Salvatore Ingrassia
Carmela Senatore

Laboratorio di Statistica I

Guida alle Attività

Facoltà di Economia, Università della Calabria
Corso di Laurea in Statistica
Anno Accademico 2002-2003

Indice

Capitolo 1

Modelli di regressione con SPSS

1.1 Correlazione

Nel capitolo relativo ai modelli di regressione è stao introdotto il concetto di legame tra i valori di due o più variabile; si è visto come calcolare indici che indicano la vicendevole dipendenza tra le variabili e indici che forniscano il verso del legame stesso. E' stato introdotto uno specifico indice per la misura della relazione tra due caratteri quantitativi, il coefficiente di correlazione lineare di Bravis-Pearson, in quella sede è stato chiarito che il coefficiente deve essere interpretato esclusivemente come misura di interdipendenza lineare che assume valori compresi tra -1 e 1 se e solo se i due valori sono legati da una relazione di tipo lineare, cioè, se e solo se noto il valore assunto da uno dei due caratteri, il valore assunto dall'altro carattere risulta essere univocamente detrminato dal legame lineare[1]. Il problema relativo al legame tra levariabili è stato anche affrontato durante l'analisi delle tavole di contingenza (vedi capitolo 5) e si è visto come in funzione del tipo di variabili analizzate è possibile calcolare diversi indici, tra i quali è anche previsto, ovviamente, il coefficiente di correlazione.

In questo paragrafo sarà trattata l'analisi dei legami attraverso un altro comando di spss le CORRELAZIONI (Questo comando può essere applicato solo a variabili di tipo quntitativo, al contrario del comando tavole di contingenza applicabile ad ogni tipo di variabile.), anche ad esso si accede dal menù analizza. Come si vede dalla figura 1.1, al sottomenù correlazioni corrispondo tre opzioni:

- correlazione bivariata: questa procedura consente di calcolare coefficiente di correlazione di Pearson, rho di Spearman e tau-b di Kendall con i rispettivi livelli di significatività;
- correlazione parziale: questa procedura consente di calcolare i coefficienti di correlazione parziale che descrivono la relazione lineare tra due variabili controllando gli effetti di una o più variabili aggiuntive;
- distanze: questa procedura consente di calcolare una grande varietà di statistiche in base alle similarità o alle dissimilarità (distanze), sia considerando coppie di variabili, sia considerando coppie di casi.

[1] Nelle situazioni reali, una tale condizione si realizza molto raramente, molto più frequenti sono, invece, le situazioni in cui è ipotizzabile un qualche legame tra i due caratteri e nelle quali la relazione lineare viene assunta come misura di prima approssimazione del lehame stesso.

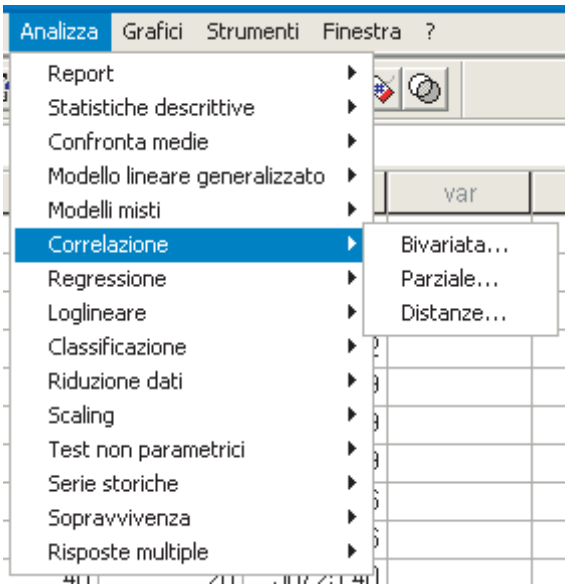

Figura 1.1: *Menù correlazioni.*

Il set di dati utilizzato per gli esempi è datinordsud.dat reperibile sul sito ————————; è un campione di 100 individui fra i quali sono sate rilevate le seguenti variabili: listruzione, in anni; la residenza, 0 = sud - 1 = nord; letà, in anni; il reddito, in milioni di Lire annue.

1.1.1 Correlazione Bivariata

Optando per la voce correlazioni bivariate dal sottomenù correlazioni, sarà visualizzata la finestra in figura 1.2. Come ormai è chiaro nel riaquadro a sinistra si trova l'elenco alfabetico di tutte le variabili del file, selezionando quelle di cui si intende conoscere le relazioni e poi cliccando sulla freccia esse sono trasferite nella casella a destra, selezionando i comandi d'interesse e poi su ok i risultati dell'elaborazione saranno visualizzati sulla finestra di uotput.

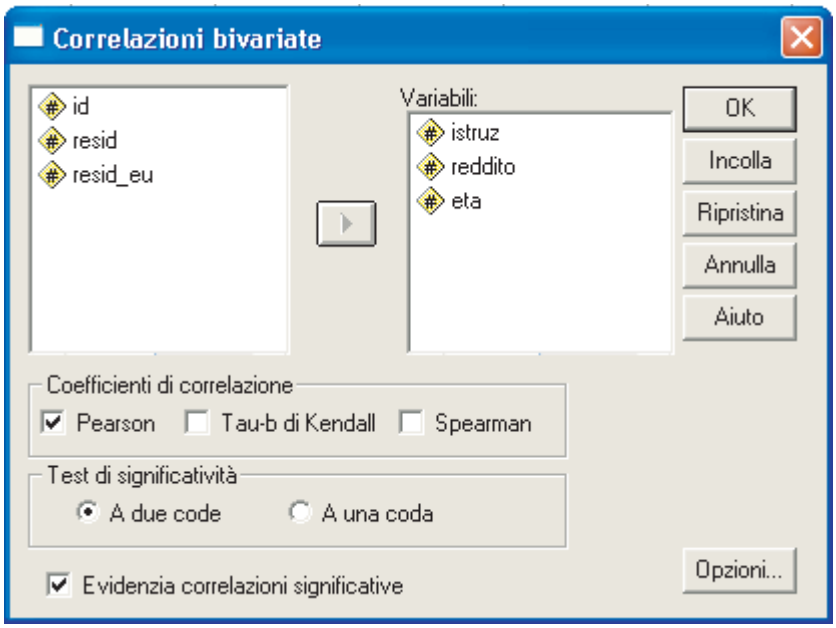

Figura 1.2: *Correlazioni bivariate.*

Sempre nella finestra correlazioni bivariate in basso si trovano due riquadri e un pulsante per le opzioni; nel primo riquadro si può selezionare il tipo di coefficiente di correlazione che si vuole calcolare, nel secondo riquadro il tipo di test di significatività che si vuole effettuare, mentre attraverso il pulsante opzioni si accede ad una sotto finestra (vedi figura 1.3, nella quale si possono scegliere le staitistiche da calcolare e come trattare i dati mancanti, da questa finestra si può decidere se calclare o meno la covarianza. Inoltre, nella finestra principale si può decidere se far evidenziare le correlazioni più significative.

Figura 1.3: *Correlazioni bivariate: opzioni.*

In figura 1.4 è visualizzato un esempio di output per le variabili reddito, istruzione ed età relative al data set datinordsud.dat.

Correlazioni

Statistiche descrittive

	Media	Deviazione std.	N
ISTRUZ	8.32	5.508	100
REDDITO	49.90	25.369	100
ETA	40.99	10.355	100

Correlazioni

		ISTRUZ	REDDITO	ETA
ISTRUZ	Correlazione di Pearson	1	.857**	.011
	Sig. (2-code)	.	.000	.913
	Somma dei quadrati e dei prodotti incrociati	3003.760	11854.200	62.320
	Covarianza	30.341	119.739	.629
	N	100	100	100
REDDITO	Correlazione di Pearson	.857**	1	.045
	Sig. (2-code)	.000	.	.658
	Somma dei quadrati e dei prodotti incrociati	11854.200	63713.000	1163.900
	Covarianza	119.739	643.566	11.757
	N	100	100	100
ETA	Correlazione di Pearson	.011	.045	1
	Sig. (2-code)	.913	.658	.
	Somma dei quadrati e dei prodotti incrociati	62.320	1163.900	10614.990
	Covarianza	.629	11.757	107.222
	N	100	100	100

**. La correlazione è significativa al livello 0,01 (2-code).

Figura 1.4: *Esempio di output per la correlazione bivariata.*

1.1.2 Correlazione Parziale

Selezionando la voce correlazioni parziali sarà visualizzata la seguente finestra (vedi figura 1.5:

Figura 1.5: *Correlazioni Parziali.*

la struttura della finestra, rispetto a quella relativa alla bivariata, cambia solo per il campo relativo alla variabile della quele si vogliono rimuovere gli effetti e all'impossibilità della scelta del coefficiente di correlazione da calcolare.

In figura **??** è visualizzato un esempio di output per le variabili reddito, istruzione, età con l'eliminazione degli effetti della variabile residenza.

```
- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S  - - -

Controlling for..    RESID

               ISTRUZ    REDDITO        ETA

ISTRUZ          1.0000      .8966     -.0099
               (    0)     (   97)    (   97)
               P= .       P= .000    P= .923

REDDITO          .8966     1.0000     -.0180
               (   97)     (    0)    (   97)
               P= .000    P= .       P= .859

ETA             -.0099     -.0180     1.0000
               (   97)     (   97)    (    0)
               P= .923    P= .859    P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Figura 1.6: *Esempio di output di correlazioni parziali.*

1.2 La regressione

La regressione, come già discusso ne paragrafo—, è una procedura statistica che stima una relazione lineare tra una variabile dipendente e un insieme di varaibili indipendenti, cercando di spiegare la variabilità della variabile dipendente in termini di quella delle esplicative e derivando, quindi, una interpretazione casuale quantitativa tra questi ultimi e la variabile dipendente stessa.

In spsss per effettuare la regerssione è necessario selezionare dalla barra dei menù, il menù analizza, da quest'ultimo la voce regressione, quindi la voce lineare (vedi figura 1.7).

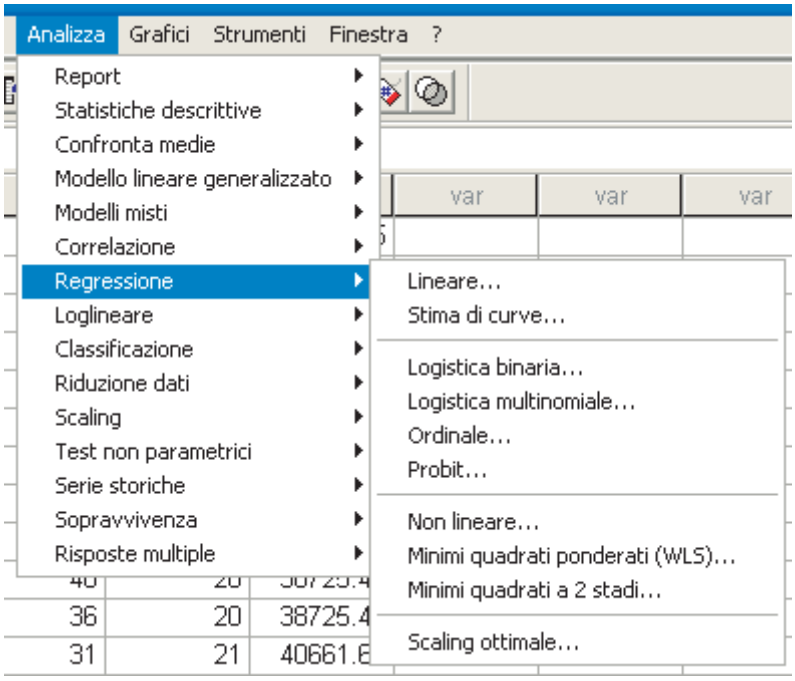

Figura 1.7: *Menù regressione.*

In questo modo compare la finestra contenete la lista delle variabili (vedi figura 1.8), in questa finestra è previsto un campo per l'inserimento della variabile dipendete e un campo per l'inserimento delle variabii indipendenti, inoltre per quest'ultimo campo è prevista la scelta del metodo di inserimento delle variabili durante a costruzione del modello; i metodi proposti sono:

- per passi: ad ogni passo, la variabile indipendente non presente nell'equazione che ha la più bassa probabilità di F viene inserita, se quella probabilità è sufficientemente piccola. Le variabili già presenti nell'equazione di regressione vengono rimosse se la loro probabilità di F diviene sufficientemente elevata. Il metodo termina quando nessuna variabile rispetta il criterio di inserimento o quello di rimozione;

- rimozione: è una procedura per la selezione di variabili in cui tutte le variabili di un blocco sono rimosse in un solo passo;

- indietro: una procedura di selezione di variabili nella quale tutte le variabili vengono inserite nell'equazione e poi rimosse sequenzialmente. La variabile con la più bassa correlazione parziale rispetto alla variabile dipendente viene considerata la prima da rimuovere. Essa viene rimossa se soddisfa il criterio di eliminazione. Dopo che la prima variabile è stata rimossa, la variabile con la più bassa correlazione parziale tra quelle rimaste nell'equazione, viene considerata come la prossima da eliminare. La procedura termina quando non ci sono più variabili nell'equazione che soddisfano il criterio di rimozione;

- avanti: una procedura di selezione delle variabili nella quale le variabili vengono inserite in modo sequenziale all'interno del modello. La prima variabile da inserire nell'equazione è quella

con la più elevata correlazione positiva o negativa con la variabile dipendente. Questa variabile viene inserita nell'equazione solo se soddisfa il criterio di inserimento. Se è stata inserita la prima variabile, viene considerata come successiva la variabile indipendente non presente nell'equazione che ha la più elevata correlazione parziale. La procedura termina quando non ci sono più variabili che soddisfano il criterio di inserimento.

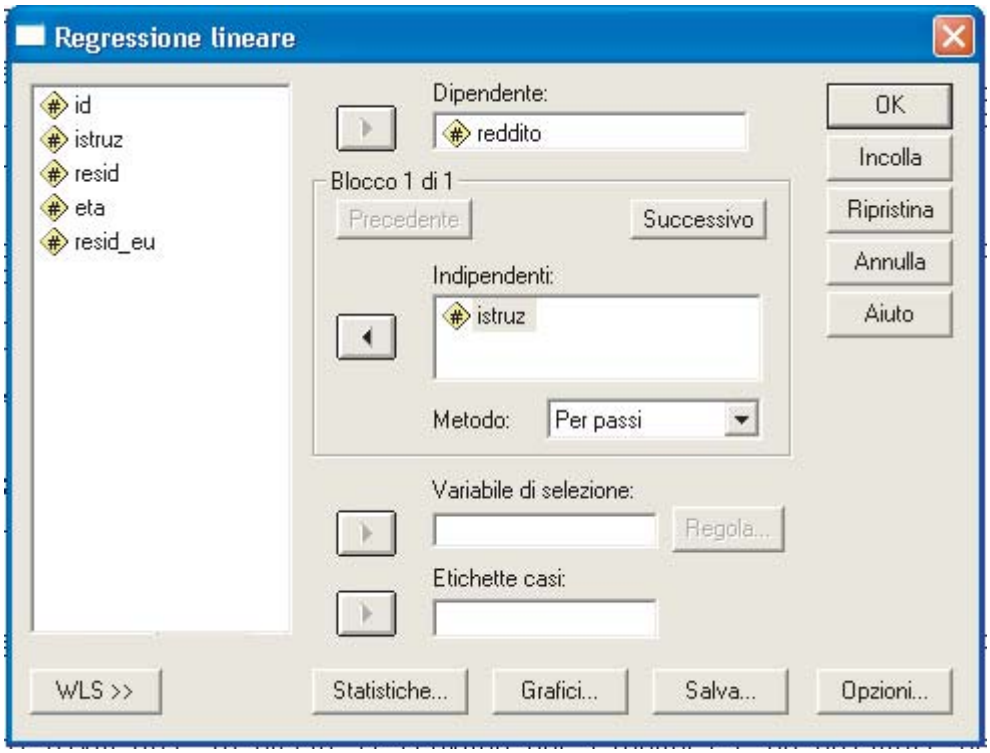

Figura 1.8: *regressione.*

Inoltre, in questa finestra, si può inserire una variabile di selezione per limitare l'analisi ad un sottoinsieme di casi che assumono un particolare valore per la variabile scelta.

Infine sempre nella finestra regressione in basso a destra si trovano quattro pulsanti: STATISTICHE, GRAFICI, SALVA, OPZIONI, cliccando si aprono le relative finestre di dialogo.

Nella finestra statistiche (vedi figura 1.9) è possibile selezionare i comandi per il calcolo di coefficienti di regressione, dei residui, statistiche descrittive, correlazioni, test di collinearità, adattamento del modello.

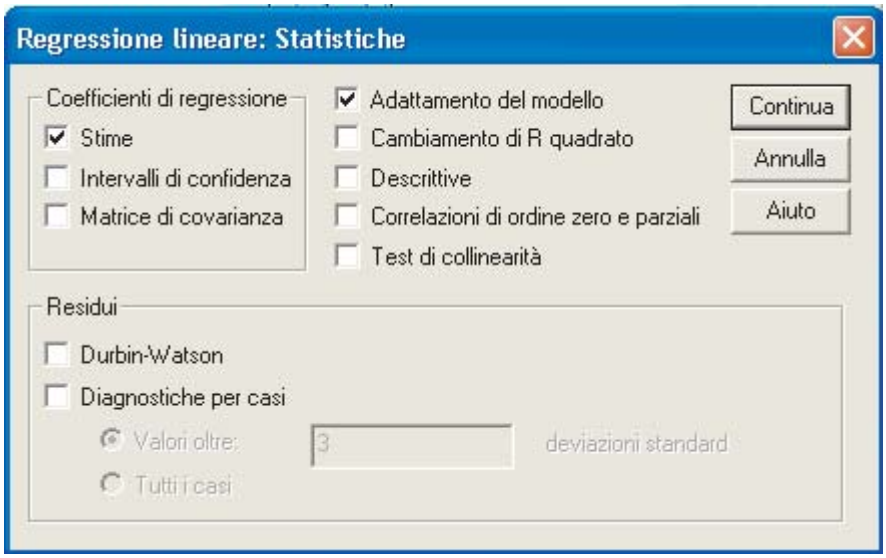

Figura 1.9: *Regressione lineare: statistiche.*

Dalla finestra grafici (vedi figura 1.10) si impostano le coordinate dei grafici a dispersione e i grafici dei residui.

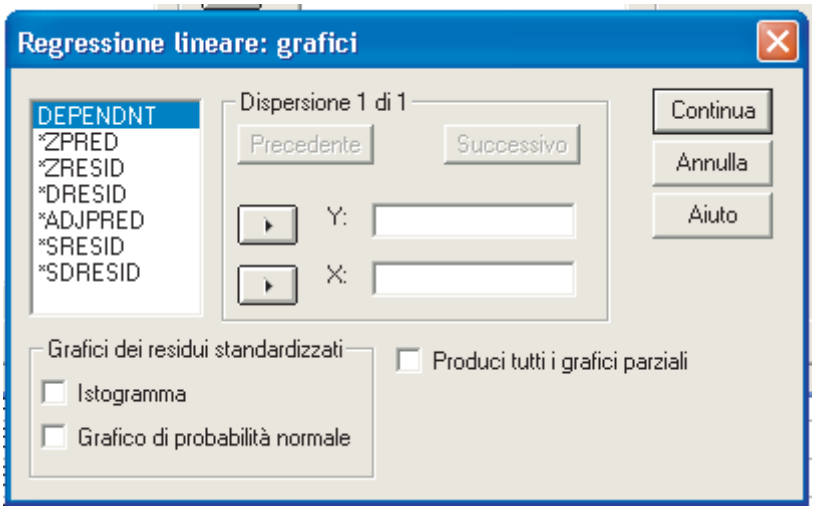

Figura 1.10: *Regressione lineare: grafici.*

La finestra salva permette di salvare nello stesso file o in un file differente .valori previsti, distanze, statistiche d'influenza e residui (vedi figura 1.11).

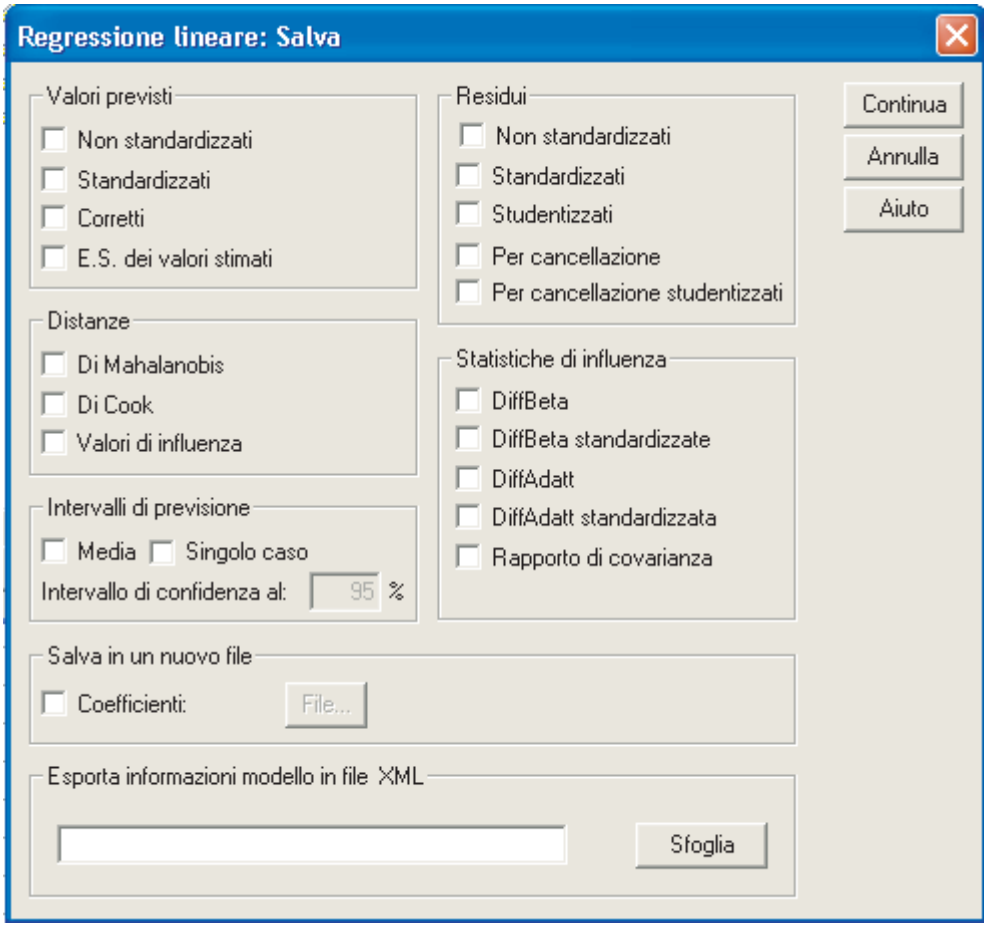

Figura 1.11: *Regressione lineare: salva.*

Infine, la finestra opzioni permette di gestire i criteri per i metodi di accettazione e rifiuto durante il processo di inserimento delle variabili e li trattamento dei dati mncanti (vedi figura 1.12).

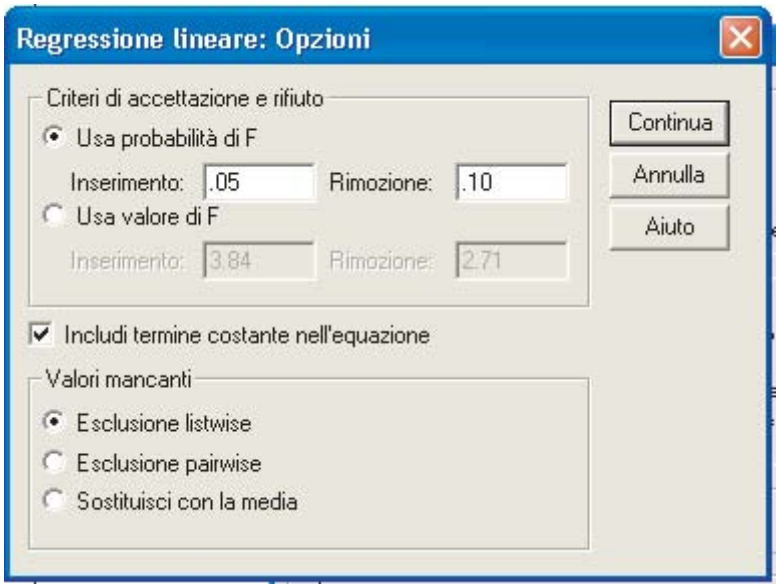

Figura 1.12: *Regressione lineare: opzioni.*

Nelle figure 1.13, 1.14, 1.15, e 1.16 si osservano grafici e dati ottenuti dall'esecuzione della regressione lineare impostando come variabile dipendente il reddito e variabile indipendente l'istruzione.

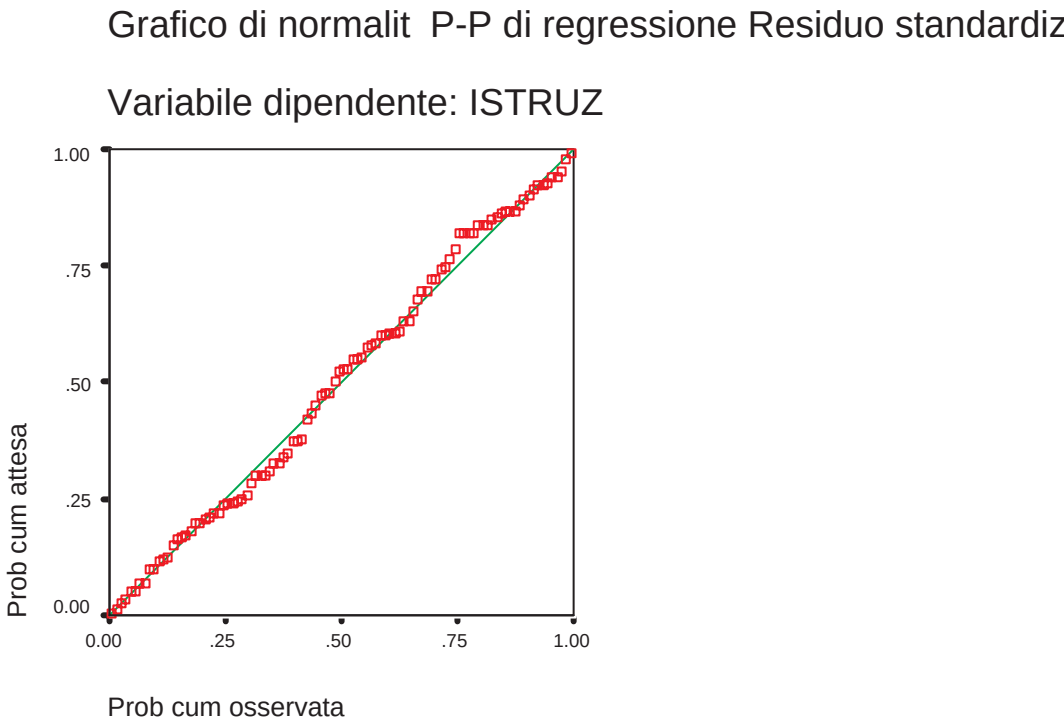

Figura 1.13: *Probability Plot.*

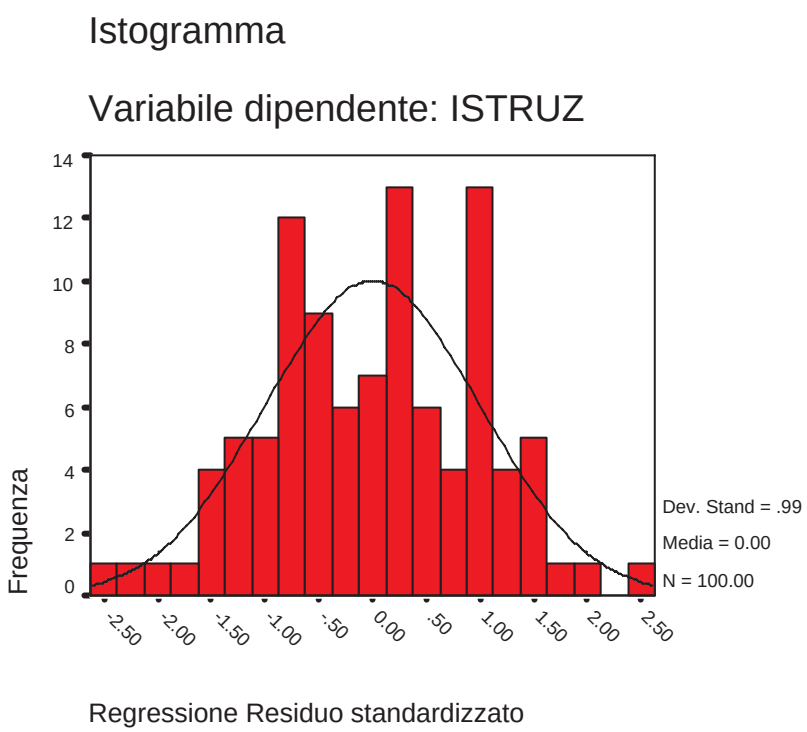

Figura 1.14: *Istogramma.*

Regressione

Variabili inserite/rimosse[b]

Modello	Variabili inserite	Variabili rimosse	Metodo
1	REDDITO[a]	.	Per blocchi

a. Tutte le variabili richieste sono state inserite

b. Variabile dipendente: ISTRUZ

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	.857[a]	.734	.732	2.854

a. Stimatori: (Costante), REDDITO

b. Variabile dipendente: ISTRUZ

ANOVA[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	2205.548	1	2205.548	270.785	.000[a]
	Residuo	798.212	98	8.145		
	Totale	3003.760	99			

a. Stimatori: (Costante), REDDITO

b. Variabile dipendente: ISTRUZ

Figura 1.15: *Esempio di output di regressione.*

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	-.964	.632		-1.525	.130
	REDDITO	.186	.011	.857	16.456	.000

a. Variabile dipendente: ISTRUZ

Statistiche dei residui[a]

	Minimo	Massimo	Media	Deviazione std.	N
Valore atteso	1.83	21.36	8.32	4.720	100
Valore atteso std.	-1.376	2.763	.000	1.000	100
Errore standard dei valori attesi	.285	.842	.388	.113	100
Valore atteso corretto	1.73	21.68	8.32	4.734	100
Residuo	-7.41	6.87	.00	2.839	100
Residuo std.	-2.596	2.407	.000	.995	100
Residuo stud.	-2.609	2.424	-.001	1.004	100
Residuo cancellato	-7.49	6.97	.00	2.892	100
Residuo studentizzato per cancellazione	-2.691	2.487	-.001	1.013	100
Distanza di Mahal.	.000	7.636	.990	1.373	100
Distanza di Cook	.000	.073	.009	.011	100
Valore d'influenza	.000	.077	.010	.014	100

a. Variabile dipendente: ISTRUZ

Figura 1.16: *Esempio di output di regressione.*