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

II Prova in Itinere di Statistica per Ingegneria Fisica
7 luglio 2017

Cognome, Nome, Matricola:

Problema 1. Il Dott. Damakutra vuole determinare la massa μ del pianeta Apolon. Dispone di due strumenti di misura, di precisione differente, ciascuno dei quali introduce un errore casuale additivo gaussiano centrato $\epsilon_k \sim N(0, \sigma_k^2)$, $k = 1, 2$. Sia X_k il risultato di una misura di μ eseguita con lo strumento k.

1. Scrivere X_k in funzione di μ e di ϵ_k e determinarne la distribuzione.

Eseguendo n_k osservazioni con lo strumento k si ottengono n_k misure indipendenti.

2. Proporre uno stimatore corretto $\widehat{\mu}_k$ di μ basato su n_k misure eseguite con lo strumento k.
3. Qual è la distribuzione di $\widehat{\mu}_k$? Quanto vale il suo errore quadratico medio?

Lo strumento 1 è più preciso. Fissate infatti le unità di misura, $\sigma_1 = 1.23$, $\sigma_2 = 3.72$.

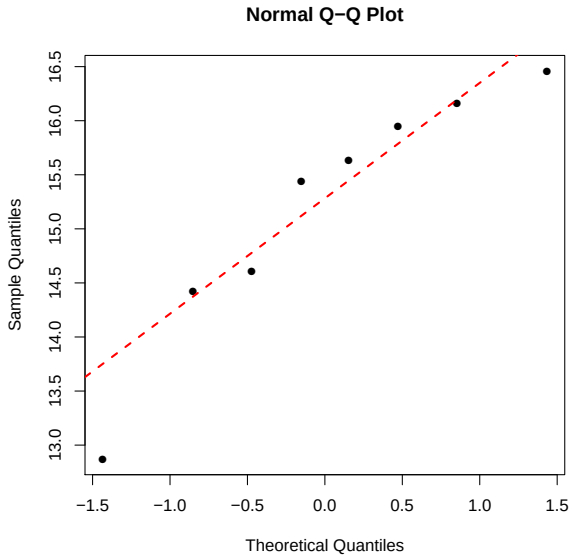
4. Quante misurazioni con lo strumento 2 servono per superare la precisione di 8 misurazioni con lo strumento 1?

Il Dott. Damakutra decide di eseguire 8 misurazioni con lo strumento 1, ottenendo il campione:

12.87 14.61 14.42 15.95 15.63 16.16 15.44 16.46

Per i dati raccolti sono riportati di seguito media campionaria, deviazione standard campionaria, p-value del test $H_0 : \mathrm{Var}(X_1) = (1.23)^2$ contro $H_1 : \mathrm{Var}(X_1) \neq (1.23)^2$, p-value di Shapiro Wilks, il grafico dei quantili normali:

$\overline{x}_8 = 15.53$ $s_8 = 1.18$ p-value varianza $= 0.9853$ p-value SW $= 0.3414$



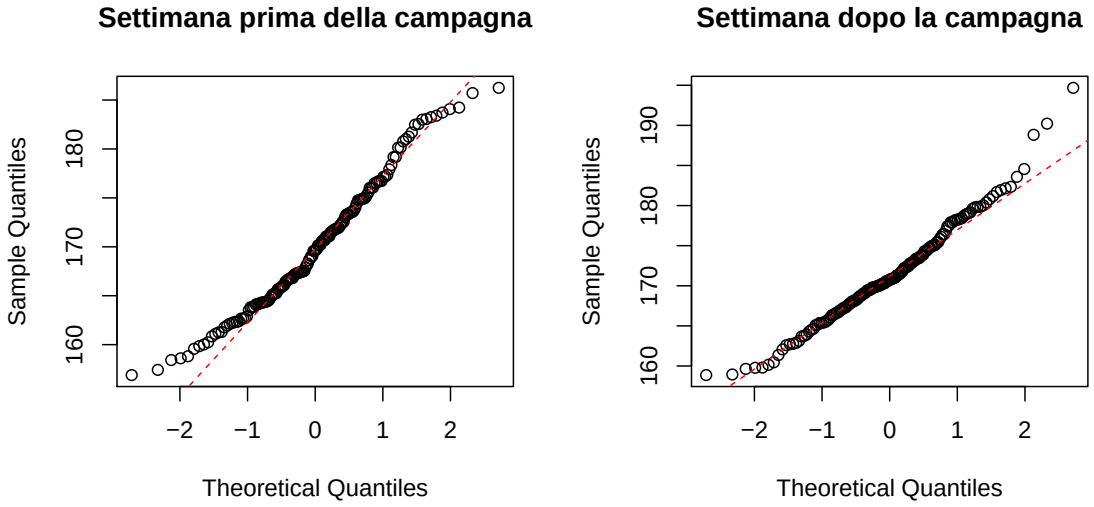
5. Si dia una stima intervallare bilatera di μ ad un livello di confidenza del 90%.
6. Siete sicuri del livello di confidenza della stima trovata? Perché?

Risultati.

1. $X_k = \mu + \epsilon_k \sim N(\mu, \sigma_k^2)$.
2. $\widehat{\mu}_k = \dfrac{X_{k1} + \cdots + X_{kn_k}}{n_k}$.
3. $\widehat{\mu}_k \sim N\left(\mu, \dfrac{\sigma_k^2}{n_k}\right), \qquad \mathrm{MSE}(\widehat{\mu}_k) = \dfrac{\sigma_k^2}{n_k}$.
4. $n_2 > \dfrac{\sigma_2^2}{\sigma_1^2}\, n_1 = 73.17 \qquad$ ovvero $\qquad n_2 \geq 74$.
5. Con una confidenza del 90% stimiamo $\mu = \overline{x}_8 \pm \dfrac{\sigma_1}{\sqrt{8}} z_{0.05} = 15.53 \pm 0.72$, ovvero $\mu \in (14.81\,,\, 16.25)$.
6. Il livello di confidenza è sotto controllo perché sappiamo che le misurazioni $X_1, \ldots, X_8$ sono un campione casuale proveniente da una popolazione gaussiana $N(\mu, 1.23^2)$. Inoltre i dati raccolti, e in particolare il grafico dei quantili normali e i p-value calcolati, non pongono in dubbio né la gaussianità né il valore di σ_1.

Problema 2. Una catena di chioschi street food ha appena avviato la sua prima campagna pubblicitaria su internet. Al fine di valutarne l'efficacia confronta la quantità di cibo venduto (in Kg) da un campione casuale di 130 fra i suoi chioschi nella settimana immediatamente precedente e in quella immediatamente successiva alla campagna.

La media campionaria del consumo nella prima settimana è di 170.13 Kg, con deviazione standard campionaria 6.82 Kg, mentre è di 171.48 Kg nella seconda settimana, con deviazione standard campionaria 6.39 Kg. I p-value dei corrispondenti test di Shapiro-Wilk sono 0.00935 e 0.02237, mentre i QQ-norm dei dati raccolti sono



La variazione di cibo venduto fra le due settimane da ciascun chiosco ha media campionaria pari a 1.35 Kg, con una deviazione standard campionaria di 9.39 Kg.

Come prima verifica degli effetti della campagna pubblicitaria sulle vendite, la catena vi chiede di stimare la differenza fra la quantità media di cibo venduto da un chiosco dopo l'avvio della campagna e la quantità media di cibo venduto prima.

(a) Stimate tale differenza con un intervallo di confidenza bilatero al 98%.

(b) Il livello di confidenza della stima trovata è sotto controllo? Perché?

Rimanendo dubbio l'effetto della campagna, la catena vi chiede di verificarne l'utilità tramite test statistico.

(c) Formulate opportunamente ipotesi nulla e ipotesi alternativa per il test di interesse, specificando in cosa consiste quindi un errore di primo tipo.

(d) Indicate la regione critica di livello α da utilizzare per le ipotesi introdotte e calcolate il p-value dei dati raccolti.

(e) Cosa concludete? La campagna è stata utile?

Risultati. Si tratta di un confronto fra medie per popolazioni accoppiate. Infatti, detta X la quantità di cibo venduta da un chiosco nella prima settimana e Y la quantità di cibo dello stesso chiosco nella seconda settimana, siamo interessati alla variazione $D = Y - X$, variabile aleatoria di distribuzione incognita, media $\mu_D = \mu_Y - \mu_X$ incognita, varianza σ_D^2 incognita.

(a) Un intervallo di confidenza al 98% per la differenza media μ_D del consumo di cibo è:

$$\overline{d}_{130} \pm \sqrt{\frac{s_D^2}{130}}\, t_{\alpha/2}(129) \approx 1.35 \pm 1.92 = [-0.57\,,\, 3.27].$$

L'intervallo di confidenza contiene lo zero, anche se è prevalentemente spostato sul semiasse positivo. L'effetto della campagna rimane dubbio (almeno per $\alpha = 0.02$), potrebbe addirittura aver avuto un effetto negativo.

(b) Il livello di confidenza nominale della stima trovata corrisponde al livello di confidenza reale, in modo esatto se D è gaussiana, oppure in modo approssimato se il campione $D_1, \ldots, D_n$ è sufficientemente numeroso da rendere $\overline{D}_{130}$ approssimativamente normale grazie al TCL.

I grafici e i p-value di Shapiro-Wilks forniti riguardano i campioni $x_1, \ldots, x_{130}$ e $y_1, \ldots, y_{130}$ e quindi non ci permettono di inferire sulla gaussianità di D. Tuttavia il campione $D_1, \ldots, D_{130}$ è sufficientemente numeroso (130>40) da ritenere sotto controllo il livello di confidenza della stima trovata.

(c) Dobbiamo verificare se i dati raccolti forniscono una forte evidenza statistica a favore dell'ipotesi $\mu_D > 0$ (campagna utile). Pertanto

$$H_0 : \mu_D \leq 0, \qquad H_1 : \mu_D > 0.$$

Un errore di primo tipo consiste quindi nel dichiarare la campagna utile quando in realtà non lo è.

(d) Grazie alla numerosità del campione raccolto, la regione critica di livello α da usare è

$$RC_\alpha : \frac{\overline{d} - 0}{\sqrt{s_D^2/n}} > t_\alpha(n-1).$$

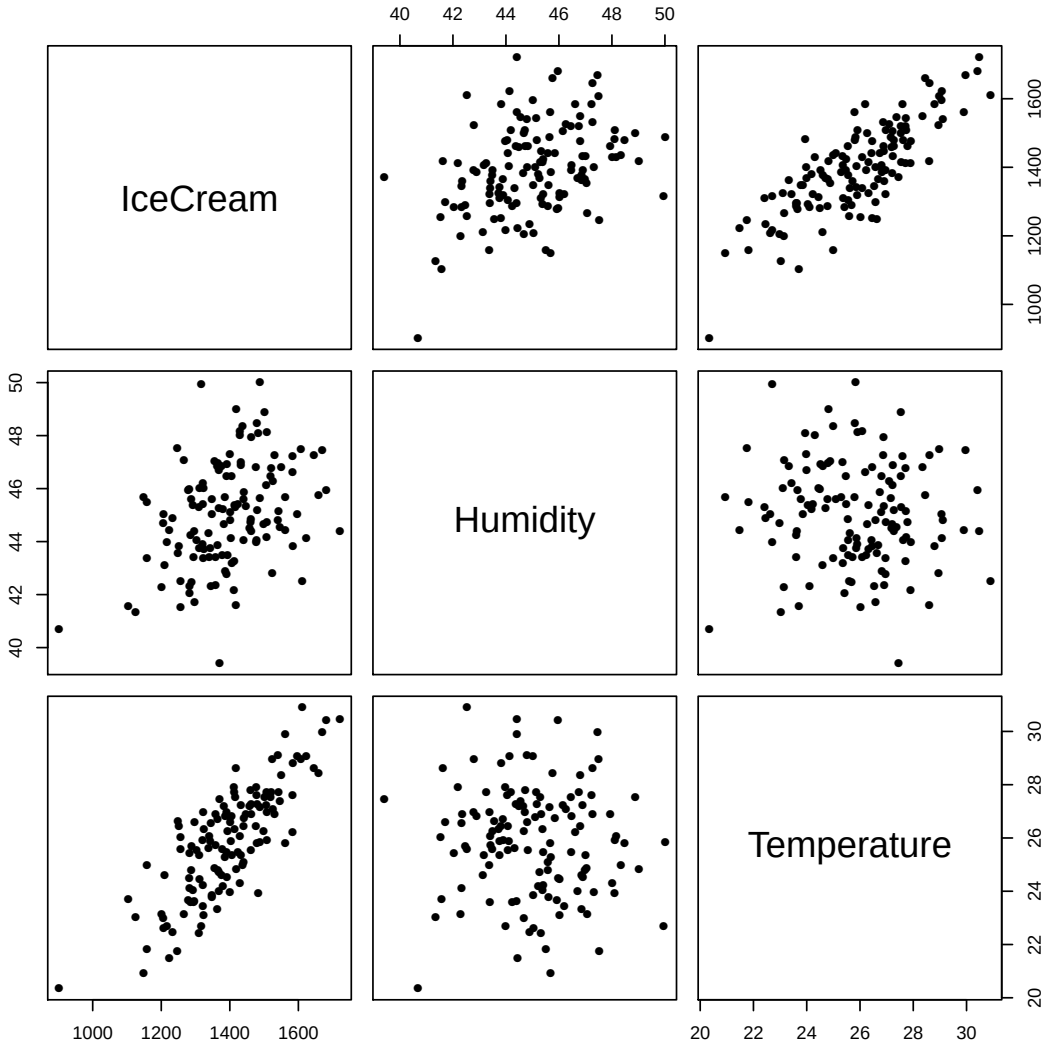
Per i dati raccolti $\dfrac{\overline{d} - 0}{\sqrt{s_d^2/n}} = 1.64$. Essendo $t_\alpha(129) \approx z_\alpha$, il p-value dei dati è

$$\alpha \approx 1 - \Phi(1.64) = 0.0505.$$

(e) Il p-value dei dati è compreso fra l'1% e il 10% e pertanto, agli usuali livelli di significatività, la conclusione non è sempre la stessa, dipende da quanto la catena voglia rischiare un errore di primo tipo. Si può quindi affermare che la campagna è stata utile solo ad un livello $\alpha > 5.05\%$.

Problema 3. Una storica gelateria di Milano sta organizzando la produzione di gelato in vista della bella stagione. Per questo motivo, mette a confronto le vendite giornaliere della stagione scorsa con alcune variabili climatiche, come l'umidità relativa dell'aria media giornaliera e la temperatura media giornaliera, per capire se siano utili a prevedere la domanda giornaliera di gelato.

I dati raccolti la scorsa stagione sono mostrati qua sotto: la quantità (in Kg) di gelato venduto ogni giorno (Y = *IceCream*), il tasso di umidità dell'aria medio (x_1 = *Humidity*) e la temperatura media in gradi Celsius (x_2 = *Temperature*) di 130 giorni consecutivi.



Per descrivere il consumo giornaliero, la gelateria studia due modelli lineari empirici gaussiani:

- M_1: IceCream su Humidity, Temperature e Humidity*Temperature,
- M_2: IceCream su Humidity e Temperature.

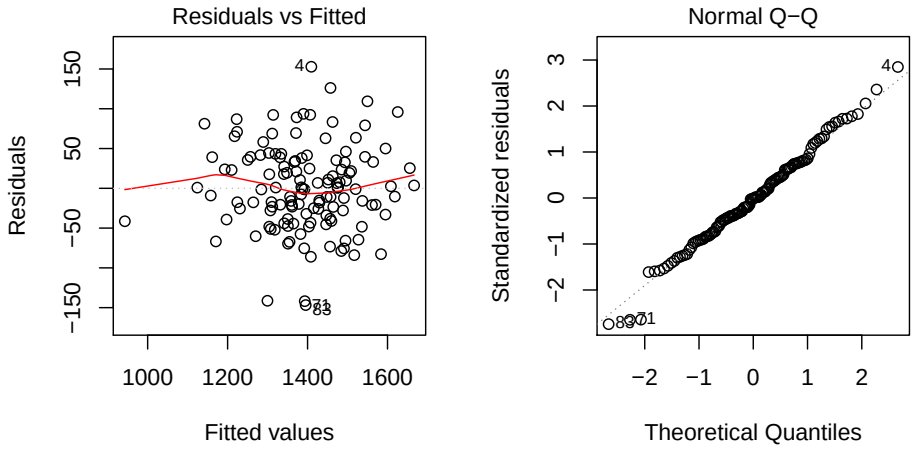
I dati raccolti vengono quindi elaborati sulla base di entrambi i modelli. Si ottengono così le seguenti tabelle riassuntive e i relativi grafici dei residui (i p-value del test di Shapiro-Wilk eseguito sui residui dei due modelli valgono 0.5221 per M_1 e 0.6549 per M_2).

```
Call:
lm(formula = IceCream ~ Humidity + Temperature + Humidity * Temperature,
    data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-146.739  -33.675   -0.732   35.090  152.826

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3010.595   1318.822  -2.283   0.0241 *
Humidity                67.466     29.355   2.298   0.0232 *
Temperature            120.619     51.260   2.353   0.0202 *
Humidity:Temperature    -1.505      1.142  -1.318   0.1899
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.87 on 126 degrees of freedom
Multiple R-squared:  0.8363,    Adjusted R-squared:  0.8324
F-statistic: 214.6 on 3 and 126 DF,  p-value: < 2.2e-16
```

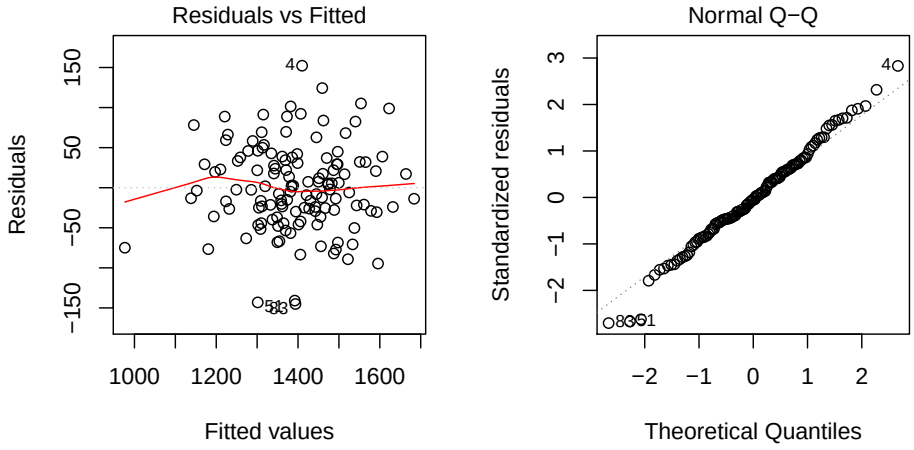


```
Call:
lm(formula = IceCream ~ Humidity + Temperature, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-145.110  -29.641   -2.544   33.114  152.281

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1280.606    128.384  -9.975   <2e-16 ***
Humidity       28.909      2.435  11.874   <2e-16 ***
Temperature    53.126      2.305  23.051   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.03 on 127 degrees of freedom
Multiple R-squared:  0.8341,    Adjusted R-squared:  0.8314
F-statistic: 319.2 on 2 and 127 DF,  p-value: < 2.2e-16
```



(a) Scrivere il legame ipotizzato fra le variabili per entrambi i modelli lineari empirici gaussiani.

(b) Scrivere il legame ipotizzato fra il consumo medio giornaliero e i predittori per entrambi i modelli.

(c) Specificare quale incremento del consumo medio giornaliero corrisponde ad un incremento di 1 grado della temperatura, con tasso di umidità costante del 40%, secondo entrambi i modelli.

(d) Stimare puntualmente tale incremento, secondo entrambi i modelli.

(e) Selezionare un modello tra M_1 e M_2, motivando la risposta.

(f) Quale percentuale della variabilità del consumo viene spiegata dal modello selezionato?

(g) Fornire una stima puntuale del consumo atteso di gelato in una bollente giornata milanese con tasso di umidità medio del 60% e temperatura media di 38 gradi, commentandone la bontà.

(h) Per il modello selezionato, si può ritenere che il coefficiente del tasso di umidità sia maggiore di 28?

Risultati.

(a)
$$M_1: \quad Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_1\, x_2 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2),$$
$$M_2: \quad Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

(b)
$$M_1: \quad \mathbb{E}[Y|x_1, x_2] = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_1\, x_2,$$
$$M_2: \quad \mathbb{E}[Y|x_1, x_2] = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2.$$

(c)
$$M_1: \quad \mathbb{E}[Y|40, x_2 + 1] - \mathbb{E}[Y|40, x_2] = \beta_2 + \beta_3\, 40,$$
$$M_2: \quad \mathbb{E}[Y|40, x_2 + 1] - \mathbb{E}[Y|40, x_2] = \beta_2.$$

(d)
$$M_1: \quad \mathbb{E}[Y|40, x_2 + 1] - \mathbb{E}[Y|40, x_2] \approx \widehat{\beta}_2 + \widehat{\beta}_3\, 40 = 120.619 - 1.505 \cdot 40 = 60.419,$$
$$M_2: \quad \mathbb{E}[Y|40, x_2 + 1] - \mathbb{E}[Y|40, x_2] \approx \widehat{\beta}_2 = 53.126.$$

(e) Senza dubbio il modelo migliore è M_2.

In entrambi i casi l'ipotesi gaussiana è confermata dall'analisi dei residui (normali e omoschedastici) e in entrambi i casi si ha un elevato valore di R^2_{corr} (praticamente identici, leggermente più alto per M_1 a dire il vero: 0.8324 e 0.8314) ed una regressione globalmente molto significativa.

Solo per M_2 però si hanno tutti i predittori significativi. Di più: M_2 è proprio il modello che si ottiene da M_1 eliminando il predittore meno significativo, l'interazione $x_1\, x_2$.

(f) La percentuale della variabilità del consumo spiegata da M_2 è $R^2 = 83.41\%$.

(g) $\widehat{Y} = -1280.606 + 28.909 \cdot 60 + 53.126 \cdot 38 = 2472.758$.

La bontà della stima puntuale è però dubbia: il modello utilizzato è buono, ma il caso $(x_1, x_2) = (60, 38)$ è lontano dai dati con cui è stato elaborato il modello e potrebbe quindi essere fuori dal suo range di validità.

(h) Si deve eseguire il test
$$H_0 : \beta_1 = 28, \qquad H_1 : \beta_1 > 28,$$
con regione critica di livello α
$$RC : \widehat{\beta}_1 > 28 + \text{se}(\hat{\beta}_1)\, t_\alpha(n - 3).$$
I dati raccolti danno
$$\frac{\hat{\beta}_1 - 28}{\text{se}(\hat{\beta}_1)} = 0.3733,$$
da cui, essendo $t_\alpha(127) \approx z_\alpha$,
$$\text{p-value} \approx 1 - \Phi(0.3733) = 0.3557,$$
per cui non consentono di rifiutare H_0 agli usuali livelli di significatività: non c'è forte evidenza statistica che $\beta_1 > 28$.