Esercitazione del 16 maggio 2014

Conclusione dell'esercizio sul consumo annuo di energia elettrica della lezione precedente.
Avevamo un campione di n=101 abitazioni di metratura confrontabile e per ciascuna si era considerato il consumo annuo di energia elettrica come indicatore della efficenza energetica degli edifici di un certo tipo. La deviazione standard campionaria delle 101 abitazioni del campione è pari a 173.8 kWh. Nell'esercitazione precedente abbiamo calcolato l'intervallo di confidenza (bilatero) del 95% per la varianza della popolazione X = *consumo annuo di energia*. Di questa variabile non sappiamo la distribuzione (e dunque anche media e varianza della popolazione sono incognite) e quindi, o la assumiamo gaussiana, oppure (e meglio) visto che il campione è grande possiamo usare i risultati asintotici derivanti dal teorema centrale del limite: la variabile aleatoria $(n-1)S^2 / \sigma^2$ ha distribuzione chi quadrato a (n-1) gradi di libertà, ove S^2 è lo stimatore non distorto e consistente della varianza. Per i dettagli rivedete il materiale della esercitazione precedente.
Ora vogliamo un intervallo di confidenza *unilatero* del tipo (0, *c*) ad un livello dell'(1-α)%. Riprendiamo la formula per l'intervallo bilatero:

$$\left(\frac{(n-1)\,S^2}{\chi^2_{\alpha/2}\,(n-1)} \;,\; \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2}\,(n-1)} \right)$$

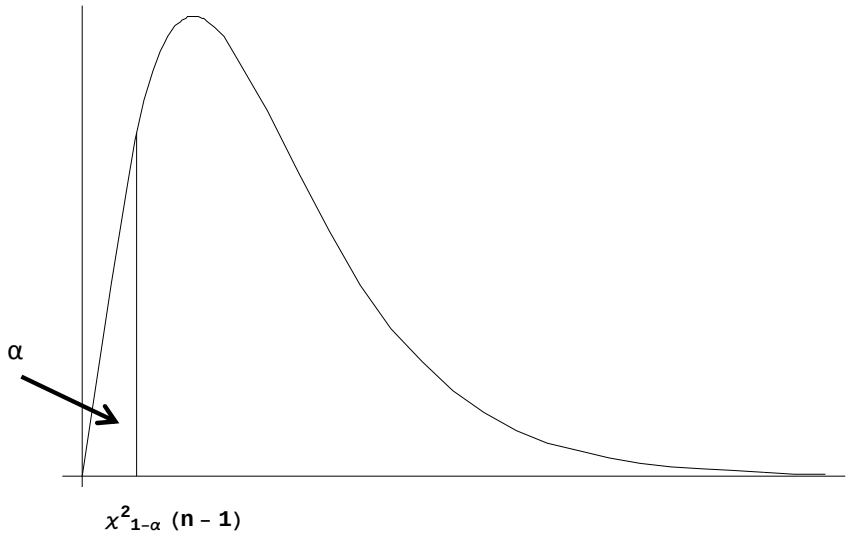
che significa

$$1-\alpha = P\left(\frac{(n-1)\,S^2}{\chi^2_{\alpha/2}\,(n-1)} \le \sigma^2 \le \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2}\,(n-1)} \right)$$

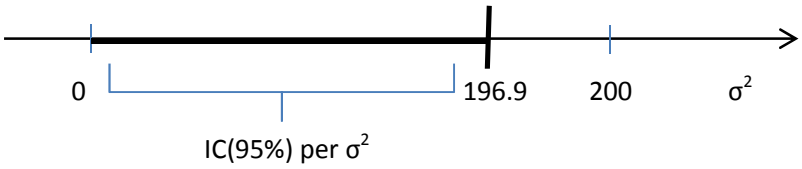
Allora, l'intervallo unilatero di livello (1- α)% sarà quello definito dal quantile $\chi^2_{1-\alpha}$ tale che

$$1-\alpha = P\left(0 \le \sigma^2 \le \frac{(n-1)\,S^2}{\chi^2_{1-\alpha}\,(n-1)} \right)$$

Infatti, dobbiamo usare il quantile "all'inizio" della distribuzione per avere il limite superiore perchè quantili e σ^2 sono inversamente proporzionali (si veda la figura sottostante).



Per i dati del campione, si ha che l'intervallo unilatero che fornisce una limitazione superiore della varianza al livello del 95% è (0, $100 \times 173.8^2/77.9294$) = (0, 3861.3). Quindi l'intervallo per la deviazione standard (che si misura in kWh) si ottiene passando alla radice quadrata degli estremi: (0, 196.9) kWh.
***Complemento*. Allora, sulla base dell'intervallo di confidenza appena trovato, *possiamo rifiutare l'ipotesi (nulla) che il vero valore della deviazione standard della popolazione, σ, sia* (*maggiore o*) *uguale a* (per esempio) 200 kWh *contro l'alternativa che la deviazione standard sia minore di* 200 kWh, al livello del 5%. Si veda la figura sottostante.**



Esercizio 4 dal tema d'esame del 4/09/2013[1]
Si vuole stabilire se una certa sostanza venduta come dietetica ha effetto sul peso. Per questo, vengono reclutati 50 individui: tra questi ne vengono scelti a caso 25 da sottoporre per un mese ad una dieta a base della sostanza dietetica (gruppo D), mentre gli altri 25 seguiranno per un mese una dieta priva della sostanza (gruppo C, controllo). Di ogni individuo viene misurata la variazione di peso allo scadere del mese di dieta (cioè, *peso a fine mese – peso a inizio mese*).
Nel gruppo D, la variazione media di peso è di -467.3 grammi, con una deviazione standard di 1732.9 grammi. Nel gruppo C, la variazione media di peso è di -167.8 grammi, con una deviazione standard di 1708.3 grammi.
1. Verificare tramite test se i dati permettono di stabilire che la sostanza ha davvero effetti dietetici.
2. La stessa conclusione si potrebbe ottenere sulla base di opportuni intervalli di confidenza?
3. Sotto quali ipotesi le conclusioni tratte sono corrette?

Soluzione
1. Siccome i due gruppi vengono formati scegliendo *a caso* tra i 50 individui, i due campioni (della variazione di peso nel gruppo D e di quella nel gruppo C) sono indipendenti. La numerosità del campione permette anche di non imporre la gaussianità delle due popolazioni, e di ricorrere ai risultati approssimati tramite teorema centrale del limite. Indichiamo con μ_D la variazione media di peso nella popolazione D e con μ_C la variazione media di peso nella popolazione C. Se c'è effettivo dimagrimento μ_D e μ_C sono negativi.
L'ipotesi nulla è la situazione *contro* cui ci chiediamo se i dati forniscono abbastanza evidenza:

$$H_0 : \mu_D \geq \mu_C \quad \textit{verso} \quad H_1 : \mu_D < \mu_C$$

Rifiuteremo l'ipotesi nulla di assenza di effetto della sostanza se la differenza $\bar{X}_D - \bar{X}_C$ tra le medie campionarie è un valore negativo grande, cioè la variazione di peso osservata nel gruppo D è molto minore (valori negativi) della variazione di peso nel gruppo C. La statistica test è, allora,

$$T = \frac{\bar{X}_D - \bar{X}_C}{\sqrt{\frac{s^2_D}{n_D} + \frac{s^2_C}{n_C}}}$$

che *sotto l'ipotesi nulla di assenza di efficacia* ha (asintoticamente) distribuzione Gaussiana standard.

[1] ATTENZIONE CHE A LEZIONE HO ERRONEAMENTE INDICATO COME DATA IL 25/06/2013.

Per i due campioni indipendenti del testo si ha che la statistica vale

$$\frac{-467.3 - (-167.8)}{\sqrt{\frac{1732.9^2}{25} + \frac{1708.3^2}{25}}} = -0.61$$

Il p-valore corrispondente (ricordiamo che il test è unilatero) è dato da $P(Z \leq -0.61) = 1-P(Z \leq 0.61) = 1-0.7291 = 0.27$ circa (si veda l'illustrazione qui sotto). Questo valore, che rappresenta il livello di significatività più piccolo che porta a rifiutare il test, è elevato, maggiore dei livelli 5% o 1% di significatività del test, e pertanto **non** possiamo rifiutare l'ipotesi che la nuova sostanza sia inefficace.

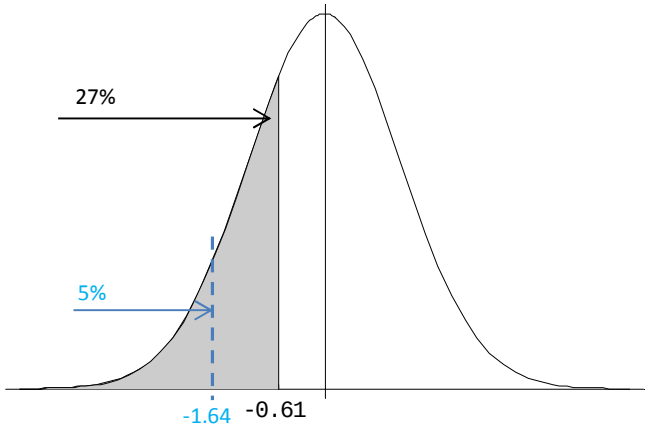
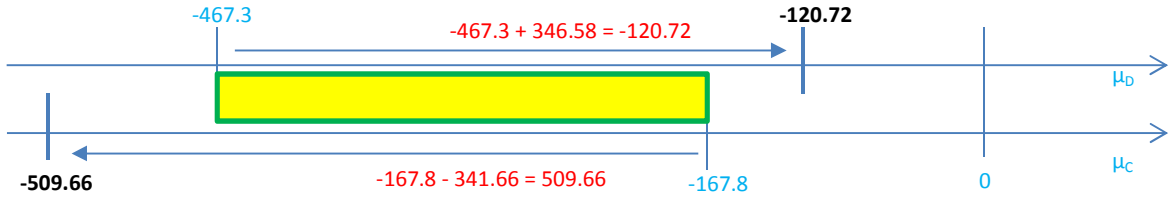


Illustrazione del calcolo del p-valore (area grigia, indicazioni in nero) e suo confronto con il livello di significatività del 5% (azzurro).

2. Determiniamo l'intervallo di confidenza (bilatero) per μ_D e confrontiamolo con quello di pari livello per μ_C. Siamo nel caso dell'approssimazione col teorema centrale del limite, e quindi l'intervallo è della forma:

$$\left(\bar{x} - z_{\alpha/2}\sqrt{s^2/n} \;,\; \bar{x} - z_{\alpha/2}\sqrt{s^2/n}\right)$$

Per il gruppo D si ha che l'intervallo è centrato sul valore -467.3 e la sua lunghezza dipende dal termine $\sqrt{s^2/n} = \sqrt{1732.9^2/25} = 346.58$ mentre l'intervallo di pari livello per il gruppo C è centrato sul valore -167.8 e la sua lunghezza dipende dal termine $\sqrt{s^2/n} = \sqrt{1708.3^2/25} = 341.66$.
Già per $z_{\alpha/2} = 1$, quindi, i due intervalli si sovrappongono (- 467.3 + 346.58 = -120.72 > -167.8) , come si vede nella figura sottostante.



da cui si vede come i due intervalli si intersechino ampiamente (area gialla) già per $z_{\alpha/2} = 1$, cioè in corrispondenza ad un quantile del livello 2 x (1-0.8413) = 0.3174, ove 0.8413 è il valore della funzione di ripartizione della gaussiana standard in z=1. Pertanto, i "veri valori" delle medie incognite nelle due popolazioni potrebbero essere uguali, e quindi non possiamo rifiutare l'ipotesi nulla di inefficacia della nuova sostanza.

Questo risultato era, magari, inatteso visto che le due medie campionarie sono abbastanza diverse. Il fatto che non si possa rifiutare l'ipotesi nulla è dovuto alla notevole ampiezza delle deviazioni standard campionarie in entrambi i gruppi. Provate a ripercorrere il punto 1 ed il punto 2 nell'ipotesi che s_D = 50.3 gr e s_C = 75.1 gr.

Nota 1. A lezione ho usato la *t* di Student senza passare per l'approssimazione gaussiana. Per coerenza con il punto 1 è più adeguato l'uso della gaussiana.

Nota 2. Lo stesso risultato si sarebbe ottenuto considerando degli intervalli di confidenza unilateri per le medie delle due popolazioni. Rifletteteci sopra.

3. Delle ipotesi abbiamo discusso all'inizio del punto 1. E' anche sottinteso che per queste popolazioni non sono note nemmeno le varianze, e dunque se qualcuno avesse scelto di ipotizzare la gaussianità delle popolazioni avrebbe dovuto tenerne conto.

Esercizio 2 dal tema d'esame del 3/9/2012

Su un campione di *n*=9 unità sono state osservate due variabili, *X* ed *Y*, per cui si ipotizza la normalità in popolazione. Sono stati ottenuti questi dati:

x_i	16	21	44	57	51	60	57	34	45
y_i	62.9	63.3	83.4	116.5	126.7	84.5	134.8	101.1	89.2

1. Per prima cosa si è interessati a testare l'affermazione che il valore atteso in popolazione di *Y,* μ_Y sia uguale a 100. E' possibile rifiutare questa ipotesi al livello di significatività dell'1%?
2. Se un'ipotesi H_0 viene accetta (non rifiutata) al livello di significatività dell'1% è sempre vero che verrà accettata (non rifiutata) anche al livello di significatività dell'5?
3. Si ritiene che la variabilità di *Y* possa essere spiegata in funzione di *X*. Rappresentare la distribuzione congiunta di *X* ed *Y* mediante un diagramma di dispersione.
4. Calcolare le stime dei minimi quadrati dei parametri β_0 e β_1 del modello di regressione lineare $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
5. Calcolare una misura della bonta' di adattamento del modello di regressione lineare ai dati (a questo scopo si tenga conto che $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 5396.06$
6. Si considera una seconda variabile *Z* che può essere usata in alternativa a *X* per spiegare la variabilità di *Y.* Sullo stesso campione considerato si ha che $\rho_{Z,Y} = -0.90$. Conviene utilizzare *Z* al posto di *X*? Perché?

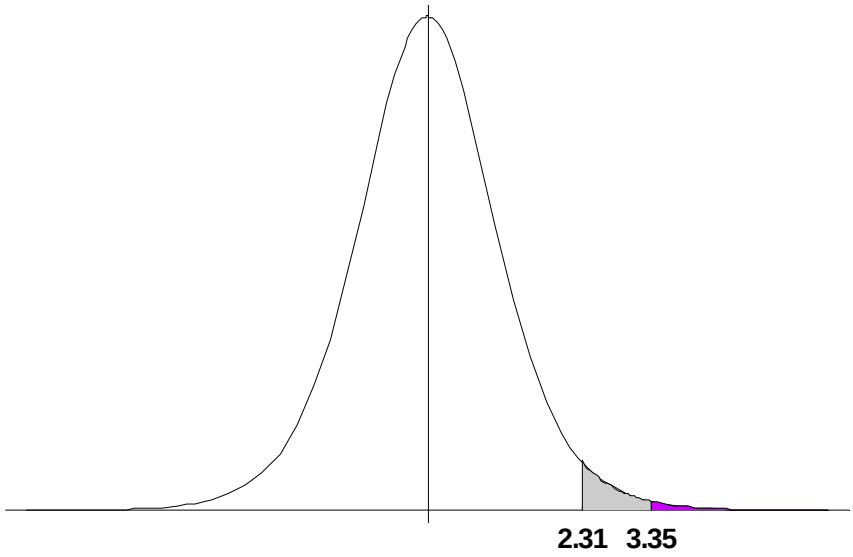
Soluzione.

1. Ipotesi che *Y* sia gaussiana con media, ed anche varianza, incognite.
Verifica d'ipotesi di $H_0 : \mu_Y = 100$ contro l'alternativa $H_1 : \mu_Y \neq 100$. Rifiuto l'ipotesi nulla solo se la media campionaria è molto diversa (più grande o più piccola) da 100. La statistica test è

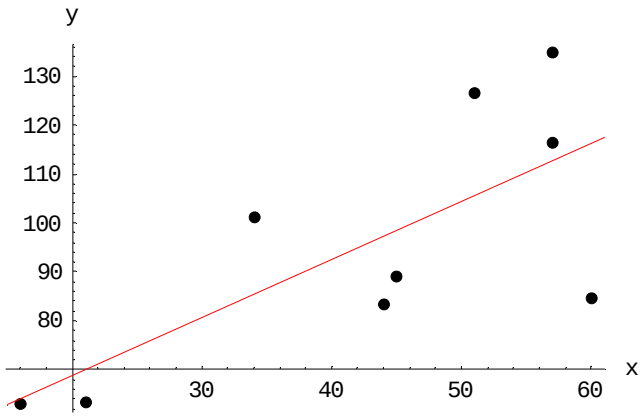
$$T = \frac{\left| \bar{Y} - 100 \right|}{\sqrt{S_Y^2 / 9}}$$

che, sotto l'ipotesi nulla che la media della popolazione sia 100 ha distribuzione *t* di Student a 8 (=9-1) gradi di libertà. Per il nostro campione, la media delle osservazioni y_i è 95.82 e la deviazione standard campionaria s_Y^2 è 674.51. Pertanto la statistica test assume il valore 0.48. Confrontiamo il valore della statistica con il quantile $t_{1-0.005}(8) = 3.3554$. Non è quindi possibile rifiutare l'ipotesi nulla al livello di significatività dell'1%.

2. In generale no, perché aumentando il livello di significatività (accettando, cioè, un errore di prima specie maggiore) si allarga la regione di rifiuto che potrebbe, così, andare ad includere il valore campionario della statistica test. Il ragionamento è illustrato nella figura seguente sulla sola coda di destra: l'area viola indica la regione di rifiuto al livello di significatività dell'1%, mentre l'area grigia indica l'allargamento alla regione di rifiuto al livello di significatività del 5%, per una distribuzione t(8).



3. Il grafico richiesto è riportato nella figura sottostante, assieme con la retta di regressione stimata al punto 4.
Per inciso, i dati sono troppo pochi per farsi un'idea visiva della eventuale relazione tra le due variabili.



4. Per i dati della tabella si ha:

$$\bar{x} = 42.78$$

$$\sum_{i=1}^{9} (x_i - \bar{x})\,(y_i - \bar{y}) = 2420.94$$

$$\sum_{i=1}^{9} (x_i - \bar{x})^2 = 2043.56$$

$$\sum_{i=1}^{9} (y_i - \bar{y})^2 = 5396.06$$

da cui si ottiene $\hat{\beta_1} = 1.185$, $\hat{\beta_0} = 95.82 - 1.185 * 42.78 = 45.12$.

5. In generale, per valutare la bontà del modello si può

a. calcolare il coefficiente di determinazione R^2;

b. testare la significatività della regressione

c. analizzare i residui.

a. $R^2 = \rho_{X,Y}^2 = \dfrac{\sum_{i=1}^{9} (x_i - \bar{x})\,(y_i - \bar{y})}{[\sum_{i=1}^{9} (x_i - \bar{x})^2]\,[\sum_{i=1}^{9} (y_i - \bar{y})^2]} = 0.53$. Indica che solo il 53% della variabilità dei dati attorno alla media è spiegato dalla variabilità attorno alla retta stimata, in rosso nella figura precedente.

b. Testiamo l'ipotesi di significatività della regressione col *t*-test per β_1

$$H_0 : \beta_1 = 0 \text{ contro } H_0 : \beta_1 \neq 0$$

Per questo dobbiamo stimare la varianza, incognita dell'errore, ε, che si suppone gaussiano con media nulla. La stima della varianza è data dalla varianza dei residui: $\hat{e}_i = \hat{y_i} - y_i = 45.12 + 1.185\, x_i - y_i$.

I residui valgono:

$$-1.18, -6.70, -13.86, 3.83, 21.14, -31.72, 22.13, 15.69, -9.24$$

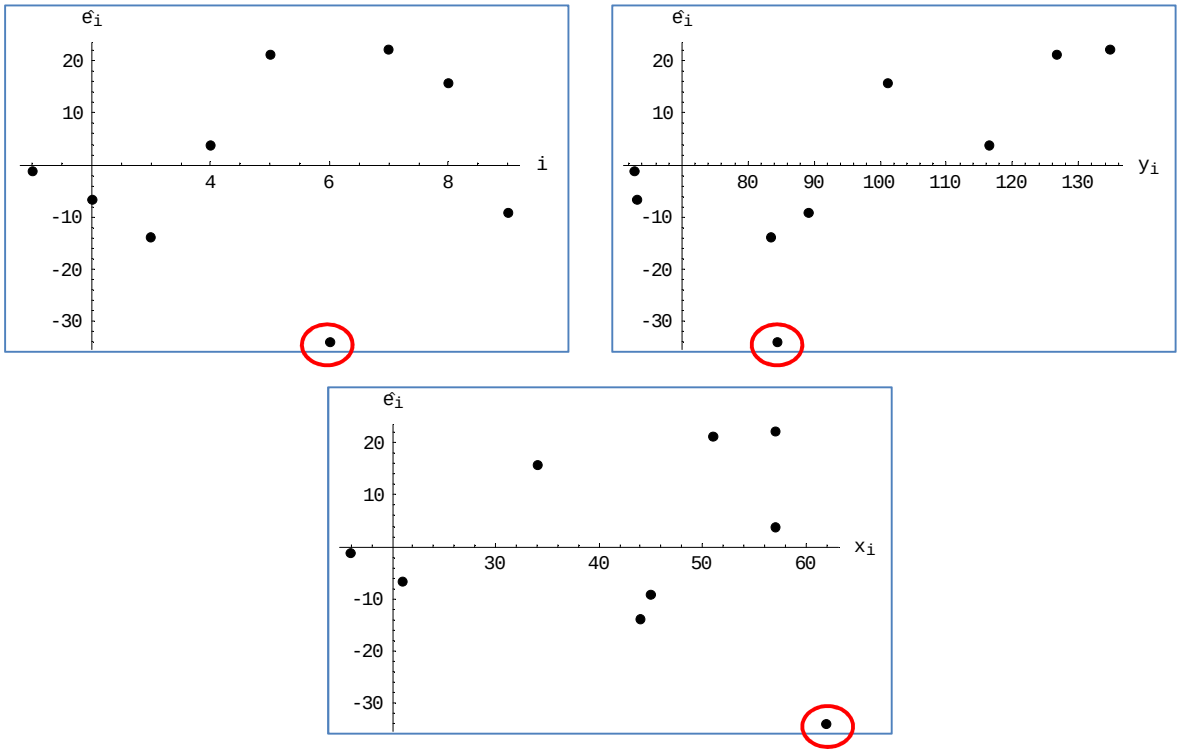
e la loro varianza è data da $s^2 = \sum_{i=1}^{9} \hat{e_i}^2 / 7 = 2527.4 / 7 = 361.057$

Il test si basa sulla statistica

$$\frac{\beta_1}{\sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2/(n-2)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}$$

che ha distribuzione *t*-Student a n-2 (=7) gradi di libertà. Nel nostro caso vale 2.82 : questo valore si colloca in mezzo ai quantili che definiscono le regioni di rifiuto al livello del 5% e dell'1% rispettivamente, pertanto non possiamo rifiutare l'ipotesi nulla al livello del 5%, ma la rifiutiamo al livello dell'1%. La situazione, realisticamente, è pertanto abbastanza incerta. *In questo caso, il p-valore quanto vale, più o meno?*

c. L'analisi dei residui serve a verificare visualmente se le ipotesi di gaussianità dell'errore possono ritenersi soddisfatte. Qui sotto tre grafici dei residui per i nostri dati:



Come accennato anche prima, il numero di punti è troppo basso per chiarire visivamente qualcosa. Tuttavia si nota in tutti i grafici un valore "anomalo", lontano dagli altri dati (evidenziato dal cerchio rosso). In generale, questa situazione richiederebbe un approfondimento dell'analisi. Un altro indicatore *negativo* emerge dall'osservazione del secondo grafico in alto a destra, che mostra valori negativi dei residui in corrispondenza ai valori inferiori alla media campionaria di *Y* e valori positivi in corrispondenza ai valori superiori alla media campionaria. Questo comportamento indica che gli errori (residui) potrebbero non avere una distribuzione casuale attorno al loro valore medio, pari a 0.

6. $\rho_{Z,Y} = -0.9$ implica che $\rho_{Z,Y}^2 = R^2 = 0.81 > 0.53$, e pertanto, la variabile *Z* spiega una quota maggiore di variabilità dei dati attorno alla media tramite la variabilità attorno alla (nuova!) retta di regressione. In questo senso, conviene utilizzare *Z* al posto di *X*, facendo anche (in presenza del campione di *Z*) tutte le opportune verifiche sulla bontà del nuovo modello. *Provate a valutare la significatività di* $R^2 = 0.81$ *con l'analisi della varianza, come nella simulazione d'esame.*

7. ***Parte aggiunta***. Quale valore si può prevedere che venga osservato per *Y* in corrispondenza ad una futura osservazione *x* = 27? Con quale incertezza?

Siccome *x*=27 rientra nell'intervallo di valori di *X* per cui si è stimata la retta di regressione (da 16 a 60), possiamo usare questa retta per prevedere una futura osservazione in corrispondenza a *x*=27: la previsione è 45.12+1.185x27 = 77.11.

L'errore nella previsione è valutato, per esempio, tramite l'intervallo di confidenza del 95% sulla previsione, che è della forma, nel nostro caso:

$$\hat{y} \mp t_{1-0.025}\ (7) * \sqrt{s^2 \left(1 + \frac{1}{9} + \frac{(27 - \bar{x})^2}{\sum_{j=1}^{9} (x_j - \bar{x})^2}\right)}$$

cioè (27.22, 127.01). Come si vede, l'intervallo è molto ampio, anche a causa del basso numero di osservazioni.