac{1-\gamma}{2}}.$
4. $\widehat{\lambda} = 21.3, \qquad \widehat{\sigma} = 4.6.$
5. $\text{se}(\widehat{\lambda}) = 1.5, \qquad \text{se}(\widehat{\sigma}) = 0.2.$
6. 21.3 ± 2.9 ovvero $(18.4\,,\, 24.2)$.

Problema 2. Un'agenzia pubblicitaria è sul punto di decidere in merito ad un ingente investimento in vista dei mondiali di calcio che si svolgeranno in Russia nel 2018. Vuole quindi capire se intraprendere una determinata campagna promozionale attraverso Twitter, e a tal fine vuole verificare se la percentuale p di tifosi che commenta su Twitter i mondiali di calcio sia in crescita dal 2010 al 2014.

L'agenzia sta per ricevere i risultati di un'intervista fatta al termine dei mondiali del 2010 a un campione casuale di 120 tifosi ai quali è stato semplicemente chiesto se durante quei mondiali abbiano pubblicato almeno un tweet per commentarli.

Analogamente il 15 luglio, al termine dei mondiali del 2014, l'agenzia porrà lo stesso quesito a un nuovo campione casuale, sempre di 120 tifosi.

(a) Si introduca un opportuno test statistico per verificare l'ipotesi di interesse per l'agenzia, sulla base dei dati campionari che saranno a disposizione il 15 luglio. Si introducano esplicitamente: variabili (o popolazioni) oggetto dell'inferenza, ipotesi nulla, ipotesi alternativa, campioni casuali, regione critica di livello α.

Arrivano i risultati della prima intervista: nel 2010, su 120 tifosi, 78 hanno pubblicato almeno un tweet sui mondiali. Impazienti di dare una risposta all'agenzia, a bordo di una DeLorean raggiungiamo il 15 luglio 2014: nel 2014, su 120 tifosi, 102 hanno pubblicato almeno un tweet sui mondiali.

(b) Dare una risposta all'agenzia ad un livello di significatività del 5%.

(c) La conclusione è forte o debole?

(d) Calcolare il p-value dei dati raccolti.

Risultati.

(a)

$$X_{2010} = \text{risposta di un tifoso nel 2010} = \begin{cases} 1, & \text{almeno un tweet} \\ 0, & \text{altrimenti,} \end{cases} \sim B(p_{2010})$$

$$X_{2014} = \text{risposta di un tifoso nel 2014} = \begin{cases} 1, & \text{almeno un tweet} \\ 0, & \text{altrimenti,} \end{cases} \sim B(p_{2014})$$

$$p_k = \text{percentuale di tifosi che commenta i mondiali con almeno un tweet nell'anno } k$$

$$H_0 : p_{2010} \geq p_{2014} \qquad \text{vs} \qquad H_1 : p_{2010} < p_{2014}$$

$$\begin{array}{lll} X_{2010,1}, \ldots, X_{2010,120} & \text{campione casuale} & B(p_{2010}) \\ X_{2014,1}, \ldots, X_{2014,120} & \text{campione casuale} & B(p_{2014}) \end{array}$$

$$R_\alpha : Z_0 = \frac{\widehat{p}_{2010} - \widehat{p}_{2014}}{\sqrt{\widehat{p} * (1 - \widehat{p}) * (\frac{1}{120} + \frac{1}{120})}} < -z_\alpha$$

dove

$$\widehat{p}_k = \frac{\sum_{j=1}^{120} X_{k,j}}{120}, \qquad \widehat{p} = \frac{\sum_{j=1}^{120} X_{2010,j} + \sum_{j=1}^{120} X_{2014,j}}{240}.$$

(b) Essendo $\widehat{p}_{2010} = 0.65$, $\widehat{p}_{2014} = 0.85$, $\widehat{p} = 0.75$, risulta

$$Z_0 = -3.577708764 < -z_\alpha = -1.645,$$

pertanto l'ipotesi nulla viene rifiutata a livello 0.05: l'agenzia ha motivo di ritenere che la percentuale di tifosi che commenta su Twitter i mondiali di calcio sia aumentata.

(c) La conclusione è forte.

(d) Il p-value dei dati raccolti è il livello α per cui $Z_0 = -3.577708764 = -z_\alpha$, per cui

$$\begin{gathered} z_\alpha = 3.577708764 \\ 1 - \alpha = \Phi(z_\alpha) = \Phi(3.577708764) = 0.9998266903 \\ \alpha = 0.0001733097 = 1.7 \cdot 10^{-4} \end{gathered}$$

Problema 3. Dopo la delusione mondiale, il difensore della Juventus e della nazionale Leonardo Bonucci, notoriamente avvezzo al gioco d'azzardo, decide di consolarsi con le scommesse. In particolare è molto interessato alle quote che i bookmaker inglesi danno per la somma dei Km percorsi dalle due squadre durante i 90 minuti della finale mondiale 2014, e vorrebbe scommettere sul fatto che tale somma sarà maggiore di 190.

Pertanto raccoglie informazioni sulla variabile

Y = somma dei Km percorsi dai giocatori delle due squadre durante una partita dei mondiali 2014

analizzando i dati relativi alle 48 partite della fase a gironi[1]. Per questi dati

$$\overline{y} = 213.6735 \qquad s_y^2 = 115.6454.$$

Inoltre il test di Shapiro-Wilk per la normalità di Y, eseguito sempre su questi dati, restituisce un p-value di 0.3974, mentre il loro Normal Probability Plot è riportato in Figura 1.

Bonucci però non si accontenta dei dati relativi alla sola variabile Y, poiché ritiene che ci possa essere una relazione tra la somma Y dei Km percorsi e la temperatura x (in gradi centigradi) alla quale è stata giocata la partita, e vorrebbe sfruttare questa relazione prima di scommettere. Per fare questo raccoglie le temperature relative alle stesse 48 partite[2] e imposta il seguente modello empirco gaussiano di regressione lineare:

$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

I risultati ottenuti elaborando i dati tramite il software statistico R sono i seguenti:

```
Coefficients:
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    213.6      8.937   23.904    <2e-16 ***

X           0.001816      0.351    0.005     0.996
---

Signif. codes:  0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Residual standard error: 10.87 on 46 degrees of freedom

Multiple R-squared: 5.82e-07, Adjusted R-squared: -0.02174
F-statistic: 2.677e-05 on 1 and 46 DF,  p-value: 0.9959
```

Il grafico di dispersione di Y su x è riportato in Figura 2, i grafici di diagnostica dei residui in Figura 3, mentre il test di Shapiro-Wilk sui residui restituisce un p-value di 0.3961.

(a) Scrivere l'equazione della retta ai minimi quadrati.

(b) Discutere la validità dell'ipotesi gaussiana del modello empirico di regressione lineare.

(c) Discutere più in generale la bontà del modello empirico di regressione lineare.

(d) Sapendo che per il 12 luglio 2014 a Rio de Janeiro alle 16:00, ora locale, è prevista una temperatura di 22°C, costruire un opportuno intervallo di previsione al 95% per la somma Y dei km percorsi dalle due squadre durante la finale dei mondiali di calcio 2014. Giustificare la validità delle formule usate.

(e) Consiglierete a Bonucci di scommettere?

[1] per questo esercizio sono stati utilizzati i dati veri raccolti direttamente dal sito ufficiale della FIFA
[2] anche le temperature utilizzate sono quelle vere

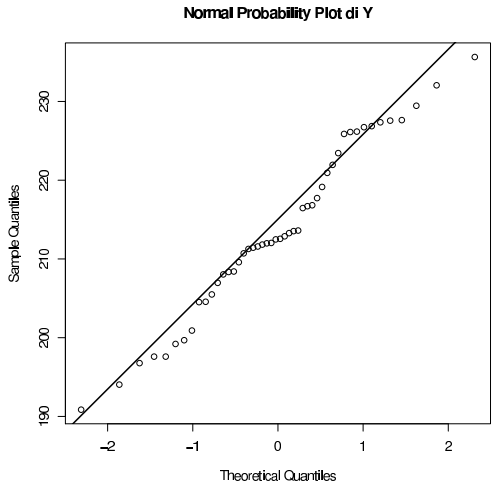

Figura 1: Normal Probability Plot di Y

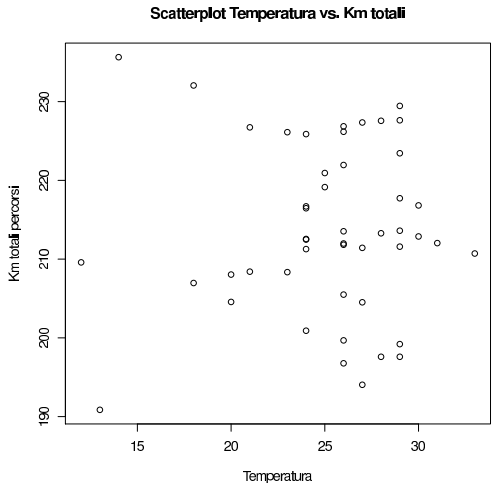

Figura 2: Diagramma di dispersione di Y su x

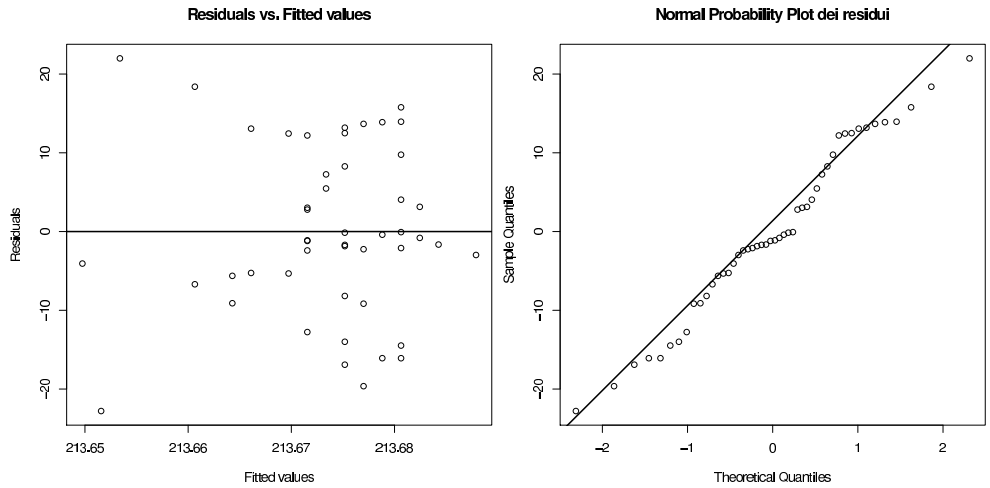

Figura 3: Grafici di diagnostica dei residui del modello lineare di Y su x

Risultati.

(a) $\widehat{y} = 213.6 + 0.001816\, x$.

(b) Le ipotesi gaussiane del modello di regressione lineare sono rispettate, infatti dal grafico dei residui osservo che non ci sono tendenze o andamenti anomali della variabilità (quindi l'ipotesi di omoschedasticità è rispettata), mentre dal Normal Probability Plot e dal valore piuttosto alto del p-value del test di Shapiro-Wilk non ho evidenza per rifiutare l'ipotesi di normalità dei residui.

(c) Questo modello di regressione non risulta essere adatto al problema in analisi: la percentuale di variabilità di Y spiegata dal modello con la variabilità di x è bassissima ($R^2 = 5.82 \cdot 10^{-07}$) e la regressione non risulta essere significativa (p-value della regressione molto alto: 0.9959).

(d) Essendo il modello non significativo non ha senso utilizzarlo per fare previsione. Fortunatamente però, non abbiamo evidenza per rifiutare l'ipotesi di normalità di Y, e possiamo quindi utilizzare l'intervallo di previsione per una popolazione gaussiana:

$$\overline{y} \pm s_y \sqrt{1 + \frac{1}{n}}\, t_{\alpha/2;n-1} = \overline{y} \pm s_y \sqrt{1 + \frac{1}{48}}\, t_{0.025;47} = 213.6735 \pm 21.8582 = (191.8153\,,\, 235.5317).$$

(e) Essendo l'intervallo di previsione completamente superiore al valore 190, cosiglieremmo a Bonucci di scommettere.