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

1.1 Dipendenza e Indipendenza distributiva

Si considerino ora alcune caratteristiche di una distribuzione doppia che non sono estensioni delle caratteristiche delle distribuzioni semplici. Tale problema riguarda l'analisi del *legame* fra i valori di X e quelli di Y, sia per misurare l'entità del legame stesso, sia per fare previsioni relative ai singoli casi. Si consideri la tabella a doppia entrata:

	y_1	$\dots$	y_j	$\dots$	y_k	
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1k}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ik}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_h	n_{h1}	$\dots$	n_{hj}	$\dots$	n_{hk}	n_{h0}
	n_{01}	$\dots$	n_{0j}	$\dots$	n_{0k}	N

Si dice che in una distribuzione doppia, *il carattere Y è indipendente dal carattere X* o che Y *non è connesso con* X, se le distribuzioni parziali secondo il carattere Y corrispondenti alle modalità di X sono fra loro tutte simili, cioè se, per $j = 1, \dots, h$, risulta:

$$\frac{n_{1j}}{n_{10}} = \frac{n_{2j}}{n_{20}} = \cdots = \frac{n_{ij}}{n_{i0}} = \cdots = \frac{n_{kj}}{n_{k0}} = \frac{n_{0j}}{N} .$$

In termini di frequenze relative, si ha indipendenza stocastica quando è verificata la relazione:

$$f_{ij} = f_{i0} f_{0j} .$$

Il caso opposto dell'indipendenza o connessione nulla è quello della *dipendenza perfetta*: il carattere Y dipende perfettamente da X se ad ogni modalità x_i di X è associata una sola madalità y_j di Y e se Y assume almeno due diverse modalità diverse al variare di X.

Se fra i due caratteri di una distribuzione doppia non sussiste né connessione nulla né perfetta dipendenza, sorge il problema di misurare il grado di connessione fra di essi.

Un'importante misura di distanza fra distribuzioni di frequenze è la *distanza* χ^2 di K. Pearson che è data da:

$$\chi^2 \;=\; N\left(\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{f_{ij}^2}{f_{i0}f_{0j}}-1\right)\;.$$

L'indice χ^2:

- è sempre non negativo;
- assume valori piccoli quando le frequenze osservate assumono valori prossimi a quelli che si avrebbero nel caso di indipendenza stocastica, in particolare è uguale a zero nel caso di connessione nulla;
- tende ad assumere valori molto grandi in situazioni molto "lontane" da quella di indipendenza;
- Si noti infine che la distanza del χ^2 non dipende dalle modalità dei caratteri in esame ma solo dalla loro distribuzione di frequenze, pertanto essa può essere utilizzata sia nell'analisi delle dipendenza fra caratteri qualitativi che fra caratteri quantitativi.

L'indice χ^2 è un indice di connessione assoluta, alrti due indici sono la *contingenza quadratica media*:

$$\Phi^2 \;=\; \frac{\chi^2}{N}\;.$$

e la *contingenza quadratica media relativa* di H. Cramér:

$$\varphi^2 \;=\; \frac{\Phi^2}{\min(k-1,h-1)}\;,$$

Un esempio di calcolo dell'indice χ^2 è fornito qui di seguito. Si consideri, a questo punto, la distribuzione relativa agli investimenti di 330 professionisti e si calcolino i tre indici sopra descritti. Pertendo dalla tabella delle frequenze assolute calcoliamo:

1. la tabella delle frequenze relative $\left(\frac{n_{ij}}{N}\right)$;
2. la tabella di connessione nulla $(f_{i0}\, f_{0j})$;
3. la tabella dei contributi $\left(\frac{f_{ij}^2}{f_{i0}f_{0j}}\right)$;

Tali tabelle sono rappresentate in figura 3.1.

Frequenze Relative					
	Medico	Avvocato	Commercialista	Altro	Totale
Medico	0.000	0.045	0.076	0.091	0.212
Avvocato	0.018	0.036	0.103	0.088	0.245
Commercialista	0.045	0.088	0.106	0.152	0.391
Altro	0.015	0.030	0.042	0.064	0.152
Totale	0.079	0.200	0.327	0.394	1.000
Connessione Nulla					
	Medico	Avvocato	Commercialista	Altro	Totale
Medico	0.017	0.042	0.069	0.084	
Avvocato	0.019	0.049	0.080	0.097	
Commercialista	0.031	0.078	0.128	0.154	
Altro	0.012	0.030	0.050	0.060	
Totale					1.000
Rapporti tra frequenze al quadrato e connessione nulla					
	Medico	Avvocato	Commercialista	Altro	Totale
Medico	0.000	0.049	0.083	0.099	
Avvocato	0.017	0.027	0.132	0.080	
Commercialista	0.067	0.099	0.088	0.149	
Altro	0.019	0.030	0.036	0.068	
Totale					1.043

Figura 1.1: *Schema per il calcolo degli indici*

Quindi otteniamo:

- $\chi^2 = 330(1,043 - 1) = 14,143$;
- $\Phi^2 = \frac{14,143}{330} = 0,043$;
- $\varphi^2 = \frac{0,043}{3} = 0,0143$.

1.2 Covarianza

Quando si è in presenza di caratteri quantitativi si possono misurare, oltre ad indici che indicano il grado della vicendevole dipendenza, anche indici che forniscano il verso del legame stesso. In altra parole, un indice che deve essere positivo se i valori più elevati di una variabile si associano con quelli più alti dell'altra, e viceversa, essere negativo se i valori più elevati di una variabile si associano con i valori più piccoli dell'altra; tali indici sono detti di concordanza.

Un'importante misura di dipendenza fra due variabili è la *covarianza*, la covarianza misura la linearità del legame; viene indicata con i simboli $\text{Cov}(X, Y)$ e σ_{xy}.

La covarianza è positiva se nella sommatoria prevalgono lewcovariazioni positive, è negativa se prevalgono le covariazioni negative ed è nulla se si hanno covariazioni positive e covariazioni negative che si bilanciano.

Riportando in un piano cartesiano i punti di coordinate (x_i, y_i), si porta l'origine del sistema di riferimento nel punto di coordinate $(\bar{x}; \bar{y})$ si osseva che:

- la covarianza positiva informa che prevalgono i punti nel primo e terzo quadrante;
- la covarianza negativa informa che prevalgono i punti nel secondo e quarto quadrante;
- la covarianza nulla informa che i punti giacciono nei quattro quadranti;

Per quanto riguarda la costruzione di un grafico del tipo appena descritto la procedura da seguire è quella descritta nel paragrafo 1.4 e il tipo di grafico da scegliere e *Dispersione (x,y)*. In figura 3.2 si osserva tale grafico costruiti per le variabili colore ed intensità colore del set di sati *wine.data*.

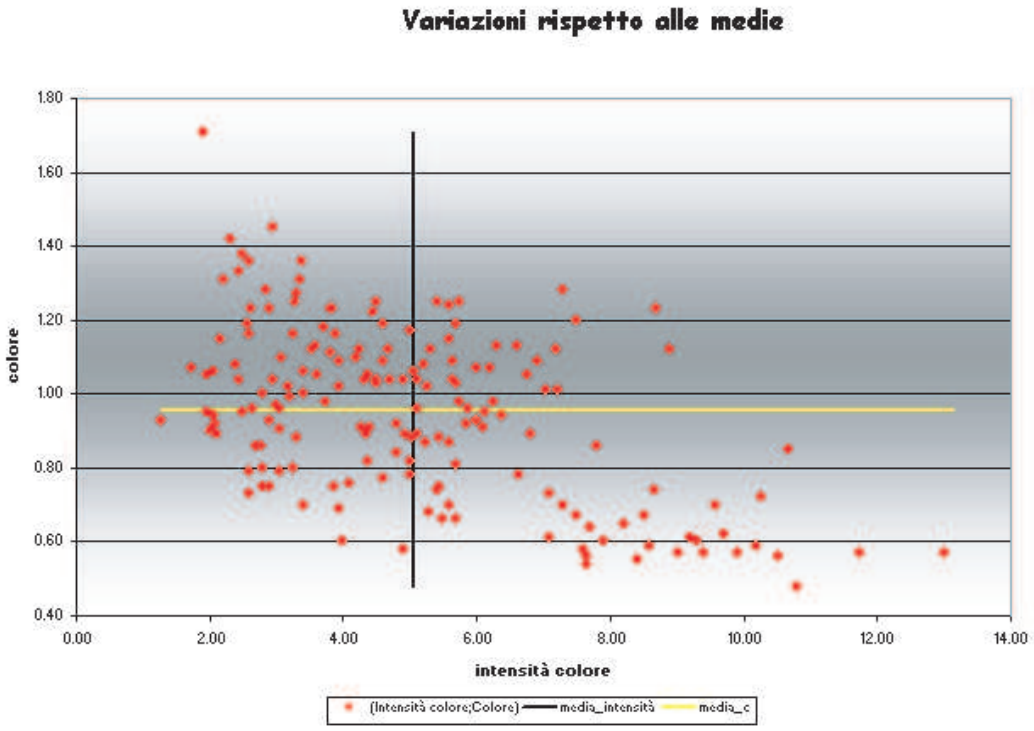

Figura 1.2: *Variazioni*

La procedura da seguire, per il calcolo della covarianza in *excel*, è quella descritta nel paragrafo 1.2 identica per il calcolo di ogni funzione. Il nome della funzione è *covarianza* e gli argomenti sono le distribuzioni delle variabili coinvolte nell'analisi.

Una misura normalizzata è il *coefficiente di correlazione lineare*, così ottenuta:

$$\rho \;=\; \frac{\sigma_{xy}}{\sigma_x \sigma_y}\,,$$

esso misura il legame lineare fra X e Y, e si ha:

$$-1 \;\leq\; \rho \;\leq\; 1\,.$$

Il coefficiente di correlazione lineare è indipendente da cambiamenti di unità di misura e di origine per le variabili X e Y. Il nome della funzione per il calcolo in *excel* è correlazione e gli argomenti sono le distribuzioni delle variabili coinvolte nell'analisi.

1.3 Matrici di Covarianza e Correlazione

Seguendo le procedure di analisi dei dati descritte nel paragrafo 1.5 è possibile costruire in modo immediato delle matrici di correlazione e di varianza e covarianza coinvolgendo più di due variabili.

Lo strumento di analisi da selezionare nel primo caso è correlazione e la relativi finestra di dialogo è visualizzata in figura 3.3

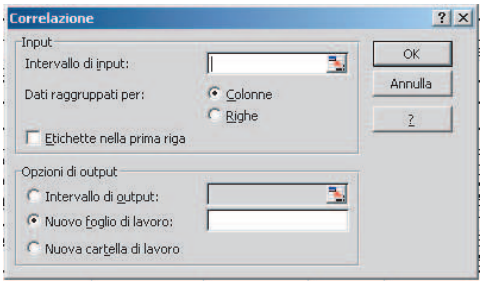

Figura 1.3: *Calcolo matrice di correlazione*

mentre, nel secondo caso lo strumento di analisi è covariaza è la finestra di dialogo è la seguente (Figura 3.4)

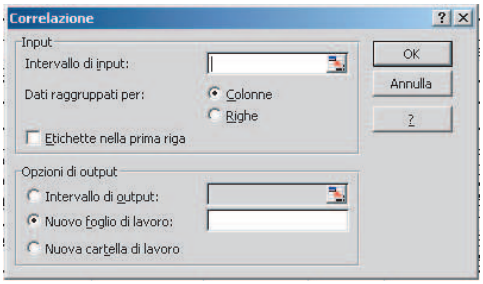

Figura 1.4: *Calcolo matrice di varianza e covarianza*

In figura 3.5 sono mostrate le due matrici costruite coinvolgendo le variabili: alcol, acido malico, magnesio, tot.fenoli, intesità colore, colore.

CORRELAZIONE						
	Alcol	*Acido Malico*	*Magnesio*	*Tot.fenoli*	*Intesità colore*	*Colore*
Alcol	1	0.094396941	0.270798	0.289101123	0.546364195	-0.07175
Acido Malico	0.094396941	1	-0.05458	-0.335166997	0.248985344	-0.5613
Magnesio	0.270798226	-0.054575096	1	0.214401235	0.199950006	0.055398
Tot.fenoli	0.289101123	-0.335166997	0.214401	1	-0.055136418	0.433681
Intesità colore	0.546364195	0.248985344	0.19995	-0.055136418	1	-0.52181
Colore	-0.071747197	-0.561295689	0.055398	0.433681335	-0.521813193	1

VARIANZA E COVARIANZA						
	Alcol	*Acido Malico*	*Magnesio*	*Tot.fenoli*	*Intesità colore*	*Colore*
Alcol	0.65535973	0.085130347	3.122238	0.146062009	1.022505676	-0.01324
Acido Malico	0.085130347	1.241004081	-0.86589	-0.233021219	0.641215496	-0.14252
Magnesio	3.122238354	-0.865887514	202.8433	1.905703194	6.583326677	0.179835
Tot.fenoli	0.146062009	-0.233021219	1.905703	0.389489032	-0.079548095	0.06169
Intesità colore	1.022505676	0.641215496	6.583327	-0.079548095	5.344255848	-0.27495
Colore	-0.013238649	-0.142520437	0.179835	0.061690343	-0.274952398	0.051951

Figura 1.5: *Matrici di varianza e covarianza*

1.4 La Regressione

La regressione è una procedura statistica che che stima una relazione lineare tra una variabile dipendente e un insieme di variabili esplicative e predittori. Lo scopo è quello di spiegare la variabilità della variabile dipendente in termini di quella delle esplicative e derivarne una interpretazione causale quantitativa tra questi ultimi e la variabile dipendente stessa.

Il fenomeno statistico di cui si vuole spiegare il comportamento in base ad una o più variabili esplicative viene definita come variabile dipendente (o risposta) e deve essere necessariamente rappresentato da una variabile statistica quantitativa continua (o discreta approssimabile ad una continua).

Nel modello di regressione compare anche un termine detto "errore" (ε) che ha una duplice valenza:

- dal punto di vista interpretativo indica che una qualsiasi relazione tra una variabile dipendente ed esplicative non può essere esatta ma contiene delle "imprecisioni" sia di tipo casuale sia dovute alle limitazioni di precisione delle misurazioni e/o all'assenza di informazioni complete sulla natura del fenomeno;

- dal punto di vista operativo consente, di stabilire la bontà dell'adattamento del modello ai dati osservati in base al confronto tra le proprietà degli errori osservati (gli scarti del modello dalla variabile dipendente) e le proprietà desiderate che il termine di errore dovrebbe possedere.

1.4.1 Regressione Lineare Semplice

L'equazione che rappresenta il modello di regressione lineare semplice è la seguente:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.1}$$

$i = 1, \ldots, n$ con n = numero totale di osservazioni; si ipotizza che ε_i sia IID (indipendente identicamente distribuite) con una Normale di media 0 è varianza σ^2.

In base al criterio di ottimizzazione che rende minima la somma dei quadrati degli scarti tra variabile dipendente e modello lineare (metodo dei minimi quadrati), cioè:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i + \varepsilon_i)^2 \tag{1.2}$$

si ottengono i valori stimati b_0 e b_1 di β_0 e di β_1. Graficamente l'equazione, rappresenta la retta "ottima" attraverso la "nuvola" di punti del diagramma di dispersione y/x: b_0 esprime lintercetta sullasse y mentre b_1 il coefficiente angolare (la pendenza) della retta.

1.4.2 Regressione Lineare Multipla

Se le variabili esplicative sono più di una, allora si è in presenza della regressione multipla e l'equazione diventa:

$$y = \beta_0 + \beta_1 x_{1i} + ... + b_p x_p + \varepsilon i \tag{1.3}$$

Si noti che in questa equazione, i coefficienti di regressione rappresentano il contributo indipendente alla previsione della variabile dipendete.

Anche in questo si ipotizza che ε_i sia IID (indipendente identicamente distribuiti) da una Normale di media 0 è varianza σ^2.

In base al criterio di ottimizzazione che rende minima la somma dei quadrati degli scarti tra variabile dipendente e modello lineare (metodo dei minimi quadrati), si ottengono i valori stimati b_0, b_1, ,b_p di $\beta_0, \beta_1, , \beta_p$. Dal punto di vista grafico l'equazione rappresenta un piano di uno spazio $p + 1$-dimensionale per cui è difficilmente visualizzabile ed interpretabile.

1.4.3 Interpretazione dei risultati

Il punto di partenza consiste nella valutazione delle proprietà dei coefficienti, corrispondenti a ciascuna variabile esplicativa, secondo le seguenti indicazioni:

1. se il p-value associato al coefficiente è superiore al livello a (solitamente 0.05) allora tale coefficiente si può considerare nullo e quindi l'apporto della esplicativa ad esso associata è non significativo nei confronti della variabile dipendente; questo regressore può essere rimosso dal modello;

2. se il p-value associato al coefficiente è inferiore al livello a (solitamente 0.05) allora l'apporto alla spiegazione della variabilità della variabile dipendente nei confronti della dipendente è significativo ed in particolare:

   - o se il segno del coefficiente è maggiore di zero allora esiste una relazione positiva tra le variabili cioè se aumenta la variabile dipendente aumenta anche la variabile indipendente;
   - o se il segno del coefficiente è minore di zero allora esiste una relazione negativa tra le variabili cioè se aumenta la variabile dipendente diminuisce invece la variabile indipendente;

- o il valore assoluto del coefficiente misura la variazione che subisce la variabile dipendente a causa di una variazione unitaria della variabile indipendente;

1.4.4 Valutazione sulla bontà del modello

Indici

Il primo elemento da valutare è il valore di s, cioè della stima dello errore standard s associato al termine di errore ε: tanto basso è questo valore tanto capace è stato il modello a spiegare la variabilità della variabile dipendente.

Il secondo elemento da valutare è la quantità di varianza che il modello riesce a catturare; in base alla scomposizione della devianza totale data da:

$$DevT = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{1.4}$$

indicando con $DevR$ la devianza dei valori *teorici*,

$$DevR = \sum_{i=1}^{n} (y_i^* - \overline{y})^2 \tag{1.5}$$

e con $DevE$ la devianza degli scarti tra valori empirici e valori teorici,

$$DevE = \sum_{i=1}^{n} (y_i^* - y_i)^2 \tag{1.6}$$

si ottiene la relazione:

$$DevT = DevR + DevE \tag{1.7}$$

è possibile, quindi, calcolare *l'indice di determinazione*, R^2, e il *coefficiente di determinazione multipla corretto* $\tilde{R^2}$:

$$R^2 = \frac{DevE}{DevT} \tag{1.8}$$

$$\tilde{R^2} = R^2 - \frac{K}{N-K-1} 1 - R^2 \tag{1.9}$$

quest'ultimo considera il numero delle K variabili esplicative in relazione al numero N delle osservazioni.

Questi indici variano tra 0 e 1 e stimano la proporzione di varianza spiegata dai predittori, quindi un valore per questi coefficienti è 1, poichè più elevati sono il valori di R tanto più le esplicative sono state capaci di catturare la variabilità della dipendente.

Analisi della varianza

Calcolando il rapporto F dato da:

$$F = \frac{DevR/p}{DevE/(n-p)} \tag{1.10}$$

si ottiene un test F che valuta la bontà globale del modello: se il p-value è inferiore al livello a (solitamente 0.05) allora l'apporto del modello di regressione alla spiegazione della variabilità della variabile dipendente nei confronti della dipendente è significativo[1]

Analisi dei residui

Se i residui, cioè gli scarti della variabile dipendente dal modello calcolato con i parametri ottenuti con il metodo dei minimi quadrati,

$$e_i = y_i - b_0 - b_1 x_i \tag{1.11}$$

rispettano una serie di condizioni allora il modello si pu dire correttamente specificato.
Le condizioni sono:

1. media pari a zero;
2. distribuzione (almeno approssimativamente) secondo la variabile casuale normale;
3. nessuna relazione con i valori della variabile dipendente.

1.5 La Regressione in *Excel*

Excel prevede diverse funzioni per la costruzione dei nodelli di regressione ed, in particolare esse sono:

- intercetta;
- pendenza;
- reg.log;
- regr.lin;
- tendenza;
- crescita;
- previsione.

In questo contesto sarà trattata solo la fuzione[2] "regr.lin" e si analizzeranno due esempi uno riguardante la fuzione lineare semplice ed uno la regressione lineare multipla.

[1] Il test F `e associato allipotesi nulla che tutti i coeffi cienti della regressione (media esclusa) siano pari a zero.
[2] In appendice sono riportate le funzionalit`a e la sintassi sulle altre funzioni

Funzione Regr.Lin

Questa funzione calcola le statistiche per una linea utlizzando il metodo dei minimi quadrati per calcolare la retta che meglio rappresenta i dati e restituisce una matrice che descrive la retta. In figura 3.6 si osserva il tipo di matrice standard restituito dalla funzione (coeff. di regr; errori standard per i coeff.; coefficiente di determinazione; errore standard della stima...):

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

Figura 1.6: *Risultati forniti dalla funzione Regr.lin.*

La procedura per il calcolo della suddetta funzione è la seguente: dalla barra dei menù standard cliccare sull'icona f_x (funzioni)[3] e quindi scegliere tra l'elenco delle *categorie* delle funzioni disponibili *statistiche* e dal *tipo di funzione* la funzione regr.lin; a questo punto si aprirà una finestra nella quale si dovranno inserire (vedi figura 3.7:

- l'insieme dei valori y gi noti (variabile dipendente);
- l'insieme dei valori x che possono essere già noti (variabila/i indipendente/i);
- *cost* un valore logico che specifica se la costante deve essere uguale a zero, impostando cost a vero la costante è calcolata secondo la normale procedura;
- *stat* un valore logico che specifica se restituire statistiche aggiuntive di regressione[4].

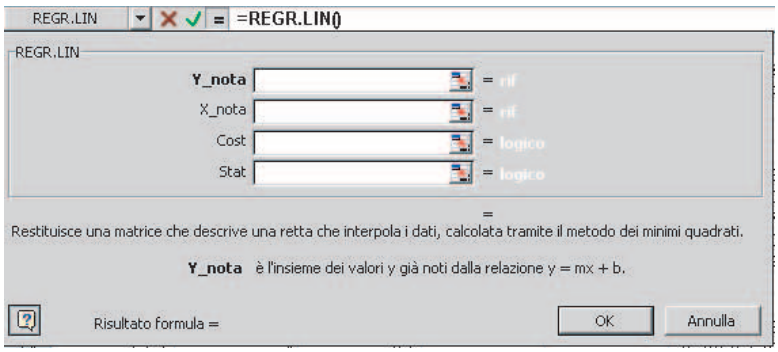

Figura 1.7: *Inserimento dati: funzione regr.lin*

Un'unica accortezza deve essere posta durante il valcolo di questa funzione, dal momento che essa restituisce una matrice di valori, deve essere immessa come formula in forma di matrice. Dopo

[3]Lo stesso risultato si ottiene selezionando dalla barra dei men`u, il men`u inserisci quindi funzioni.
[4]In appendice si trova una derragliata descrizione delle stitistiche aggiuntive.

aver confermato l'operazione attraverso la selezione del pulsante ok è quindi necessario, per poter calcolare tutti i valori, digitare CTRL+MAIUSC+INVIO.

Quando si ha una sola variabile indipendente i valori di pendenza (coeff.angoare) e intercetta posso rapidamente essere calcolati con le seguenti formule:

- INDICE(REGR.LIN(Y;X);1)
- INDICE(REGR.LIN(Y;X);2)

i valori ottenuti con l'applicazione di queste formule sono gli stessi contenuti nella prima riga della matrice di out-put dela funzione regr.lin.

A questo punto diventa banale calcolare la funzione di regressione, avendo i coefficienti di regressione, e inoltre si può interpretazione e valutare la bontà del modello costruito. Si possono calcolare i residui confrontando i valori stimati e osservati, osservando il valore del coefficiente di determinazione si può verificare l'intensità di relazione tra le variabili e se il valore assunto dalla statistica osservata F è maggiore del valore critico[5] di F si può affermare che alti valori del coefficiente di determinazione sono casuali. Nel paragrafo 3.6 sarà illustrato come ottenere in modo automatico questo tipo di analisi.

Si consideri a questo punto il set di dati rocket.xls relativo a materiali utilizzati per la costruzione del motore di un missile. Il motore di un missile viene costruito miscelando insieme un combustibile esplosivo ed uno di origine naturale all'interno di un serbatoio metallico. La forza d'urto della miscela è un'importante caratteristica di qualità del sistema. Si ipotizza che la forza d'urto della miscela possa dipendere dall'età del combustibile di origine naturale (rispetto alla data di produzione). Vengono quindi raccolte $n = 20$ coppie di osservazioni concernenti la forza d'urto (y_i) e l'età del combustibile in settimane (x_i)e riportate nella seguente tabella:

[5]Per trovare il valore critico di F, `e suffi ciente consultare la tabella dei valori critici di F in un manuale di statistica. Per leggere la tabella, nel caso di un test a una coda, utilizzare un valore α di 0,05 e per i gradi di libert, indicati nella maggior parte delle tabelle con le abbreviazioni v1 e v2, utilizzare v1 = k e v2 = n - (k + 1), dove k il numero di variabili nell'analisi di regressione e n `e il numero di dati.

i	y_i	x_i
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

Si applichi la procedura appena illustrata ai dati di questo campione, al termine della procedura la matrice ottenuta sarà del tipo (figura **??**):

-37.1535909	2627.82
2.889106549	44.18
0.901841432	96.11
165.3767577	18
1527482.743	166254.9

Figura 1.8: *Risultati relativi al modello di regressione.*

Calcolando i valori di pendenza ed intercetta con le funzioni INDICE(REGR.LIN(Y;X);1); INDICE(REGR.LIN(Y;X);2), per l'esempio precedente, si verifica che i valori ottenuti sono identici a quelli della matrice in figura 3.8.

Infine osservando i valori di R^2 e F si verifica che tra le due variabili in questione vi è una forte relazione.

Si consideri adesso il set di dati deliverytimedata.xls. Il campione è stato raccolto da un'azienda produttrice di bevande analcoliche desidera analizzare le prestazioni del servizio di assistenza e manutenzione dei distributori automatici della propria rete distributiva. In particolare si è interessati alla previsione del tempo richiesto da un operatore per effettuare l'approvigionamento e l'ordinaria

manutenzione dei distributori automatici in un generico punto vendita (edificio, o complesso di edifici,che in genere contiene più distributori automatici). Le attività da svolgere comprendono l'approvvigionamento dei distributori con le bevande previsite, oltre ad alcune semplici operazioni di ordinaria manutenzione. Il responsabile dello studio, ha suggerito che le due variabili più importanti che influenzano il tempo richiesto per effettuare il servizio completo in ciascun punto vendita comprendono il numero di casse di prodotti (lattine o bottiglie in plastica) richiesti dal distributore e la distanza percorsa a piedi dall'operatore. Il responsabile ha raccolto 25 osservazioni concernenti le seguenti variabili:

Y : tempo richiesto per l'espletamento del servizio, in minuti (variabile dipendente)
X_1 : numero di casse di prodotti
X_2 : distanza percorsa dall'operatore, in piedi (un piede = 30,48 cm)

i	Y	X_1	X_2
1	16,68	7	560
2	11,50	3	220
3	12,03	3	340
4	14,88	4	80
5	13,75	6	150
6	18,11	7	330
7	8,00	2	110
8	17,83	7	210
9	79,24	30	1460
10	21,50	5	605
11	40,33	16	688
12	21,00	10	215
13	13,50	4	255
14	19,75	6	462
15	24,00	9	448
16	29,00	10	776
17	15,35	6	200
18	19,00	7	132
19	9,50	3	36
20	35,10	17	770
21	17,90	10	140
22	52,32	26	810
23	18,75	9	450
24	19,83	8	635
25	10,75	4	150

Si applichi, ancora una volta la procedura illustrata ai dati di questo campione, al termine della procedura la matrice ottenuta sarà del tipo (Figura 3.9):

0.01438483	1.61590721	2.34123115
0.00361309	0.17073492	1.09673017
0.95959375	3.25947345	#N/D
261.235109	22	#N/D
5550.81092	233.731677	#N/D

Figura 1.9: *Risultati relativi al modello di regressione.*

in questo secondo esempio le variabili indipendenti sono 2 e risultano anche'esse risultano avere una forte relazione con la variabile indipendete.

1.6 Analisi dati: Regressione

Utilizzando lo strumento di analisi "Regressione" disponibile nel comando di analisi dei dati, illustrato nel paragrafo 1.5, si possono ottenere in maniera immediata i risultati relativi ad un'analisi di regressione. Lo strumento da selezionare in questo caso, ovviamente, è regressione la selezione porterá alla visualizzazione della finestra di dialogo illustrata in figura 3.10.

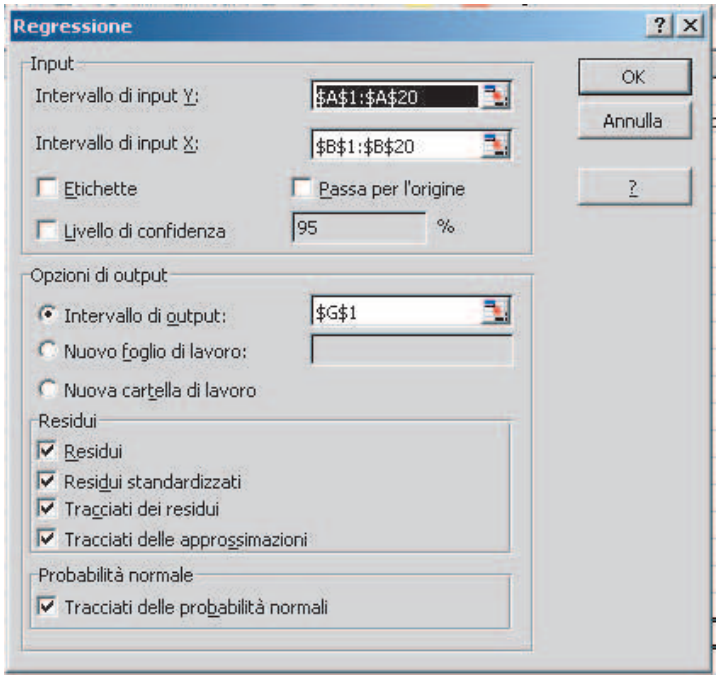

Figura 1.10: *Finestra di dialogo:Regressione.*

In questa finestra dovranno essere inseriti gli intervalli dei valori delle variabili oggetto di studio, il tipo do output che si vuole ottenere e il posizionamento dello stesso. L'output fornisce tabelle relativamente a:

- statistiche di riepilogo (vedi Figura 3.11)

OUTPUT RIEPILOGO	
Statistica della regressione	
R multiplo	0.975307
R al quadrato	0.951224
R al quadrato corretto	0.905625
Errore standard	3.502466
Osservazioni	25

Figura 1.11: *Out-put di Regressione.*

- analisi della varianza (vedi Figura3.12)

ANALISI VARIANZA					
	gdl	SQ	MQ	F	i
Regressione	2	5502.395	2751.198	224.2715	
Residuo	23	282.1471	12.26727		
Totale	25	5784.543			
Intercetta					
X_1	1.707902	0.177523	9.620737	1.58E-09	
X_2	0.016115	0.003783	4.259337	0.000295	

Figura 1.12: *Out-put di Regressione.*

- dati e residui (vedi Figura 3.13)

OUTPUT RESIDUI				OUTPUT DATI	
Oss	Previsto Y	Residui	Residui standard	Percentile	Y
1	20.97978625	-4.299786	-1.279908905	2	8
2	8.669034335	2.8309657	0.842687974	6	9.5
3	10.60285011	1.4271499	0.424816897	10	10.75
4	8.120817734	6.7591823	2.011992473	14	11.5
5	12.66468055	1.0853195	0.323064905	18	12.03
6	17.27330601	0.836694	0.249057051	22	13.5
7	5.18846807	2.8115319	0.836903172	26	13.75
8	15.33949024	2.4905098	0.741345136	30	14.88
9	74.76514607	4.4748539	1.332020957	34	15.35
10	18.28916356	3.2108364	0.95576336	38	16.68
11	38.41363932	1.9163607	0.570439311	42	17.83
12	20.54377131	0.4562287	0.135804696	46	17.9
13	10.94096574	2.5590343	0.761742688	50	18.11
14	17.69260156	2.0573984	0.612421741	54	18.75
15	22.59069513	1.4093049	0.419505004	58	19
16	29.58436006	-0.58436	-0.173945308	62	19.75
17	13.47043712	1.8795629	0.559485781	66	19.83
18	14.08250998	4.91749	1.463779568	70	21
19	5.703850145	3.7961499	1.129992451	74	21.5
20	41.4429819	-6.342982	-1.88810293	78	24
21	19.33513645	-1.435136	-0.42719424	82	29
22	57.4587034	-5.138703	-1.529627719	86	35.1
23	22.6229254	-3.872925	-1.152846075	90	40.33
24	23.89632291	-4.066323	-1.210414333	94	52.32
25	9.248876937	1.5011231	0.446836346	98	79.24

Figura 1.13: *Out-put di Regressione.*

Si possono anche ottenere interessanti rappresemtazioni grafiche quali:

- tracciato delle probabilità normali;
- tracciato delle approssimimazioni;
- tracciato dei residui.

In figura 3.14 è riportato un grafico relativemente ai residui.

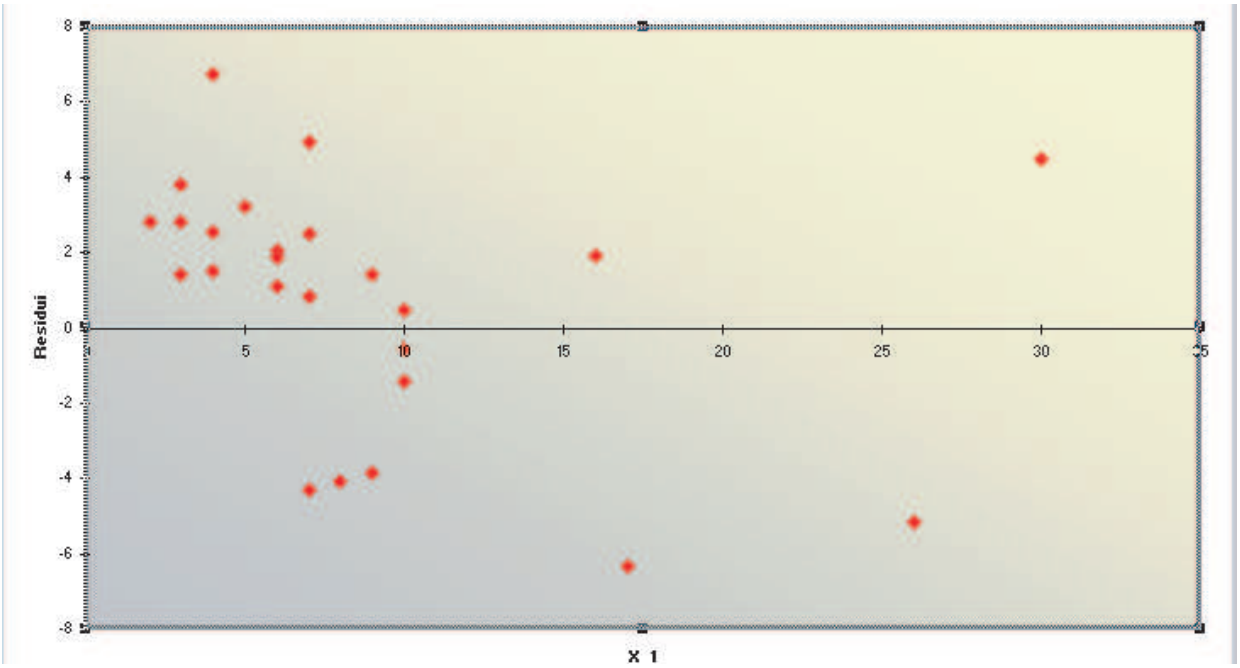

Figura 1.14: *Residui.*