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

II Appello di Statistica per Ingegneria Energetica
19 luglio 2016

Cognome, Nome e Numero di matricola:

Problema 1. L'Ordine degli Alchimisti deve preparare 10·000 ampolle di altofuoco da 450 cc ciascuna. In realtà in ogni ampolla viene versata una quantità X di sostanza che vale 450 cc solo in media, ma che per il resto è casuale, con una distribuzione gaussiana, la cui deviazione standard σ dipende dalla precisione del processo di produzione. Una ampolla è considerata "ben preparata" se 448 cc $\leq X \leq$ 452 cc.

1. Calcolare, in funzione di σ, la probabilità che una ampolla risulti ben preparata.
2. Quanto deve valere σ se gli Alchimisti vogliono che tale probabilità sia almeno del 99%?

Gli alchimisti ottengono il valor critico σ_0 che rende la probabilità considerata esattamente pari al 99%.

3. Con quale probabilità vengono versati più di 452 cc in un'ampolla?

Alla fine gli Alchimisti producono ben 13·000 ampolle di altofuoco, grazie ad un aumento del loro potere che non si vedeva dalla scomparsa dei draghi. Le quantità di sostanza versate nelle diverse ampolle sono tutte indipendenti. Sia N il numero di ampolle "ben preparate" sulle 13·000 prodotte.

4. Qual è la distribuzione di N?
5. Con quale probabilità ci saranno almeno 12·850 ampolle "ben preparate"?

Risultati.

1. Ponendo $Z = \dfrac{X - 450}{\sigma} \sim N(0,1)$ si trova

$$\mathbb{P}(448 \text{ cc} \leq X \leq 452 \text{ cc}) = \mathbb{P}\left(|Z| \leq \frac{2 \text{ cc}}{\sigma}\right) = 2\,\Phi\left(\frac{2 \text{ cc}}{\sigma}\right) - 1$$

2.

$$2\,\Phi\left(\frac{2 \text{ cc}}{\sigma}\right) - 1 \geq 0.99 \iff \Phi\left(\frac{2 \text{ cc}}{\sigma}\right) \geq 0.995 \iff \frac{2 \text{ cc}}{\sigma} \geq z_{0.005} = 2.576 \iff \sigma \leq 0.7764 \text{ cc}$$

3. Per simmetria, essendo 450 il punto medio di $[448\,,\,452]$, si ha

$$\mathbb{P}(X > 452 \text{ cc}) = \frac{1 - \mathbb{P}(448 \text{ cc} \leq X \leq 452 \text{ cc})}{2} = 0.005$$

4. $N \sim B(13{\cdot}000\,,\,0.99)$.

5. Dato che $13{\cdot}000 \cdot 0.99 = 12{\cdot}870 > 5$ e $13{\cdot}000 \cdot 0.01 = 130 > 5$, vale l'approssimazione normale $N \simeq N(12{\cdot}870\,,\,128.7)$, per cui

$$\begin{aligned}\mathbb{P}(N \geq 12{\cdot}850) = \mathbb{P}(N > 12{\cdot}849.5) = \mathbb{P}\left(\frac{N - 12{\cdot}870}{\sqrt{128.7}} \geq -\frac{20.5}{\sqrt{128.7}}\right) \\ \simeq \Phi\left(\frac{20.5}{\sqrt{128.7}}\right) = \Phi(1.81) = 0.964852\end{aligned}$$

Problema 2. La tabella seguente riporta i prezzi, in euro, di 9 modelli di moto, praticati da due diffferenti rivenditori:

Modello	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Rivenditore 1	4585	4736	4262	4440	4398	4823	4459	4320	4268
Rivenditore 2	4516	4550	4203	4285	4408	4570	4348	4385	4231

(a) Si ritiene che il primo rivenditore sia in media più caro del secondo. Verificare la veridicità di questa ipotesi, sulla sola base dei 9 modelli considerati, usando un opportuno test di livello $\alpha = 0.05$. Specificare ipotesi nulla, ipotesi alternativa, regione critica, assunzioni eventualmente necessarie per il test, conclusione (H_0 o H_1), forza della conclusione (debole o forte).

(b) Calcolare il p-value dei dati raccolti per il test appena eseguito.

(c) Calcolare un intervallo di confidenza bilatero al 95% per la differenza media fra i prezzi praticati dai due rivenditori per uno stesso modello di moto.

Risultati.

(a) Serve un test per la differenza di medie per dati accoppiati. Sia quindi X la differenza di prezzo fra i due rivenditori (1-2) per uno stesso modello di moto, e sia $d = \mathbb{E}[X]$ la differenza media. Dobbiamo verificare l'ipotesi

$$H_0 : d = 0 \qquad \text{contro} \qquad H_1 : d > 0$$

Rifiutiamo H_0 se

$$t_0 = \frac{\overline{d}}{s_d/\sqrt{9}} > t_{0.05;8}$$

che è effettivamente un test di livello $\alpha = 0.05$ se i 9 dati raccolti sono frutto di un campionamento casuale e se la differenza dei prezzi è gaussiana (cioeè $X \sim N(d, \sigma^2)$), non potendo invocare il TCL con solo 9 dati.

Per i dati raccolti otteniamo $\overline{d} = 88.3333$ e $s_d = 99.16022$, per cui

$$t_0 = \frac{88.3333 - 0}{99.16022/\sqrt{9}} = 2.6724 > t_{0.05;8} = 1.86$$

e quindi rifiutiamo H_0: il Rivenditore 1 è mediamente più caro del Rivenditore 2. Conclusione forte ad un livello del 5%.

(b) Troviamo il p-value α dei dati raccolti per il test appena eseguito:

$$t_{0.025;8} = 2.306 < t_0 = 2.6724 = t_{\alpha;8} < 2.896 = t_{0.01;8}$$

$$0.01 < \alpha < 0.025$$

(c) Costruiamo l'intervallo di confidenza:

$$\overline{d} \pm t_{1-\alpha/2;n-1}\frac{s_d}{\sqrt{n}}$$

$$88.3333 \pm 2.306\frac{99.16022}{\sqrt{9}}$$

$$(12.1118\,,\,164.554)$$

Problema 3. L'agenzia immobiliare PoliAffitti vuole elaborare un modello che spieghi il prezzo di affitto mensile Y degli appartamenti che ha in gestione in funzione della metratura x_1 (in m^2), della distanza dal centro x_2 (in km) e dell'età dello stabile x_3 (in anni).

(a) Si scriva il modello empirico gaussiano di regressione lineare multipla che spiega il prezzo di affitto in relazione alle suddette caratteristiche dell'appartamento.

(b) Quanto vale, in funzione dei parametri del modello, la variazione media del prezzo di affitto di un appartamento se, a parità delle altre caratteristiche, la distanza dal centro aumenta di 1 km?

In Figura (i) è riportato il risultato del modello ipotizzato al punto (a), elaborato sulla base di un campione 150 appartamenti. Per il momento la bontà di tale modello non viene messa in discussione. Risolvere quindi i seguenti punti sulla base del modello elaborato.

(c) Stimare la variazione media del prezzo di affitto di un appartamento se, a parità delle altre caratteristiche, la distanza dal centro aumenta di 1 km. Rispondere con:

- una stima puntuale,
- una stima intervallare al 95%.

(d) I dati raccolti permettono di affermare che, aumentando di 1 km la distanza dal centro, a parità delle altre caratteristiche, si ha una diminuizione media del prezzo medio di affitto di almeno 100 €?

Rispondere tramite un opportuno test di ipotesi, specificando: ipotesi nulla, ipotesi alternativa, regione critica, p-value dei dati, conclusione.

All'interno dell'agenzia, non tutti sono d'accordo sul modello costruito. In particolare, Giovanni Giovannelli sostiene che la distanza dal centro non sia importante per stabilire il prezzo, Marco Marchini invece ritiene che l'età dello stabile sia ininfluente, infine Francesca Franceschiello è convinta che tutte le variabili siano necessarie. Nelle Figure (ii) e (iii) sono riportati gli output dei modelli ridotti.

(e) Per valutare la bontà dei modelli proposti, indicare:

- quali modelli hanno residui omoschedastici,
- per quali modelli appare soddisfatta l'ipotesi gaussiana,
- per quale modello è minimo il rumore,
- quali modelli sono significativi,
- quali modelli hanno tutti i predittori significativi.

Alla luce di queste osservazioni, chi supportate tra gli agenti dell'agenzia?

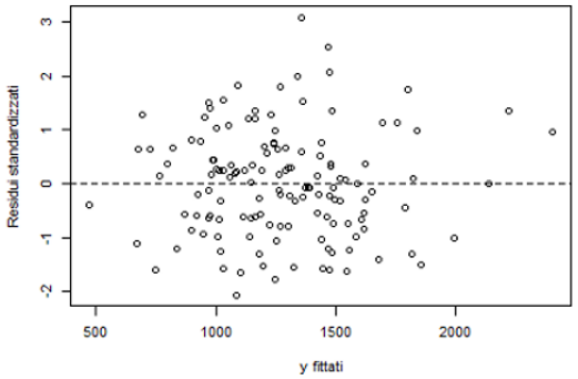
(f) Quali fra le risposte (a)-(d) andrebbero riviste?

```
Residuals:
    Min      1Q  Median      3Q     Max
-346.48 -111.90    0.84  108.71  515.42

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  601.2961    90.6701   6.632 7.61e-10 ***
metratura     14.2846     0.7208  19.816  < 2e-16 ***
distanza    -102.4066    14.1398  -7.242 3.19e-11 ***
eta           -1.2862     1.0115  -1.272    0.206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 167.5 on 133 degrees of freedom
Multiple R-squared:  0.7906,    Adjusted R-squared:  0.7859
F-statistic: 167.4 on 3 and 133 DF,  p-value: < 2.2e-16
```

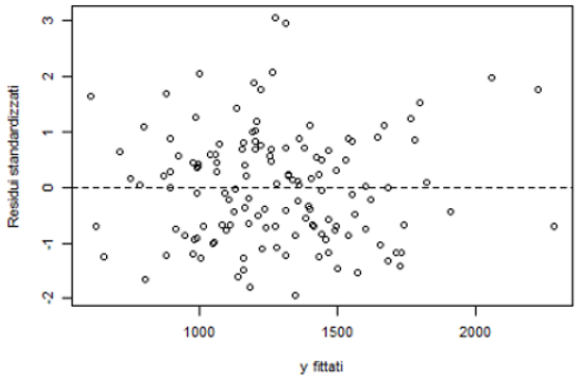
```
Residuals:
   Min     1Q Median     3Q    Max
-382.6 -148.3    0.4  134.3  601.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 143.8116    76.5219   1.879   0.0624 .
metratura    15.0254     0.8394  17.899   <2e-16 ***
eta          -1.2503     1.1900  -1.051   0.2953
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 197 on 134 degrees of freedom
Multiple R-squared:  0.708,     Adjusted R-squared:  0.7037
F-statistic: 162.5 on 2 and 134 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-342.31 -117.11   -5.95   99.56  497.51

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  561.3805    85.2582   6.584 9.50e-10 ***
metratura     14.3402     0.7212  19.884  < 2e-16 ***
distanza    -102.3185    14.1721  -7.220 3.51e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 167.9 on 134 degrees of freedom
Multiple R-squared:  0.7881,    Adjusted R-squared:  0.7849
F-statistic: 249.1 on 2 and 134 DF,  p-value: < 2.2e-16
```

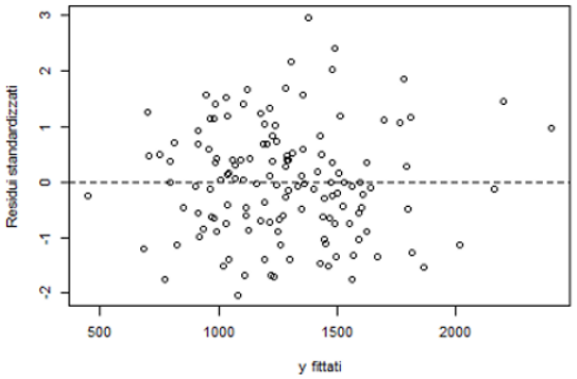
Modello Francesca



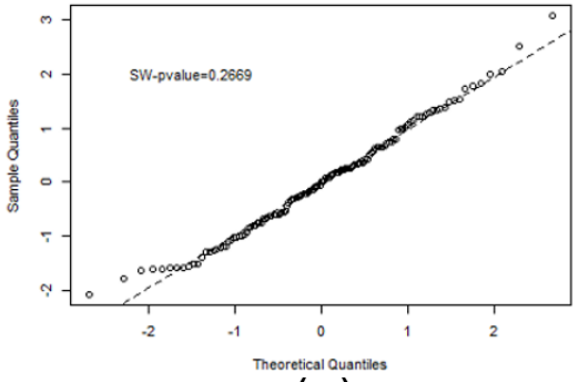
Modello Giovanni



Modello Marco

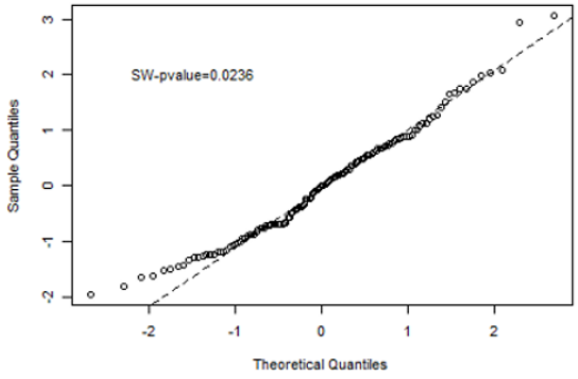


Normal Q-Q Plot



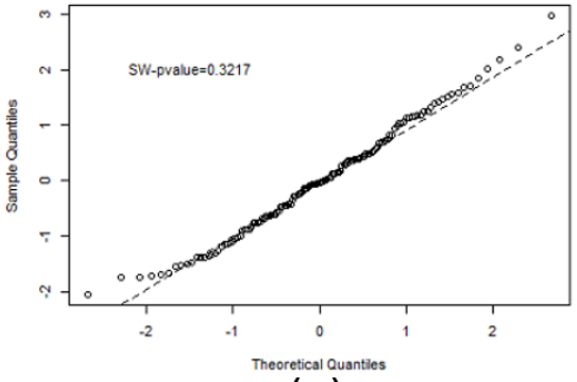
(i)

Normal Q-Q Plot



(ii)

Normal Q-Q Plot



(iii)

Risultati.

(a) $Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$.

(b) Variazione media $= \beta_2$.

(c) Stima della variazione media:

- stima puntuale: $\widehat{\beta}_2 = -102.4066$,
- stima intervallare al 95%:

$$\widehat{\beta}_2 \pm \mathrm{se}(\widehat{\beta}_2)\, t_{0.025}(146) = -102.4066 \pm 14.1398 \cdot 1.96 = -102.4066 \pm 27.7140.$$

(d)

$$H_0 : \beta_2 > -100 \qquad \text{vs} \qquad H_1 : \beta_2 \leq -100$$
$$R_\alpha : \widehat{\beta}_2 < -100 - \mathrm{se}(\widehat{\beta}_2)\, t_\alpha(146)$$

per cui il p-value è il valore di α tale che

$$\widehat{\beta}_2 = -100 - \mathrm{se}(\widehat{\beta}_2)\, t_\alpha(146)$$

Usando l'approssimazione $t_\alpha(146) \simeq z_\alpha$ giustificata dall'alto numero di gradi di libertà (146>120), troviamo

$$z_\alpha = \frac{-\widehat{\beta}_2 - 100}{\mathrm{se}(\widehat{\beta}_2)} = 0.17 \iff \alpha = 1 - \Phi(0.17) = 1 - 0.567495 = 0.432505 = 43.2505\%$$

che è un valore decisamente alto che non consente di rifiutare H_0 agli usuali livelli di significatività: i dati raccolti **NON** permettono di affermare che, aumentando di 1 km la distanza dal centro, a parità delle altre caratteristiche, si ha una diminuizione media del prezzo medio di affitto di almeno 100 €.

(e) Bontà dei modelli proposti:

- residui omoschedastici: modelli 1, 2 e 3
  (con qualche riserva in realtà, ma che vale allo stesso modo per tutti i modelli),
- ipotesi gaussiana soddisfatta: modelli 1 e 3
  (oltre ai residui omoschedastici hanno dei NPP e dei p-value di SW abbastanza buoni),
- minimo rumore: modello 1
  ($R^2_{\mathrm{corr}} = 0.7859$)
- modelli significativi: modelli 1 e 3
  (il modello 2 non soddisfa l'ipotesi gaussiana, quindi i p-value perdono di significato)
- predittori tutti significativi: modello 3
  (il modello 1 ha x_3 non significativo, mentre il modello 2 non soddisfa l'ipotesi gaussiana, quindi i p-value perdono di significato)

Si preferisce quindi il modello 3.

Perde rispetto al modello 1 solo sul valore di R^2_{corr}, ma con un valore comunque molto simile (0.7849 vs 0.7859)

(f) Andrebbero riviste le risposte (a), (c), (d).