GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Regressione lineare

https://www.xlstat.com/en/solutions/features/linear-regression
https://help.xlstat.com/customer/en/portal/articles/2062230-running-a-simple-linear-regression-with-xlstat?b_id=9283
https://help.xlstat.com/customer/en/portal/articles/2062231-running-a-multiple-linear-regression-with-xlstat?b_id=9283
Stesura e adattamento a cura di Luca Biasibetti.

La ***regressione lineare*** è senza alcun dubbio uno dei metodi di modellazione statistica più utilizzati. Viene solitamente fatta una distinzione tra ***regressione semplice*** (con una sola variabile esplicativa) e ***regressione multipla*** (con più variabili esplicative) nonostante i concetti di base e i metodi di calcolo siano identici.
Il metodo di regressione lineare appartiene a una più ampia famiglia di modelli denominata GLM (Generalized Linear Models), così come ANCOVA e ANOVA.
Il principio su cui si basa la regressione lineare è quello di modellizzare una variabile (quantitativa) dipendente Y attraverso una combinazione lineare di p variabili (quantitative) esplicative, X_1, X_2, …, X_p. L'equazione deterministica che rappresenta la regressione può essere scritta nel modo seguente ed è valida per ogni i-esima osservazione:

$$y_i = a_1x_{1i} + a_2x_{2i} + \ldots + a_px_{pi} + e_i$$

dove y_i è il valore osservato per la variabile dipendente relativa all'osservazione i, x_{ki} è il valore assunto per ogni k-esima variabile per l'osservazione i-esima, ed e_i è l'errore del modello.
Il modello viene determinato utilizzando il metodo dei minimi quadrati (minimizzando la somma degli errori quadratici e_i^2).
Le ipotesi assunte per il modello di regressione lineare sono:

- e_i indipendenti;
- e_i seguono una distribuzione normale *N(0, s)*.

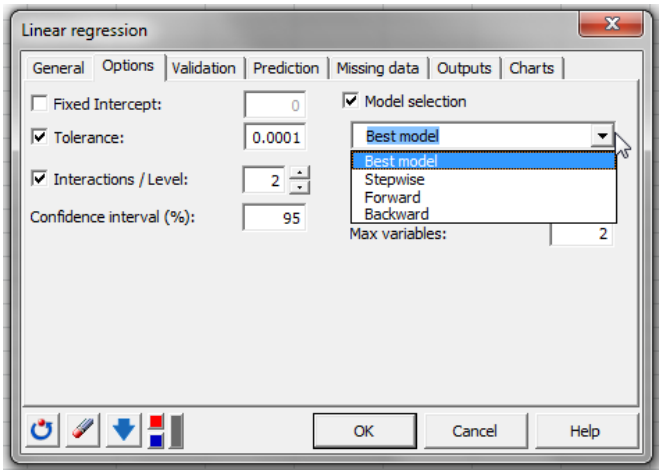
SELEZIONE DELLE VARIABILI

E' possibile selezionare le variabili del modello utilizzando uno dei quattro metodi disponibili in XLSTAT.

- **Best model**: questo metodo ti permettere di scegliere il miglior modello tra tutti i modelli che possono gestire un numero di variabili che può variare tra un valore minimo (Min variables) e un valore massimo (Max Variables). Inoltre l'utente può scegliere numerosi "criteri" per determinare il modello migliore: Adjusted R^2, Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
- **Stepwise**: la selezione del processo inizia aggiungendo la variabile che da il maggior contributo al modello (il criterio utilizzato si basa sulla statistica t-Student). Se una seconda variabile è tale che la probabilità associata con la sua "t" è inferiore alla "Probabilità di entrata", allora viene aggiunta al modello. Lo stesso vale per un'eventuale terza variabile.

Dopo aver aggiunto la terza variabile, viene valutato l'impatto generato dalla rimozione dal modello di ogni singola variabile precedentemente aggiunta (sempre utilizzando la statistica t-Student). Se la probabilità è maggiore della "Probabilità di rimozione", la variabile è rimossa. La procedura continua finché non sono più presenti variabili da poter rimuovere o aggiungere.

- **Forward**: la procedura è la stessa della selezione stepwise ad eccezione che le variabili vengono solo aggiunte e mai rimosse.
- **Backward**: la procedura inizia con l'aggiunta simultanea di tutte le variabili. Le variabili vengono poi rimosse dal modello seguendo la procedura utilizzata per la selezione stepwise.



VALIDAZIONE DELLE IPOTESI

Per verificare se le ipotesi sottostanti al modello sono state correttamente soddisfatte, è possibile utilizzare uno dei i vari test proposti nei risultati della regressione lineare (controllo retroattivo).
La normalità dei residui può essere verificata analizzando i grafici o utilizzando un test di normalità. L'indipendenza dei residui può essere anch'essa verificata analizzando i grafici oppure utilizzando il test di Durbin-Watson.

RISULTATI IN XLSTAT

- **Riepilogo delle variabili selezionate**: se viene scelto un metodo di selezione, XLSTAT mostra il riepilogo della selezione effettuata. Per la selezione stepwise vengono visualizzate le statistiche corrispondenti ai diversi passaggi. Se è stato selezionato il best model con un numero di variabili che può variare da p a q, viene visualizzato il miglior modello per ciascun numero o variabile corredato con le statistiche corrispondenti e, in grassetto, il miglior modello per il criterio scelto.

- **Goodness of fit statistics**: mostriamo nella seguente tabella le statistiche relative all'adattamento del modello di regressione:

- *Observations:* numero di osservazioni utilizzate nei calcoli. Nelle formule mostrate sotto, n è il numero di osservazioni.

Goodness of fit statistics:	
Observations	237,000
Sum of weights	237,000
DF	234,000
R^2	0,630
Adjusted R^2	0,627
MSE	140,858
RMSE	11,868
MAPE	9,049
DW	2,177
Cp	3,000
AIC	1175,598
SBC	1186,002
PC	0,379

- *Sum of weights:* somma dei pesi delle osservazioni utilizzate nei calcoli. Nelle formule mostrate sotto, W è la somma dei pesi.

- *DF:* numero di gradi di libertà per il modello scelto (corrispondente alla parte relativa all'errore).

- *R^2:* coefficiente di determinazione del modello. Questo coefficiente, il cui valore è compreso tra 0 e 1, viene visualizzato solo se la costante del modello non è stata inserita dall'utente. Il coefficiente R^2 può essere interpretato come proporzione tra la variabilità dei dati e la correttezza del modello. Esso misura la frazione della varianza della variabile dipendente espressa dalla regressione. Più il valore di R^2 è vicino a 1, migliore è il modello. Il problema del coefficiente di determinazione è che non tiene conto del numero di variabili utilizzate per l'adattamento del modello.

- *Adjusted R^2:* coefficiente di determinazione adattato per il modello. Se il valore di R^2 è prossimo allo zero il coefficiente R^2-Adjusted (o R^2 corretto) può essere anche negativo. Questo coefficiente viene calcolato solo se la costante del modello non è stata fissata a priori dall'utente. Il valore di R^2-Adjusted rappresenta una correzione del valore di R^2 che tiene conto del numero di variabili utilizzate nel modello.

- *MSE:* errore quadratico medio (Mean of the Squares of the Errors).

- *RMSE:* radice dell'errore quadratico medio (Root Mean Squares of the Errors); è la radice quadrata dell'errore quadratico medio.

- *MAPE:* errore percentuale medio assoluto (Mean Absolute Percentage Error).

- *DW:* statistica di Durbin-Watson. E' un test statistico utilizzato per rilevare la presenza di autocorrelazione nei residui; in un'analisi di regressione lineare è molto importante in quanto una delle ipotesi alla base di questo tipo di analisi è proprio l'indipendenza dei residui. L'utente può fare riferimento alla tabella relativa alla statistica di Durbin-Watson per verificare se l'ipotesi d'indipendenza dei residui può essere accettata.

- *Cp:* indice Cp di Mallows. Più il coefficiente Cp è vicino a p*, meno il modello è distorto.

- *AIC:* criterio di informazione di Akaike. Questo criterio, proposto da Akaike (1973), deriva dalla teoria dell'informazione ed utilizza la misura di Kullback e Leibler (1951). È un criterio utilizzato per la selezione del modello e penalizza i modelli per i quali l'aggiunta di nuove variabili esplicative non fornisce informazioni sufficienti al modello

(tali informazioni vengono rilevate attraverso l'MSE). L'obiettivo è minimizzare il criterio (AIC).

- *SBC:* criterio Bayesiano di Schwarz. Questo criterio, proposto da Schwarz (1978), è simile al criterio AIC e l'obiettivo è minimizzarlo.

- *PC:* criterio di previsione di Amemiya. Questo criterio, proposto da Amemiya (1980), è utilizzato come l'R^2-Adjusted per considerare il livello di moderazione del modello.

- *Press RMSE:* la statistica Press viene visualizzata solo se la relativa opzione è stata preventivamente attivata nella finestra di dialogo. Il Press RMSE può quindi essere confrontato con l'RMSE. Una grande differenza tra i due valori mostra che il modello è sensibile alla presenza o assenza di alcune osservazioni nel modello stesso.

- **Type I SS table**: viene utilizzata per visualizzare l'influenza che l'aggiunta progressiva di variabili esplicative ha sul fitting del modello e mostra la somma degli errori quadratici (SSE), l'errore quadratico medio (MSE), il test F di Fisher, o la probabilità associata al test F di Fisher. Minore è la probabilità, maggiore è il contributo della variabile al modello (tutte le altre variabili sono già presenti nel modello). Le somme dei quadrati nella tabella di "Tipo I" si aggiungono sempre al modello SS.
  Nota: l'ordine in cui sono selezionate le variabili nel modello ha influenza sui valori ottenuti.

- **Type III SS table**: viene utilizzata per visualizzare l'influenza che la rimozione di una variabile esplicativa ha sul fitting del modello (tutte le altre variabili vengono mantenute) e mostra la somma degli errori quadratici (SSE), l'errore quadratico medio (MSE), il test F di Fisher, o la probabilità associata al test F di Fisher. Minore è la probabilità, maggiore è il contributo della variabile al modello (tutte le altre variabili sono già presenti nel modello).
  Nota: a differenza della tabella SS di Tipo I, l'ordine in cui le variabili sono selezionate nel modello non ha alcuna influenza sui valori ottenuti.

- **Analysis of variance table**: è utilizzata per valutare il potere esplicativo delle variabili indipendenti. Nei casi in cui alla costante del modello non è stato assegnato un valore dato, il potere esplicativo viene valutato confrontando l'adattamento del modello finale (effettuato con i minimi quadrati) con l'adattamento del modello rudimentale ideato includendo solo una costante uguale alla media del variabile dipendente. Laddove la costante del modello sia stata invece impostata a priori, il confronto viene effettuato rispetto al modello per cui la variabile dipendente è uguale alla costante impostata.

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	56233,254	28116,627	199,610	< 0,0001
Error	234	32960,761	140,858		
Corrected Total	236	89194,015			

Computed against model Y=Mean(Y)

Type III Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Age	1	2678,226	2678,226	19,014	< 0,0001
Height	1	20309,170	20309,170	144,182	< 0,0001

- **The parameters of the model table**: mostra la stima dei parametri, l'errore standard corrispondente, la statistica t di Student, la probabilità corrispondente e l'intervallo di confidenza.

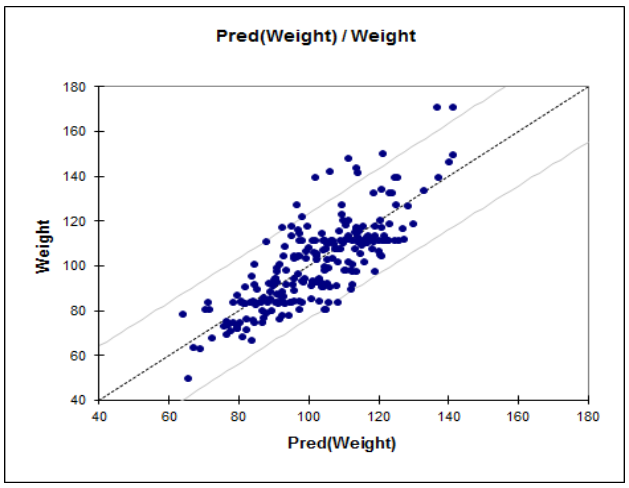
- **Model equation:** l'equazione del modello è mostrata per rendere semplice la lettura o il riutilizzo del modello.

Equation of the model:
Weight = -127,819907444769+0,240274914264462*Age+3,09004803935285*Height

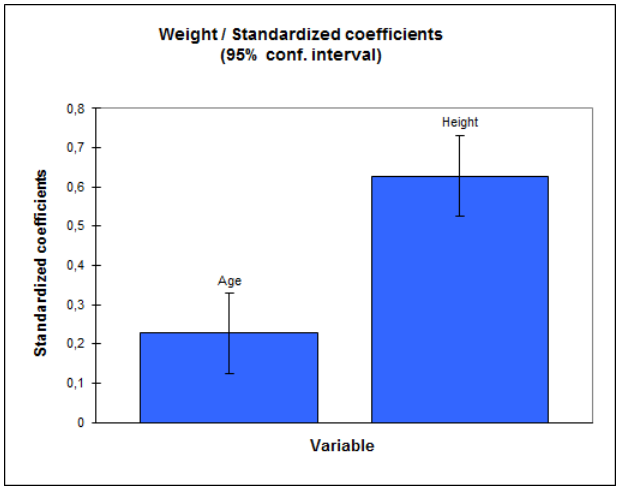
- **Standardize coefficients table:** la tabella dei coefficienti standardizzati è utilizzata per confrontare i pesi relativi delle variabili. Più il valore assoluto di un coefficiente è elevato, più è importante il peso della variabile corrispondente. Quando l'intervallo di confidenza attorno ai coefficienti standardizzati ha valore 0 (può essere visto facilmente sul grafico dei coefficienti normalizzati), il peso di una variabile nel modello non è significativo.

- **Prediction and residuals table:** la tabella delle previsioni e dei residui mostra, per ogni osservazione, il suo peso, il valore della variabile esplicativa qualitativa (se ne è presente una sola), il valore osservato della variabile dipendente, la previsione del modello, i residui, gli intervalli di confidenza insieme alla previsione adattata e la D di Cook se le opzioni opportune sono state precedentemente attivate nella finestra di dialogo. Vengono mostrati due tipi di intervallo di confidenza: un intervallo di confidenza attorno alla media (corrispondente al caso in cui la previsione viene fatta per un numero infinito di osservazioni con un insieme di valori dati per le variabili esplicative) e un intervallo attorno alla previsione isolata (corrispondente al caso di una previsione isolata per i valori dati delle variabili esplicative). Il secondo intervallo è sempre maggiore del primo e i valori casuali sono più elevati. Se sono stati selezionati i dati di convalida, verranno visualizzati alla fine della tabella.

Output Grafici

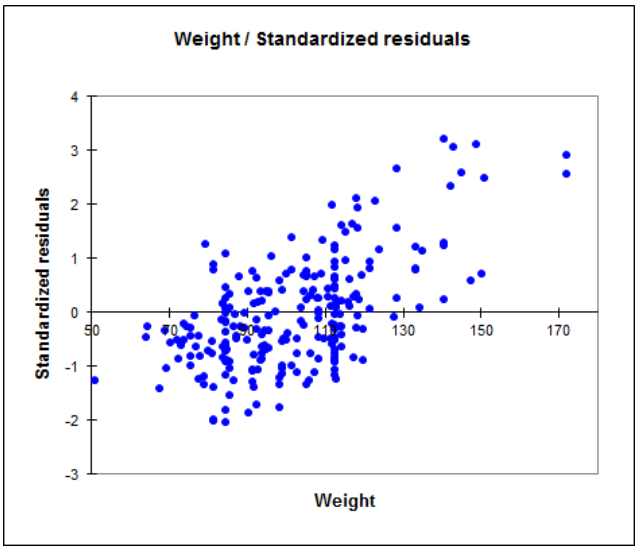


Regressione Lineare: Predicted vs Measured



Regressione Lineare: parametri del modello standardizzato

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it



Regressione Lineare: Standardized residuals vs Measured

USE CASE – REGRESSIONE LINEARE
Regressione lineare semplice in XLSTAT

Il seguente tutorial può essere utile per effettuare ed interpretare una regressione lineare semplice in Excel utilizzando il software XLSTAT. La regressione lineare semplice condotta in questo esempio è basata sul metodo dei minimi quadrati.

Dataset

I dati utilizzati per questo esempio (ricavati da Lewis T. e Taylor L.R. (1967), Introduction to Experimental Ecology, New York: Academic Press, Inc.) riguardano 237 bambini, descritti in base al loro sesso, età (espressa in mesi), altezza in pollici (1 pollice = 2,54 cm) e peso in libbre (1 libbra = 0,45 kg). E' possibile scaricare un file Excel contenente il dataset e i risultati della regressione utilizzando il seguente link:

https://help.xlstat.com/customer/portal/kb_article_attachments/%20108217/original.xls?1489751782

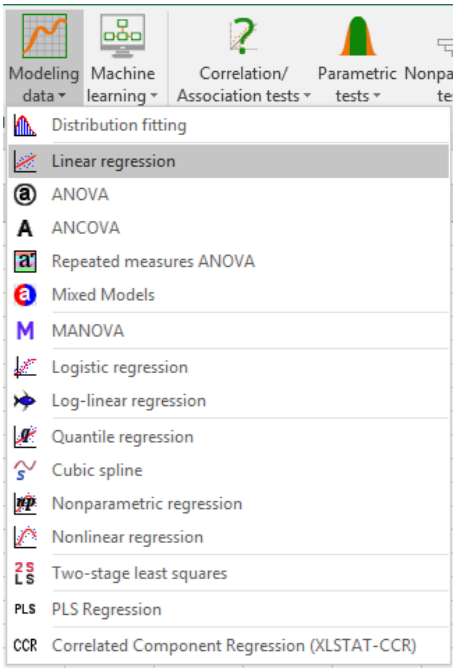
Obiettivi del tutorial

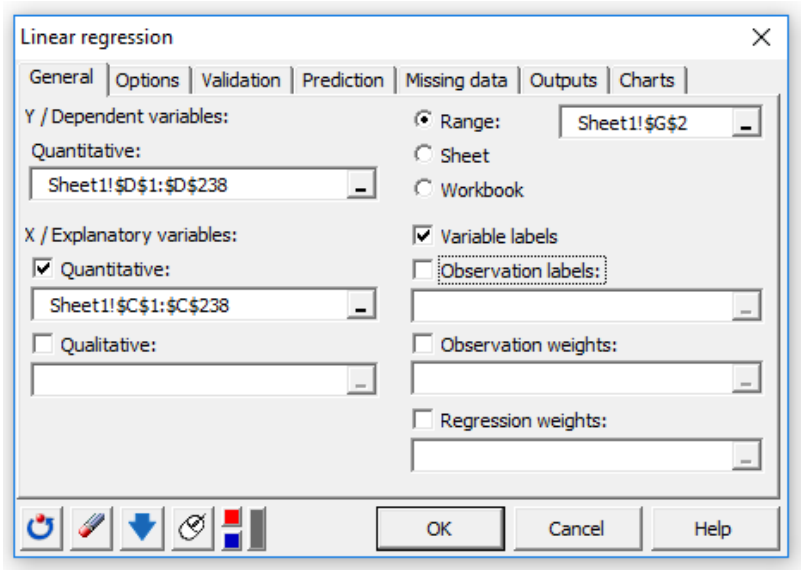
Utilizzando la regressione lineare semplice, vogliamo scoprire come il peso dei bambini possa variare con l'altezza e verificare se è sensato l'utilizzo di un modello lineare per descrivere la relazione tra le due grandezze.

Eseguire una regressione lineare semplice

Dopo aver aperto XLSTAT, selezioniamo il comando **XLSTAT / Modeling data / Regression**; verrà subito visualizzata la finestra di dialogo relativa alla Regressione lineare.
Selezioniamo i dati sul foglio Excel; nel nostro caso la **variabile dipendente** è il Peso ("Weight"). La variabile esplicativa quantitativa è invece l'Altezza ("Height").
Poiché all'interno del file contenente il dataset l'intestazione della colonna è stata assegnata alla descrizione delle variabili, è necessario attivare l'opzione **Variable label**.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it





Comandi utilizzati per eseguire una regressione lineare in XLSTAT. I calcoli relativi alla regressione iniziano dopo aver premuto il tasto **OK** (nella finestra sull'immagine di destra).

Interpretazione dei risultati di una regressione lineare semplice

Goodness of fit statistics:	
Observatio	237,000
Sum of we	237,000
DF	235,000
R²	0,600
Adjusted R	0,599
MSE	151,655
RMSE	12,315
MAPE	9,320
DW	2,125
Cp	2,000
AIC	1192,113
SBC	1199,049
PC	0,406

A questo punto vengono mostrati i risultati della regressione. La prima tabella mostra la bontà dei coefficienti di adattamento del modello. Il coefficiente R^2 (o coefficiente di determinazione) indica la percentuale di variabilità della variabile dipendente spiegata dalle variabili esplicative. Più R^2 è vicino a 1, migliore è la misura.
Nel nostro caso, il 60% della variabilità del Peso è spiegata dall'Altezza. Il resto della variabilità è dovuta ad altri effetti (altre variabili esplicative) non considerati in questa analisi.
È importante a questo punto esaminare i risultati relativi all'analisi della varianza (vedi sotto). I risultati ci permettono di determinare se le variabili esplicative portano o meno informazioni significative (ipotesi nulla H_0) al modello. In altre parole, è un modo per chiedersi se è corretto utilizzare la media per descrivere l'intera popolazione o se le informazioni fornite dalle variabili esplicative hanno valore o meno.

Analysis of variance:					
Source	DF	um of squar	lean square	F	Pr > F
Model	1	53555,028	53555,028	353,137	< 0,0001
Error	235	35638,987	151,655		
Corrected	236	89194,015			
Computed against model Y=Mean(Y)					

Output relativo all'analisi della varianza.

Dato che la probabilità corrispondente al valore F è inferiore a 0,0001, si considera un rischio inferiore allo 0,01% nell'assumere che l'ipotesi nulla (nessun effetto delle due variabili esplicative) sia errata. Pertanto, possiamo concludere con sicurezza che le tre variabili portino una quantità significativa di informazioni.
La seguente tabella fornisce invece i dettagli sul modello, i quali risultano essere molto utili in caso di previsioni o confronti tra coefficienti di diversi modelli (relativi a popolazioni diverse). Possiamo osservare che l'intervallo di confidenza del 95% della variabile Height è molto "stretto", mentre quello dell'intercetta del modello è più ampio.
Osserviamo ora l'equazione del modello riportata sotto la tabella: possiamo notare che nel range considerato per la variabile Height, quando l'Altezza aumenta di un pollice, il Peso aumenta di 3,8 libbre.

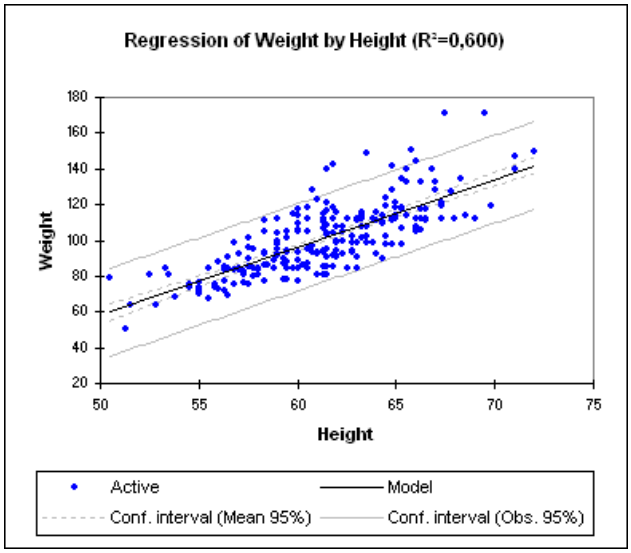
Model parameters:

Source	Value	tandard erro	t	Pr > |t|	er bound (9	er bound (95%)
Intercept	-132,991	12,494	-10,645	< 0,0001	-157,605	-108,377
Height	3,818	0,203	18,792	< 0,0001	3,418	4,218

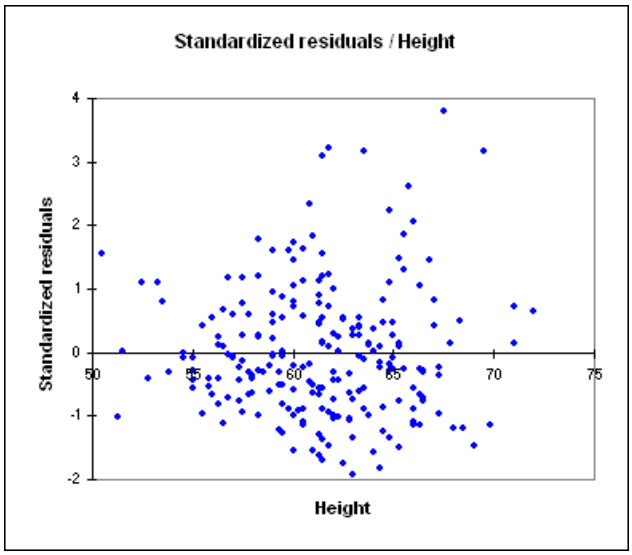
Equation of the model:

Weight = -132,991006806788+3,8181490307087*Height

Un altro possibile output XLSTAT è la tabella che mostra i residui e ci permette di controllare da vicino ogni residuo standardizzato. Questi residui, secondo le ipotesi del modello di regressione lineare, dovrebbero essere normalmente distribuiti, il ciò significa che il 95% dei residui dovrebbe essere nell'intervallo [-1,96; 1,96].
Tutti i valori fuori da questo intervallo sono potenziali outlier o potrebbero suggerire che l'assunzione di normalità sia errata. E' possibile utilizzare la funzione di XLSTAT DataFlagger per evidenziare i residui che non appartengono all'intervallo [-1,96; 1,96].
Su oltre 237 dati, possiamo identificare 9 residui (26, 38, 64, 69, 77) al di fuori dell'intervallo [-1,96; 1,96] pertanto non siamo portati a respingere l'ipotesi di normalità. Un'analisi più dettagliata dei residui può essere effettuata utilizzando il metodo ANCOVA.
Il primo grafico (vedi sotto) ci consente di visualizzare i dati, la retta di regressione (modello adattato) e due intervalli di confidenza; l'intervallo di confidenza sulla media della previsione per un dato valore dell'altezza è quello più vicino alla retta di regressione.
L'altro intervallo rappresentato è invece l'intervallo di confidenza su una singola previsione per un dato valore dell'altezza. Possiamo vedere chiaramente che c'è una tendenza lineare, ma è anche presente un'alta variabilità attorno alla retta di regressione. Inoltre, le 9 osservazioni che si trovano all'esterno dell'intervallo [-1,96; 1,96], sono anche al di fuori del secondo intervallo di confidenza.

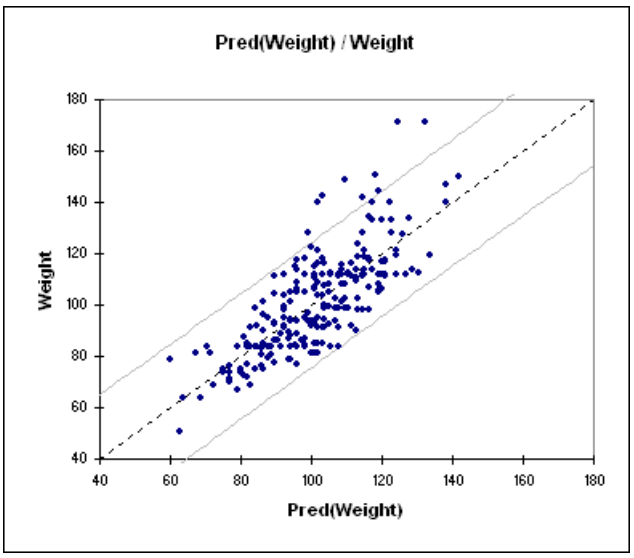


Il secondo grafico (vedi sotto) ci consente di visualizzare i residui standardizzati rispetto all'altezza. Non è questo il caso, ma quando si traccia il grafico dei residui rispetto alla variabile esplicativa, se viene identificato un trend particolare, questo indica che il modello non è corretto ed è presente un'autocorrelazione nei residui, infrangendo quindi una delle ipotesi alla base del modello di regressione lineare parametrica.

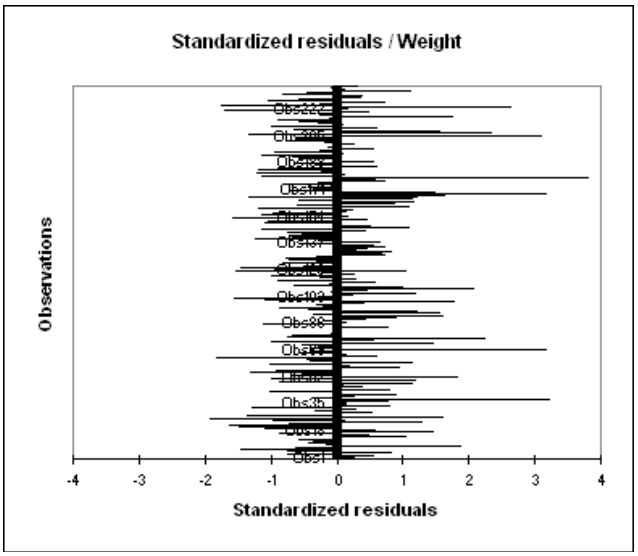


Il grafico riportato di seguito consente di confrontare le previsioni con i valori osservati. I limiti di confidenza consentono, come per il grafico di regressione visualizzato sopra, di identificare gli outlier.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it



L'istogramma dei residui ci consente di visualizzare rapidamente i residui che si trovano fuori dal range [-2, 2].



Conclusioni

Possiamo concludere che la variabile Height ci consente di spiegare il 60% della variabilità di Weight. Una quantità significativa di informazioni però non è spiegata dal modello utilizzato. Vediamo quindi ora, in un tutorial sulla regressione lineare multipla, come l'aggiunta di una nuova variabile "Age" al modello, può migliorare la qualità della misura.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Regressione lineare multipla in XLSTAT

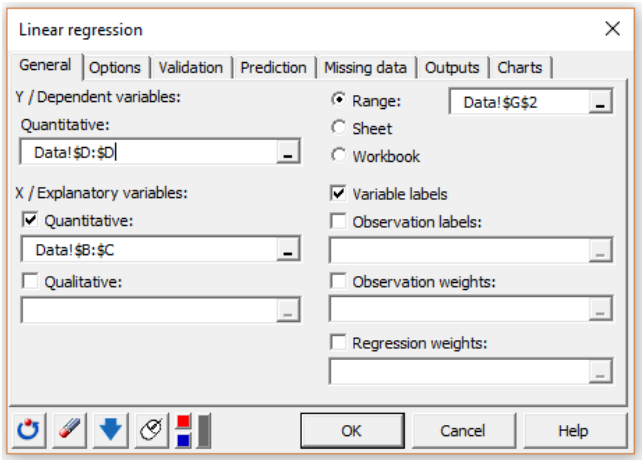
Effettuiamo ora una regressione lineare multipla. Utilizziamo lo stesso dataset dell'esempio precedente sulla regressione lineare semplice; in questo caso considereremo però ai fini della nostra indagine anche la variabile Age (Età - espressa in mesi).
E' possibile scaricare un file Excel contenente il dataset e i risultati della regressione utilizzando il seguente link:

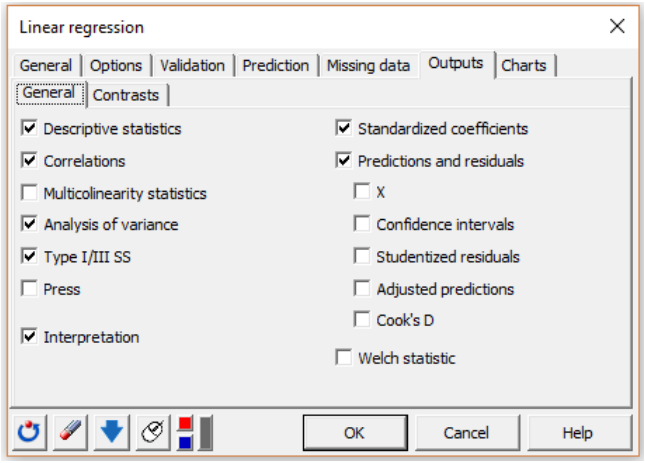
https://help.xlstat.com/customer/portal/kb_article_attachments/%20108218/original.xls?1489752915

Utilizzando la regressione lineare multipla, vogliamo scoprire come il peso dei bambini possa variare con l'altezza e l'età e verificare se è sensato l'utilizzo di un modello lineare per descrivere la relazione tra le due grandezze. Anche in questo caso, la regressione lineare effettuata è basata sul metodo dei minimi quadrati.

Eseguire una regressione lineare multipla

Apriamo XLSTAT e selezioniamo il comando **XLSTAT/Modeling data/Regression;** verrà subito visualizzata la finestra di dialogo relativa alla Regressione lineare.
Selezioniamo quindi i dati sul foglio Excel; la **variabile dipendente** è il Peso ("Weight"). Le **variabili esplicative quantitative** sono invece l'Altezza ("Height") e l'età ("Age").
Poiché nel dataset l'intestazione della colonna è stata assegnata alla descrizione delle variabili, è necessario attivare l'opzione **Variable label**.
Nella scheda **Output** attiviamo l'opzione **Type I / III SS** per poter visualizzare poi i risultati corrispondenti.





Passiamo ora all'analisi dei risultati mostrati come output dell'analisi di regressione multipla.

Interpretazione dei risultati nella regressione lineare multipla

Anche in questo caso la prima tabella mostra la bontà dei coefficienti di adattamento del modello. Il coefficiente R^2 (o coefficiente di determinazione) indica la percentuale di variabilità della variabile dipendente spiegata dalle variabili esplicative. Più R^2 è vicino a 1, migliore è la misura.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

Goodness of fit statistics:	
Observations	237,000
Sum of weights	237,000
DF	234,000
R^2	0,630
Adjusted R^2	0,627
MSE	140,858
RMSE	11,868
MAPE	9,049
DW	2,177
Cp	3,000
AIC	1175,598
SBC	1186,002
PC	0,379

In questo caso, il 63% della variabilità del Peso è spiegata dall'Altezza e dall'Età. Il resto della variabilità è dovuta ad altri effetti (altre variabili esplicative) non considerati in questa analisi.
È importante a questo punto esaminare i risultati relativi all'analisi della varianza (vedi sotto). Tali risultati ci permettono di determinare se le variabili esplicative portano o meno informazioni significative (ipotesi nulla H_0) al modello. In altre parole, è un modo per chiedersi se è corretto utilizzare la media per descrivere l'intera popolazione o se le informazioni fornite dalle variabili esplicative hanno valore o meno.
Durante l'analisi viene effettuato un test di Fisher. Dato che la probabilità corrispondente al valore F è inferiore a 0,0001, si considera un rischio inferiore allo 0,01% nell'assumere che l'ipotesi nulla (nessun effetto delle due variabili esplicative) sia errata. Pertanto, possiamo ancora concludere con sicurezza che le tre variabili portino una quantità significativa di informazioni.

Analysis of variance:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	56233,254	28116,627	199,610	< 0,0001
Error	234	32960,761	140,858		
Corrected Total	236	89194,015			

Computed against model Y=Mean(Y)

Le seguenti tabelle mostrano i risultati della Sum of Squares analysis di tipo I e di tipo III. Ricordiamo che questi risultati servono ad indicare se una determinata variabile porta informazioni significative al modello (una volta che tutte le variabili sono già incluse nel modello stesso).

Type I Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Age	1	35924,084	35924,084	255,038	< 0,0001
Height	1	20309,170	20309,170	144,182	< 0,0001

Type III Sum of Squares analysis:

Source	DF	Sum of squares	Mean squares	F	Pr > F
Age	1	2678,226	2678,226	19,014	< 0,0001
Height	1	20309,170	20309,170	144,182	< 0,0001

Possiamo osservare nella tabella riportata di seguito e relativa ai dettagli sul modello, che l'intervallo di confidenza del 95% del parametro Height è molto "stretto". Inoltre il p-value per la variabile Age è più grande rispetto a quello di Height e l'intervallo di confidenza include quasi lo 0. Questo indica che l'effetto di Age è più debole rispetto all'effetto di Height.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it

L'equazione del modello è riportata sotto la tabella. Possiamo constatare che per una data altezza, l'età ha un effetto positivo sul peso: quando l'età aumenta di 1 mese, il peso aumenta di 0,23 libbre.

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-127,820	12,099	-10,565	< 0,0001	-151,657	-103,983
Age	0,240	0,055	4,360	< 0,0001	0,132	0,349
Height	3,090	0,257	12,008	< 0,0001	2,583	3,597

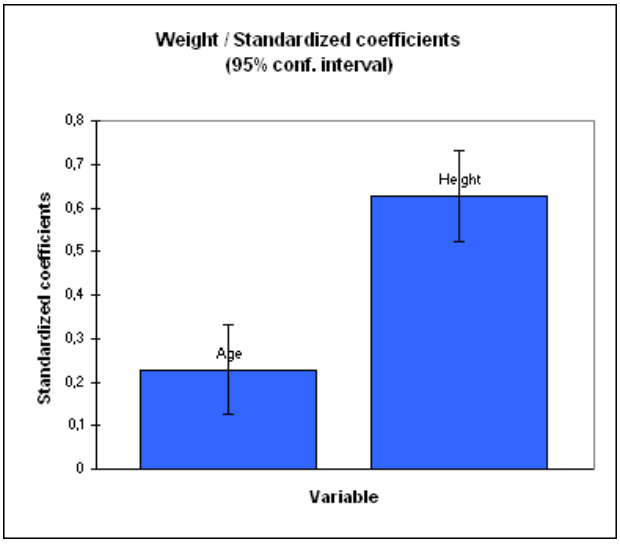
Equation of the model:

Weight = -127,819907444769+0,240274914264462*Age+3,09004803935285*Height

La tabella e il grafico sottostante riguardano i coefficienti di regressione standardizzati (spesso indicati come coefficienti beta). Consentono di confrontare in modo diretto l'influenza relativa delle variabili esplicative sulla variabile dipendente e la loro significatività.

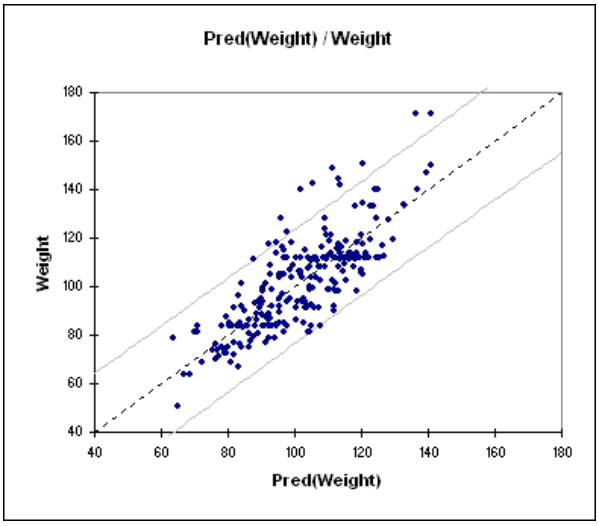
Standardized coefficients:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Age	0,228	0,052	4,360	< 0,0001	0,125	0,331
Height	0,627	0,052	12,008	< 0,0001	0,524	0,730

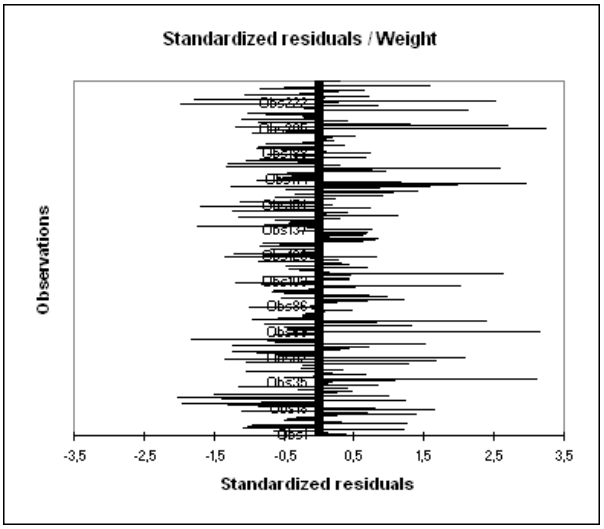


Per l'analisi dei residui, abbiamo anche in questo caso utilizzato il DataFlagger di XLSTAT per evidenziare i residui che non si trovano nell'intervallo [-1,96; 1,96].
Su 237, possiamo identificare 15 residui fuori dall'intervallo [-1,96; 1,96], ossia il 6,3% dei residui invece del 5%.
Il seguente grafico ci consente di confrontare i valori previsti con i valori osservati.

GMSL S.r.l.
Via Giovanni XXIII, 21 – 20014 – Nerviano (MI)
Tel: 0331.587511 – Fax: 0331.415772
Email: contact@gmsl.it – Web: www.gmsl.it



L'istogramma dei residui ci consente di visualizzare rapidamente i residui che si trovano fuori dal range [-2, 2].



Conclusioni

In conclusione, l'altezza, l'età e il genere ci permettono di spiegare il 63% della variabilità del peso. Anche utilizzando la regressione lineare multipla una quantità significativa di informazioni non viene spiegata dal modello. Utilizzando invece un'analisi ANCOVA, potrebbe essere aggiunta la variabile "genere" al modello per migliorare la qualità del fitting.