3. REGRESSIONE LINEARE

3.1 Introduzione

Date due variabili quantitative X e Z, il modello di regressione lineare considera il valore medio della **variabile dipendente** Z come funzione lineare del **regressore** X. Supponiamo che la collettività oggetto di studio sia suddivisa in q gruppi distinti all'interno di ciascuno dei quali tutte le unità statistiche presentano un unico valore x di X. Le condizioni standard prevedono che la variabile $Z|x$ si distribuisca in modo normale con una media che è una funzione lineare di X

$$E(Z|x)=\beta_0+\beta_1 x$$

e con una varianza σ^2 costante per tutti i gruppi (questa è la cosiddetta varianza within, varianza all'interno dei gruppi o varianza residua).

Per avere informazioni sui valori dei parametri β_0 e β_1 e sulla varianza σ^2 si estrae in modo indipendente un campione bernoulliano da ciascuno dei q gruppi.

Indicato con n ($n \geq q$) la numerosità campionaria complessiva, restano definite n variabili casuali (v.c.) indipendenti

$$Y_i = Y|x_i \ (i = 1,2,...,n)$$

"valore di Z sull'individuo che presenta la determinazione x_i di X" dove, se $n > q$, alcuni valori x_i risultano uguali fra di loro. Ogni Y_i ha una distribuzione

$$Y_i \sim N(\beta_0+\beta_1 x_i;\sigma^2) \qquad i = 1,2,...,n$$

e il modello teorico assume la forma

$$Y_i = \beta_0+\beta_1 x_i+\varepsilon_i \qquad i = 1,2,...,n$$

dove le ε_i sono n v.c. indipendenti fra di loro, ciascuna con distribuzione normale di media zero e varianza costante σ^2, che misurano l'effetto dei fattori casuali.

3.2 PROC REG

La sintassi della procedura che consente di effettuare l'analisi di regressione è

```
proc reg <opzioni>;
model <var. dipendente>=<regressore/i> / <opzioni>;
```

Le `opzioni` più comunemente usate all'interno della **`proc reg`** sono

`data = <nome DSS>`
per specificare il DSS su cui si vuole far eseguire la **`proc reg`**

`plot <var. dipendente>*<regressore>`
per ottenere lo scatter fra la variabile dipendente e ciascun regressore

Le `opzioni` più comunemente usate all'interno dell'istruzione `model` sono

`noint`
per eliminare l'intercetta (altrimenti inclusa nel modello per default).

`influence`
per l'individuazione di eventuali outliers

`cli`
per ottenere gli intervalli di confidenza a livello 1-α per il valore previsto della variabile dipendente per ciascuna osservazione

`clb`
per ottenere gli intervalli di confidenza a livello 1-α per i coefficienti del modello

`alpha =`
per assegnare il livello di significatività per gli intervalli di confidenza e i test

L'output fornito dal SAS consiste in più tabelle, nella prima delle quali è indicato il numero di osservazioni lette e di quelle usate nell'analisi

Tabella 3.2.1
Osservazioni lette e usate

Numero osservazioni lette	*n*
Numero osservazioni usate	*n*

Nella seconda tabella, analoga alla 3.3.2, sono riportati i risultati relativi a:

- scomposizione della devianza complessiva della variabile dipendente (*SST*) in devianza spiegata (*SSA*) e devianza residua (*SSE*), con *SST=SSA+SSE*;
- gradi di libertà associati
- valori della varianza spiegata (*MSA*) e della varianza residua (*MSE*);
- test *F* sulla bontà complessiva del modello (pari al rapporto *MSA*/*MSE*) e *p*-valore associato

Tabella 3.2.2
Scomposizione della varianza

Analisi della varianza					
Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Modello	h-1	*SSA*	$MSA = SSA/(h\text{-}1)$	$F_{(h\text{-}1),(n\text{-}h)} = MSA/MSE$	*p*-valore
Errore	n-h	*SSE*	$MSE = SSE/(n\text{-}h)$		
Totale corretto	n-1	*SST*			

dove *h* rappresenta il numero di parametri presenti nel modello.
Scelto un livello di significatività α per la verifica dell'ipotesi nulla H_0 che il coefficiente di determinazione lineare è pari a zero, se la probabilità risultante nell'ultima colonna della tabella precedente è minore di α l'ipotesi H_0 va rifiutata.

Successivamente il SAS riporta le seguenti informazioni (cfr. Tabella 3.2.3):

- Radice quadrata della varianza residua *MSE*
- Media della variabile dipendente ($\bar{y}$)
- Coefficiente di variazione della variabile dipendente (CV_y)
- Valore dell'indice di determinazione lineare (R^2)
- Valore dell'indice $\bar{R}^2$ (R^2 corretto), dato da $\bar{R}^2 = 1 - \frac{n-1}{n-h}\left(1-R^2\right)$, Questo indice $\bar{R}^2$ è sempre minore o uguale a R^2, viene talvolta utilizzato al posto di R^2 in quanto tiene conto del numero di regressori presenti nel modello, ma ha l'inconveniente di poter risultare negativo

Tabella 3.2.3
Statistiche della variabile dipendente e valore del coefficiente di determinazione lineare

Radice dell'MSE	$\sqrt{MSE}$	**R-quadro**	R^2
Media dip.	$\bar{y}$	**R-quadro corr**	$\overline{R}^2$
Var coeff	CV_y		

Nell'ultima tabella sono contenute le informazioni sui parametri del modello, secondo lo schema seguente

Tabella 3.2.4
Analisi dei parametri del modello

Stime dei parametri					
Variabile	**DF**	**Stima dei parametri**	**Errore standard**	**Valore t**	**Pr > |t|**
Intercept	**1**	a	$\hat{s}_a$	$\frac{a}{\hat{s}_a}$	p-valore
$\mathbf{x_1}$	**1**	b_1	$\hat{s}_{b_1}$	$\frac{b_1}{\hat{s}_{b_1}}$	p-valore
…	**…**	…	…	…	…
$\mathbf{x_{h-1}}$	**1**	b_{h-1}	$\hat{s}_{b_{h-1}}$	$\frac{b_{h-1}}{\hat{s}_{b_{h-1}}}$	p-valore

In questo caso nell'ultima colonna è riportata la somma delle probabilità isolate alla destra del valore positivo del test t e alla sinistra di $-t$ nella distribuzione t_{n-2}. Scelto un livello di significatività α, se questa probabilità è minore di α si rifiuta l'ipotesi di nullità del parametro corrispondente.
Dopo queste tabelle il SAS restituisce anche diversi grafici utili per verificare le ipotesi sottostanti il modello e per individuare eventuali outliers.

3.3 Regressione lineare semplice: un esempio numerico

Nel successivo listato del programma SAS sono indicati i valori di tre variabili rilevate su 24 diversi personal computer portatili (dati tratti da una rivista specializzata del febbraio del 2000). Nell'ordine, le variabili sono

- mod (qualitativa, da colonna 1 a colonna 21) che indica modello del pc
- y (quantitativa) che indica il prezzo in euro
- x (quantitativa) che indica la capacità del disco fisso misurata in Gigabyte

Figura 3.3.1
Listato di un programma SAS con la procedura reg

```
data a;
input mod $ 1-21 y x;
cards;
ACER 512dx cd         1729     4.3
ACER 723txv dvd       4332     10.0
ABUS 17200            2138     4.1
ABUS 17300            3111     6.4
ABUS m8300            3409     6.4
COMPAQ 1500c          2270     4.0
COMPAQ 1750           3333     6.4
COMPAQ e700           7430     18.0
MICRODATA 1239        1785     4.3
COMPASS solo          3865     6.0
DATA E. mirage        1546     2.1
DELL latitude         4002     6.4
IBM thinkpad 390e     2454     4.8
IBM thinkpad 770z     6537     14.1
IDEA PROGRESS mti     1332     3.2
IMAGE dream.m p-sc    2225     4.8
IMAGE p-slk           4952     10.0
MICRODATA 1332        2634     4.3
MICRODATA 2000        4208     6.4
MONOLITH geo focus    3464     10.0
MONOLITH geo itinera  2349     4.8
NEC ITALIA 4012       1920     4.3
TOSHIBA satellite     2008     4.0
TOSHIBA tecra         6191     10.1
;
proc reg;
model y=x;
run;
```

Le tabelle di output fornite dal SAS risultano

Tabella 3.3.1
Osservazioni lette e usate

Numero osservazioni lette	24
Numero osservazioni usate	24

Tabella 3.3.2
Scomposizione della varianza

Analisi della varianza					
Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Modello	1	54291019	54291019	148.83	<.0001
Errore	22	8025027	364774		
Totale corretto	23	62316046			

Tabella 3.3.3
Statistiche della variabile dipendente e valore del coefficiente di determinazione lineare

Radice dell'MSE	603.96519	**R-quadro**	0.8712
Media dip.	3301.00000	**R-quadro corr**	0.8654
Var coeff	18.29643		

Tabella 3.3.4
Analisi dei parametri del modello

Stime dei parametri					
Variabile	**DF**	**Stima dei parametri**	**Errore standard**	**Valore t**	**Pr > |t|**
Intercept	**1**	552.08903	256.84642	2.15	0.0429
x	**1**	414.40869	33.96853	12.20	<.0001

In base a questi risultati si può concludere che circa l'87% della variabilità complessiva della variabile "prezzo" viene spiegata dal legame lineare con la variabile "capacità del disco fisso" e, in linea con questo risultato, il test *F* ed il corrispondente test *t* associato al regressore sono entrambi altamente significativi.

Sulla base alla stima del coefficiente angolare della retta, ad un incremento unitario, di un Gigabyte, della variabile dipendente corrisponde, in media, un incremento di oltre 400 euro della variabile prezzo.

Come si è ricordato nel paragrafo precedente, il SAS restituisce anche vari grafici che, nel caso in esame, assumono la forma riportata nelle tre figure successive

Figura 3.3.2
Analisi dei residui

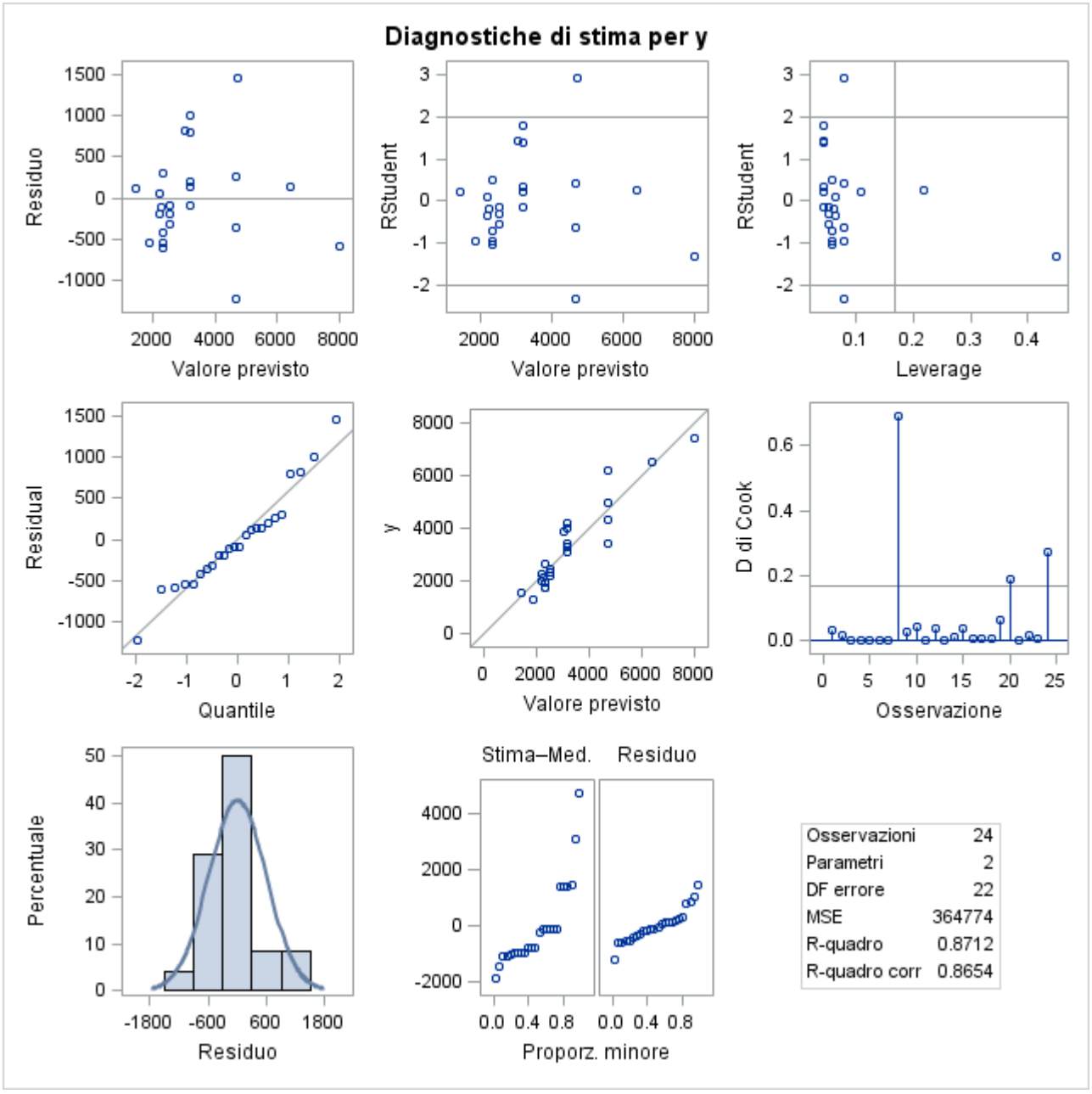


Figura 3.3.3
Grafico dei residui

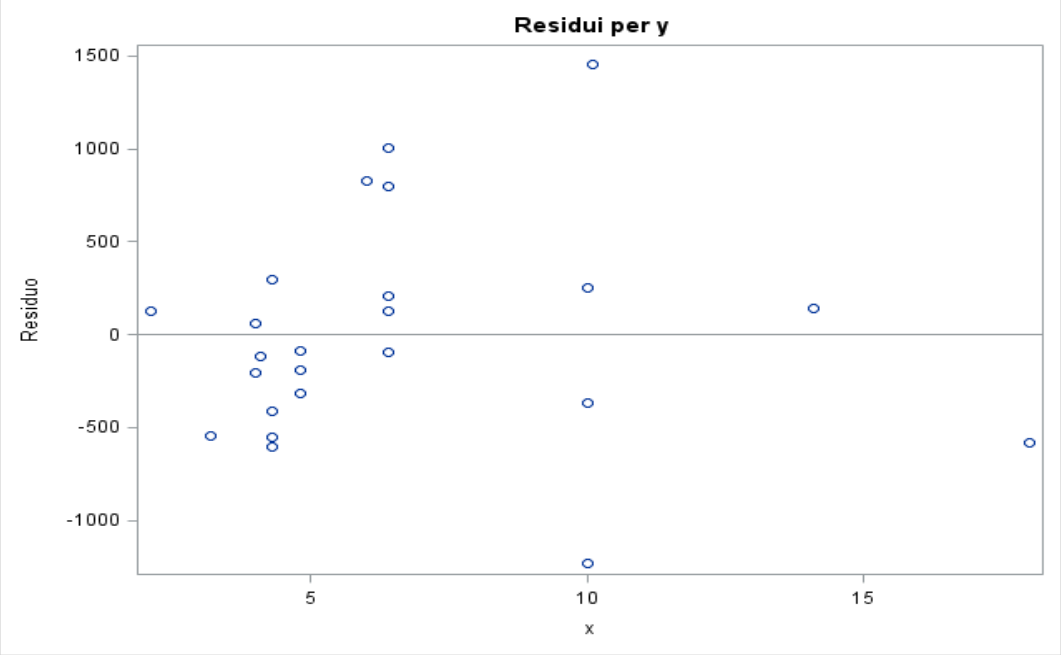
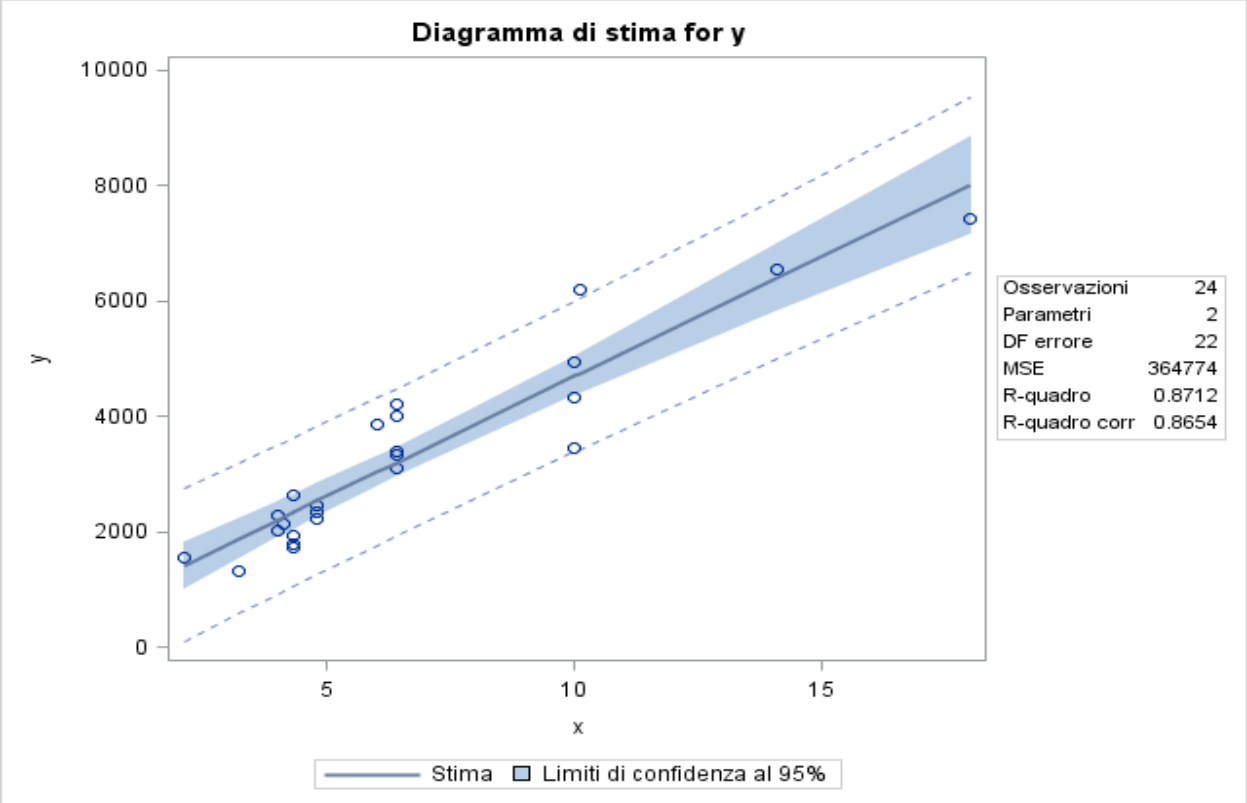


Figura 3.3.4
Bande di confidenza



Il grafico più interessante è il secondo sulla prima riga di quelli presenti nella figura 3.3.2, che mostra il diagramma di dispersione fra i valori stimati riportati sulle ascisse ed i **residui studentizzati** (che il SAS chiama "**RStudent"**) riportati sulle ordinate. Questo grafico è infatti normalmente utilizzato sia per la verifica delle ipotesi sottostanti il modello, sia per l'individuazione degli outliers.

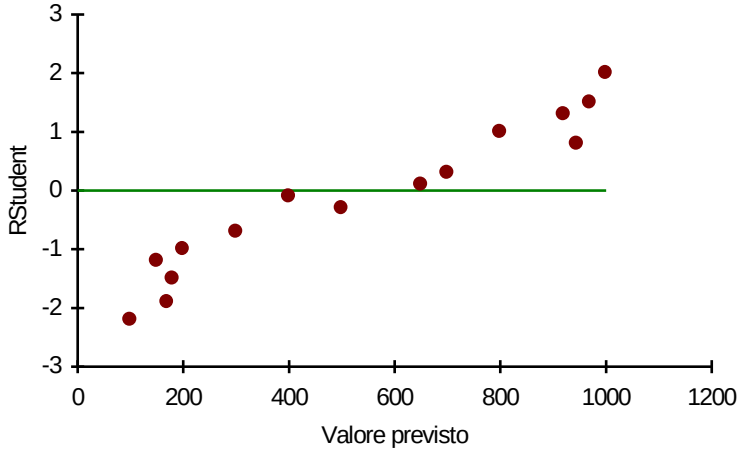
In base alle ipotesi sottostanti il modello, la distribuzione della variabile dipendente in ogni gruppo omogeneo rispetto al regressore (o ai regressori) ha una distribuzione normale con una media che è una funzione lineare del regressore (o dei regressori) e con una varianza costante.

L'**adeguatezza del modello lineare** viene verificata analizzando le caratteristiche dei residui studentizzati che corrispondono a trasformazioni dei residui (dati dalla differenza fra i valori della variabile dipendente effettivamente rilevati e quelli stimati mediante il modello di regressione). Sotto le condizioni standard questi residui studentizzati si distribuiscono come *t* di Student con $n–h–1$ gradi di libertà (g.d.l.), dove *h* corrisponde al numero di parametri presenti nel modello.

Quando i punti sul secondo grafico della prima riga della figura precedente risultano disposti in modo casuale e non mostrano delle "regolarità" si può concludere che il modello lineare è appropriato, come si osserva per il caso in esame.

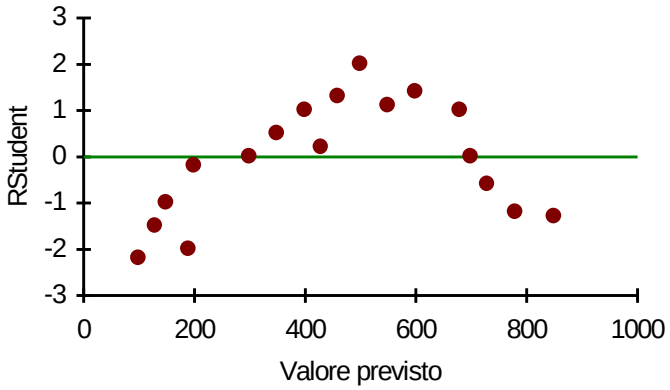
Nella figura successiva, invece, si ha un esempio di grafico con un addensamento dei residui di uno stesso segno in corrispondenza degli estremi del campo di variazione della variabile dipendente ed in questo caso il modello lineare non risulta adatto a descrivere l'eventuale relazione esistente fra le variabili.

Figura 3.3.5
Esempio di diagramma di dispersione fra valori stimati e residui studentizzati



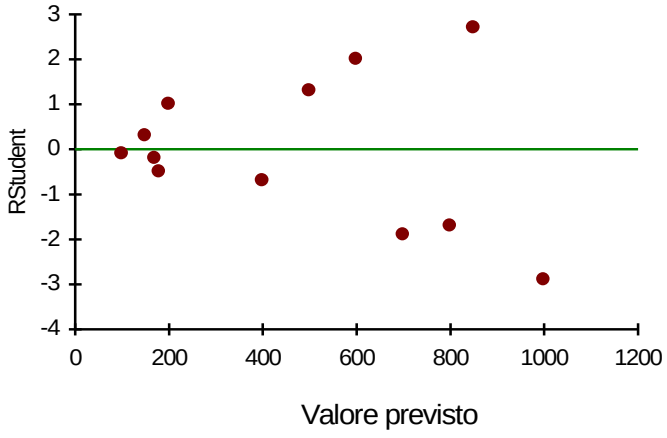
Il grafico evidenzia infatti che il modello lineare fornisce una sistematica sovrastima della variabile dipendente in corrispondenza dei suoi valori più bassi ed una sistematica sovrastima per valori alti.
Un altro esempio di grafico che evidenzia un andamento non casuale è indicato nella figura successiva.

Figura 3.3.6
Esempio di grafico dei residui studentizzati



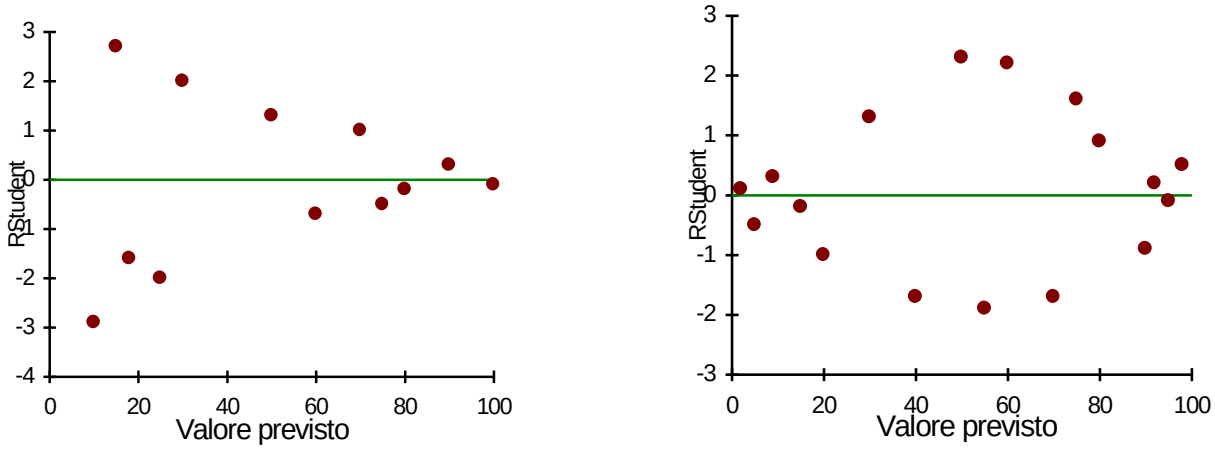
In situazioni come quelle illustrate nelle figure 3.3.5 e 3.3.6 risulta opportuno verificare se un modello diverso da quello lineare può essere più adatto a descrivere la relazione fra le variabili considerate.
Anche la condizione di **omoschedasticità** delle distribuzioni della variabile dipendente si basa sull'analisi dello stesso grafico, sempre controllando se si rilevano "regolarità" di fondo. Nella figura successiva, per esempio, i punti si dispongono "a ventaglio", a segnalare come la variabilità dei residui tenda ad aumentare all'aumentare del valore della variabile dipendente stimata sotto ipotesi di linearità.

Figura 3.3.7
Esempio di grafico dei residui studentizzati



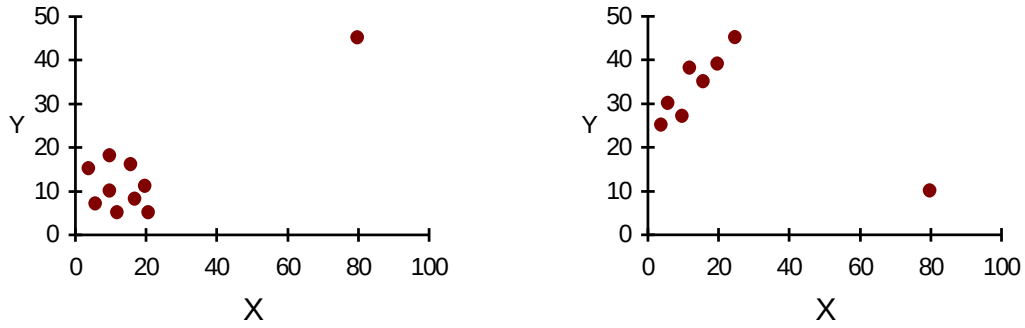
Allo stesso modo si avrebbe un chiaro segno di violazione della condizione di omoschedasticità se il grafico assumesse una delle forme indicate nelle due figure successive

Figura 3.3.8
Esempi di grafico dei residui studentizzati



che indicano una variabilità decrescente oppure una variabilità prima crescente e poi decrescente.
Un ulteriore passo dell'analisi consiste nell'individuazione dei cosiddetti **outliers** (o valori anomali), ossia di osservazioni tanto discordanti dalle altre da far sospettare che il modello ipotizzato non sia valido per le unità statistiche sulle quali sono state effettuate quelle osservazioni. Nel caso dei modelli di regressione semplice la presenza di eventuali valori anomali può essere suggerita già dall'analisi dello scatter della variabile dipendente contro il regressore, come si osserva nei due grafici della figura successiva.

Figura 3.3.9
Esempi di scatter con un outlier



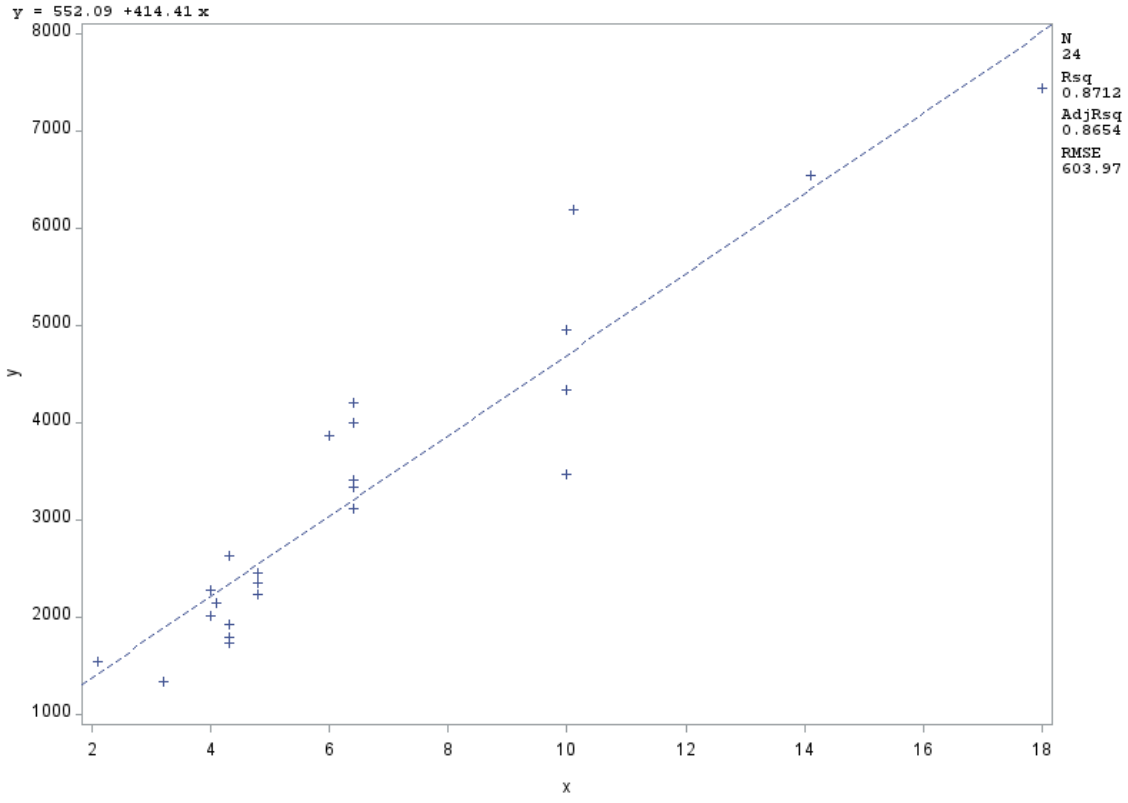
Da queste figure risulta anche chiaro come la presenza di valori anomali possa alterare in maniera sensibile i risultati dell'analisi. Nella prima figura è la presenza del punto in alto a destra a determinare un grado elevato di correlazione lineare fra le due variabili mentre, in assenza di questa osservazione, la correlazione lineare fra le due variabili sarebbe pressoché nulla. Nella seconda figura invece, la presenza del punto in basso a destra genera una correlazione che ha un segno negativo mentre, in assenza di questo punto, le due variabili risulterebbero correlate positivamente.

Nel caso in esame il diagramma di dispersione fra le due variabili, ottenuta aggiungendo l'opzione `plot <var. dipendente>*<regressore>` nella **`proc reg`**, ossia mediante le istruzioni seguenti

```
proc reg;
model y=x;
plot y*x;
run;
```

assume la forma riportata nella figura seguente, che non mostra la presenza di valori anomali evidenti.

Figura 3.3.10
Scatter delle variabili considerate nella Figura 3.3.1 e della corrispondente retta di regressione



Quando il modello è più complesso e si considera più di un regressore, l'individuazione degli eventuali outliers non può più essere effettuata semplicemente mediante gli scatter fra la variabile dipendente e ciascun regressore. In questi casi si analizza la differenza fra il valore osservato della variabile dipendente ed il corrispondente valore stimato e per valutare la rilevanza di questa differenza si fa riferimento alle distribuzioni di probabilità dei residui studentizzati, in base alla quale il valore di un residuo, con una probabilità del 95% circa, si trova all'interno dell'intervallo (-2, +2).
I possibili outliers corrisponderebbero quindi a quelle osservazioni il cui residuo studentizzato è situato all'esterno dell'area compresa fra le due linee tratteggiate che, per questo motivo, sono evidenziate dal SAS nel secondo grafico presente nella Figura 3.3.2. Nel caso preso in esame si nota la presenza di un paio di residui potrebbero segnalare la presenza di osservazioni anomale.
È però necessario tenere presente che un residuo maggiore di 2 in valore assoluto non indica necessariamente la presenza di un'osservazione anomala in quanto, per il solo effetto del caso, lavorando ad un livello di probabilità di 0.95, il 5% dei residui studentizzati presi in valore assoluto dovrebbe risultare maggiori di 2. Per questo motivo per individuare i valori anomali si utilizzano contemporaneamente diverse statistiche test e si conclude che un'osservazione è anomala se viene segnalata dalla maggioranza di questi test, che sono di seguito.
Molti dei test proposti per individuare le osservazioni anomale misurano l'influenza che ciascuna osservazione esercita sulle stime. In generale l'influenza della *i*-esima osservazione viene valutata confrontando i risultati che si ottengono in base alle *n* osservazioni originarie e quelli che si ottengono eliminando la *i*-esima osservazione. Se i risultati ottenuti nei due casi risultano pressoché uguali fra di loro non si ha motivo di ritenere che la *i*-esima osservazione eserciti un'influenza rilevante sulle stime mentre, al crescere delle differenze fra i due risultati, l'osservazione potrà essere considerata un outlier.
Nel SAS gli indici che confrontano i risultati ottenuti con *n* e con *n*-1 osservazioni sono

- **Rapporto cov**, che valuta il cambiamento della varianza residua stimata,
- **DFFITS**, che misura il cambiamento nei valori stimati della variabile dipendente
- **DFBETAS**, che misura la variazione nelle stime dei parametri della retta, per cui si ha un indice per ciascuno dei parametri presenti nel modello di regressione (nel caso del modello di regressione semplice, vengono calcolati due indici: per l'intercetta e per il coefficiente angolare della retta)

In generale, considerato un modello con *h* parametri, devono essere controllate quelle osservazioni per le quali risulta

$$|\text{Rapporto cov} - 1| \geq 3\frac{h}{n}, \quad \text{valore soglia valido quando } n > 3h$$

$$|\text{DFFITS}| \geq 2\sqrt{\frac{h}{n}}$$

$$|\text{DFBETAS}| \geq 2\sqrt{\frac{1}{n}}.$$

Nell'esempio numerico considerato in precedenza risulta h=2 e n=24 per cui i valori critici sono

$$|\text{ Rapporto cov } - 1| \geq 0.25$$

$$|\text{DFFITS}| \geq 0.5773$$

$$|\text{DFBETAS}| \geq 0.4082$$

I valori di tutti questi indici, utili per individuare le eventuali osservazioni anomale, si ottengono aggiungendo l'opzione `influence` all'interno dell'istruzione `model` della **`proc reg`**, per cui il PROC STEP considerato nella figura 3.3.1 assume la forma

```
proc reg;
model y=x/influence;
run;
```

e l'output corrispondente del SAS è costituito dalla seguente tabella 3.3.5 in cui la prima colonna indica il numero identificativo dell'osservazione, la seconda il residuo e la terza il residuo studentizzato. Il valore del Rapporto cov è riportato nella quinta colonna, mentre nella sesta si trova il valore del DFFITS e nelle ultime due i valori del DFBETAS per l'intercetta e il coefficiente angolare della retta di regressione.

Per una maggiore leggibilità dei risultati, i diagnostici che risultano esterni ai valori soglia precedentemente calcolati sono stati evidenziati in giallo, per cui si nota subito come per tre particolari osservazioni (l'ottava, la ventesima e l'ultima) ci sono 3 diagnostici su 5 che risultano maggiori del valore soglia.

Tabella 3.3.5
Analisi dei diagnostici per l'individuazione degli outliers

Statistiche di output							
Oss	**Residuo**	**RStudent**	**Diag Hat H**	**Rapporto cov**	**DFFITS**	**DFBETAS**	
						Intercept	**x**
1	-605.0464	-1.0343	0.0589	1.0559	-0.2587	-0.2272	0.1399
2	-364.1759	-0.6189	0.0775	1.1475	-0.1794	0.0439	-0.1220
3	-113.1647	-0.1892	0.0620	1.1660	-0.0486	-0.0436	0.0278
4	-93.3046	-0.1543	0.0418	1.1428	-0.0322	-0.0173	0.0021
5	204.6954	0.3392	0.0418	1.1330	0.0709	0.0379	-0.0045
6	60.2762	0.1008	0.0636	1.1709	0.0263	0.0237	-0.0154
7	128.6954	0.2129	0.0418	1.1405	0.0445	0.0238	-0.0029
8	-581.4454	-1.3203	0.4504	1.7024	-1.1951	0.8243	-1.1385
9	-549.0464	-0.9344	0.0589	1.0749	-0.2337	-0.2053	0.1264
10	826.4588	1.4317	0.0429	0.9518	0.3032	0.1891	-0.0521
11	123.6527	0.2119	0.1067	1.2233	0.0732	0.0721	-0.0572
12	797.6954	1.3765	0.0418	0.9637	0.2876	0.1540	-0.0185
13	-87.2507	-0.1451	0.0523	1.1558	-0.0341	-0.0281	0.0154
14	141.7485	0.2597	0.2180	1.3945	0.1371	-0.0794	0.1233
15	-546.1968	-0.9398	0.0790	1.0973	-0.2752	-0.2618	0.1891
16	-316.2507	-0.5290	0.0523	1.1278	-0.1243	-0.1024	0.0560
17	255.8241	0.4328	0.0775	1.1688	0.1255	-0.0307	0.0853
18	299.9536	0.5032	0.0589	1.1386	0.1259	0.1105	-0.0681
19	1004	1.7794	0.0418	0.8650	0.3718	0.1990	-0.0239
20	-1232	-2.3277	0.0775	0.7518	-0.6748	0.1651	-0.4589
21	-192.2507	-0.3202	0.0523	1.1468	-0.0752	-0.0620	0.0339
22	-414.0464	-0.6984	0.0589	1.1138	-0.1747	-0.1534	0.0945
23	-201.7238	-0.3381	0.0636	1.1594	-0.0881	-0.0796	0.0518
24	1453	2.9003	0.0797	0.6079	0.8534	-0.2209	0.5895

In generale, una volta identificati gli outliers, è opportuno rimuovere tali osservazioni dall'insieme dei dati originali e ripetere tutta l'analisi. Questo procedimento può essere ripetuto più volte.

Per quanto riguarda la verifica dell'**ipotesi di normalità** delle distribuzioni delle variabili Y_i si usa esaminare la distribuzione dei residui studentizzati che, sotto le condizioni standard, risulta essere una t_{n-h-1} e quindi, per un numero di g.d.l. moderatamente elevato, simile alla normale.

Uno dei metodi più frequentemente utilizzati per la verifica della normalità si basa sull'esame della forma dell'istogramma della distribuzione di frequenza dei residui studentizzati che, affinché l'ipotesi di normalità possa essere accettata, deve risultare pressoché simmetrico. L'ipotesi deve essere invece respinta se l'istogramma è fortemente asimmetrico o presenta più mode. Si osservi che, in realtà, i residui studentizzati non sono indipendenti fra di loro tuttavia, se il campione è moderatamente numeroso, la distorsione introdotta da questa circostanza può essere ritenuta trascurabile.

Affinché il SAS possa costruire l'istogramma dei residui studentizzati occorre innanzitutto memorizzarli in un DSS, ossia occorre creare un dataset SAS che contenga l'output della regressione. Questo può essere fatto mediante le seguenti istruzioni

```
proc reg;
model y=x/r;
output out=b
       rstudent=res_stud
       p=stime;
proc print data=b;
run;
```

che creano un DSS, visibile nella finestra dell'output mediante la **`proc print`**, che associa a ciascuna osservazione oltre ai valori delle variabili originarie (modello, prezzo e capacità del disco), considerati per default, i residui studentizzati (contenuti nella variabile `res_stud` creata con l'opzione `rstudent=res_stud`). In più, per completezza, con il listato precedente ad ogni osservazione della variabile dipendente è stato associato il corrispondente valore stimato sulla base del modello lineare (questi valori sono riportati nella colonna relativa alla variabile `stime`, creata con l'opzione `p=stime`).

I dati contenuti nel dataset creato con le istruzioni appena descritte assumono la forma riportata nella tabella successiva

Tabella 3.3.6
DSS contenente i valori delle variabili originarie (mod, y, x),
i valori stimati sotto ipotesi di linearità (stime) e i residui studentizzati (res_stud) e

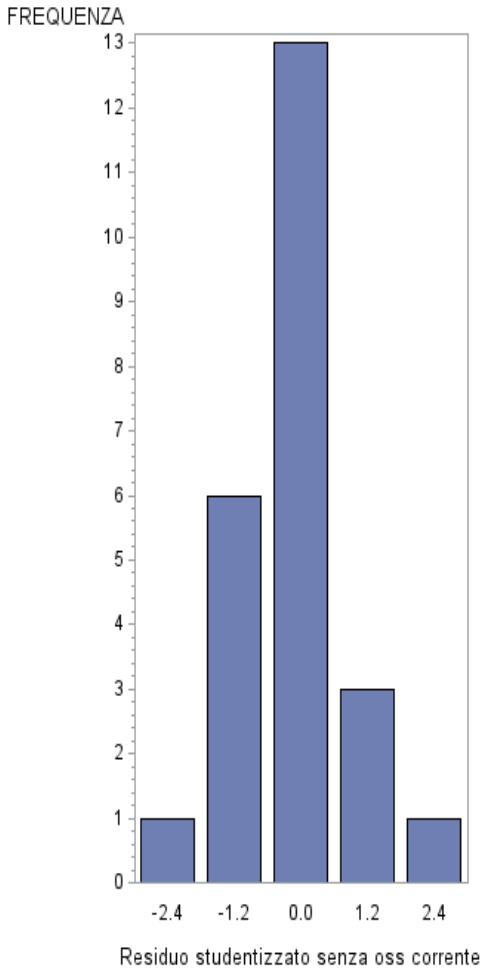
Oss	mod	y	x	stime	res_stud
1	ACER 512dx cd	1729	4.3	2334.05	-1.03429
2	ACER 723txv dvd	4332	10.0	4696.18	-0.61893
3	ABUS 17200	2138	4.1	2251.16	-0.18917
4	ABUS 17300	3111	6.4	3204.30	-0.15428
5	ABUS m8300	3409	6.4	3204.30	0.33920
6	COMPAQ 1500c	2270	4.0	2209.72	0.10079
7	COMPAQ 1750	3333	6.4	3204.30	0.21291
8	COMPAQ e700	7430	18.0	8011.45	-1.32030
9	MICRODATA 1239	1785	4.3	2334.05	-0.93437
10	COMPASS solo	3865	6.0	3038.54	1.43173
11	DATA E. mirage	1546	2.1	1422.35	0.21186
12	DELL latitude	4002	6.4	3204.30	1.37645
13	IBM thinkpad 390e	2454	4.8	2541.25	-0.14506
14	IBM thinkpad 770z	6537	14.1	6395.25	0.25972
15	IDEA PROGRESS mti	1332	3.2	1878.20	-0.93981
16	IMAGE dream.m p-sc	2225	4.8	2541.25	-0.52900
17	IMAGE p-slk	4952	10.0	4696.18	0.43279
18	MICRODATA 1332	2634	4.3	2334.05	0.50318
19	MICRODATA 2000	4208	6.4	3204.30	1.77936
20	MONOLITH geo focus	3464	10.0	4696.18	-2.32767
21	MONOLITH geo itinera	2349	4.8	2541.25	-0.32024
22	NEC ITALIA 4012	1920	4.3	2334.05	-0.69839
23	TOSHIBA satellite	2008	4.0	2209.72	-0.33814
24	TOSHIBA tecra	6191	10.1	4737.62	2.90035

Una volta ottenuto questo dataset per ottenere l'istogramma dei residui studentizzati è sufficiente scrivere le seguenti istruzioni

```
proc gchart data=b;
vbar res_stud;
```

che danno origine al grafico riportato nella figura seguente

Figura 3.3.11
Istogramma dei residui studentizzati



Sotto l'ipotesi di normalità gli stimatori ottenuti sono i più efficienti fra gli stimatori corretti, per cui la violazione di questa ipotesi può influire sull'efficienza degli stimatori, ma in genere non la modifica in modo rilevante e per questo motivo l'analisi di regressione si dice **robusta** rispetto a deviazioni dalla

condizione di normalità. Solo nel caso di distribuzioni che si discostano molto dalla normale si verifica un'influenza notevole sull'efficienza degli stimatori e questa situazione si presenta in particolare quando la distribuzione della *Y* presenta una densità elevata sulle code. In questi casi si verificano conseguenze di un certo rilievo per quanto riguarda i test e gli intervalli di confidenza dei parametri.
L'esempio che è stato analizzato non è sicuramente il migliore fra quelli possibili per ottenere stime attendibili della variabile dipendente, in quanto una previsione più accurata del prezzo dei computer può essere ottenuta sulla base di informazioni addizionali, come per esempio la RAM, la dimensione del display, la risoluzione massima, il peso. Nelle pagine successive il modello di regressione verrà generalizzato al caso in cui sia presente un maggior numero di regressori.

3.4 Cenni sulla regressione multipla

Una previsione più accurata della variabile dipendente *Z* può essere generalmente ottenuta sula base dei valori di più variabili che si ha motivo di ritenere correlate con la prima.
In generale, se si considerano $h-1$ variabili esplicative X_i ($i = 1, 2, \ldots, h-1$), la cosiddetta **funzione di regressione multipla** fa corrispondere ad ogni associazione di determinazioni di queste $h-1$ variabili il valore della media della variabile *Z* condizionata a quella particolare associazione. Si considera cioè il valore medio di Z separatamente per tutti i gruppi che risultano omogenei nelle variabili rimanenti.
L'incremento nel numero di regressori tende a far migliorare l'attendibilità della stima del valore di *Z* perché la sua varianza residua all'interno dei sottogruppi omogenei rispetto a due variabili X_1 e X_2 risulta minore o tutt'al più uguale alla varianza residua all'interno dei sottogruppi omogenei solo rispetto a X_1.
Se sono $h-1$ le variabili esplicative considerate e se indichiamo con **x** la generica associazione di determinazioni di queste variabili esplicative, supporremo che la variabile *Z* si distribuisca in modo normale con una media

$$E(Z|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{h-1} x_{h-1}, \qquad 3.4.1$$

che è una funzione lineare delle variabili X_i, e con varianza σ^2 costante per tutti i gruppi omogenei.
Il parametro β_0 rappresenta il valore medio teorico della variabile dipendente quando ciascuno dei regressori assume un valore pari a zero, mentre il generico β_i (i=1,2, ...,$h-1$) associato alla variabile

esplicativa X_i misura la variazione della variabile dipendente in corrispondenza di un incremento unitario della X_i a parità di valore dei rimanenti regressori.

Come per il caso bivariato, per avere informazioni sui parametri β_i (i=0,1,…h–1) che compaiono nel modello e sul valore della varianza σ^2 di Z si estrae in modo indipendente un campione bernoulliano da ciascuno dei gruppi omogenei.

Se è n la numerosità campionaria, restano definite n v.c. indipendenti $Y_i = Y|\mathbf{x}_i$ (i = 1,2,...,n) "valore di Z sull'individuo che presenta la determinazione $\mathbf{x}_i$ delle h – 1 variabili X_i "che si distribuiscono in modo normale con media $\beta_0 + \beta_1 x_1 + \ldots + \beta_{h-1} x_{h-1}$ e varianza σ^2.

Il modello viene indicato con l'espressione

$$Y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_{h-1} x_{h-1} + \varepsilon_i,$$

e l'ipotesi nulla globale è che tutti i coefficienti di regressione siano uguali a zero

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_{h-1} = 0$$

per cui, se l'ipotesi non viene respinta, si conclude che non vi sono relazioni lineari fra la variabile dipendente ed i regressori. La statistica utilizzata, basata sempre sul rapporto fra la varianza spiegata dal modello e la varianza residua, sotto ipotesi nulla si distribuisce come una $F_{(h\text{-}1),(n\text{-}h)}$.

Se il valore della statistica è alto e l'ipotesi viene respinta si passa alla verifica di ipotesi sui singoli coefficienti di regressione

$$H_0 : \beta_i = 0 \quad i=1,...,h-1.$$

Se l'ipotesi nulla sull'i-esimo parametro viene accettata, questo indica che il contributo dell'i-esimo regressore alla spiegazione della variabile dipendente non è significativamente diverso da zero.

A differenza del modello di regressione semplice, un'ulteriore verifica che deve essere effettuata nella regressione multipla riguarda la valutazione dell'intensità del legame lineare esistente fra le X_i, ossia l'esistenza di una situazione di **collinearità** fra i regressori. In questo caso è infatti difficile, se non impossibile, scindere le influenze che le singole X_i esercitano sulla variabile dipendente: le stime diventano inattendibili perché gli stimatori presentano una variabilità molto elevata e possono quindi

assumere valori molto lontani dai veri valori dei parametri (in caso di elevata collinearità, per esempio, è possibile ottenere stime che presentano segno opposto a quello dei coefficienti di correlazione fra i singoli regressori e la variabile dipendente, o è possibile ottenere stime che risultano tutte non significativamente diverse da zero anche se il test *F* globale è significativo).

Una misura della collinearità fra i regressori si ottiene effettuando le regressioni fra ciascun regressore e gli *h*-2 rimanenti. Se il valore dell'R^2 risulta elevato esiste una stretta relazione lineare fra il regressore preso come variabile dipendente ed i rimanenti, per cui si conclude che quel regressore non apporta informazioni significativamente rilevanti ai fini dell'analisi e viene eliminato.

Uno dei più comuni indici utilizzati per misurare la collinearità la cosiddetta **tolerance** che, per l'*i*-esimo regressore, corrisponde alla differenza fra 1 e l'R^2 calcolato fra quel regressore e gli *h*-2 rimanenti. La tolerance assume valori compresi fra 0 ed 1 e tanto più è vicino a zero tanto più elevata è la collinearità.

Un altro indice che viene frequentemente utilizzato è il Variance-inflation factor la cosiddetta (**VIF**), che corrisponde semplicemente al reciproco della tolerance.

Non esistono dei valori soglia a cui fare riferimento per la verifica della collinearità, ma il valore di riferimento per la tolerance è 0,10 (e di conseguenza è pari a 10 per il VIF).

Anche nel caso della regressione multipla le verifiche delle condizioni standard vengono effettuate nel modo già descritto per la regressione semplice.

3.5 Regressione multipla: un esempio numerico

Nel successivo listato del programma SAS sono riportati i dati relativi a 30 modelli di telefoni cellulari (dati tratti da una rivista specializzata del gennaio del 2000). Le variabili considerate sono la variabile dipendente *Y*, prezzo espresso in migliaia di lire, il peso in grammi (X_1), la lunghezza in millimetri (X_2), il numero di suonerie disponibili (X_3), il numero di minuti di autonomia della batteria in standby (X_4) ed infine il numero dei minuti di conversazione che è possibile effettuare prima che la batteria si scarichi (X_5).

Nella figura successiva si riporta il listato del programma SAS che calcola la matrice di correlazione su tutte le variabili quantitative presenti nel dataset "`data a`" e successivamente esegue l'analisi di regressione multipla, la verifica delle ipotesi sottostanti il modello, l'individuazione degli eventuali outliers e misura la collinearità fra i regressori.

Figura 3.5.1
Listato di un programma SAS con le procedure corr e reg

```
data a;
input mod $ y x1 x2 x3 x4 x5;
cards;
alcaeasy 359 150 122 15 7800 285
alcapock 590 125 116 15 4800 210
bosch509 400 150 134 27 5220 276
bosch909 1080 99 112 27 10200 300
ericsT18 690 146 105 25 6000 240
ericsT28 1450 89 97 35 4200 250
kenwooEM 499 140 125 5 6000 150
maxo3205 299 170 135 9 3000 120
moto7089 795 108 130 20 6000 195
moto3688 1450 83 83 35 4200 180
necDB500 450 135 134 14 4200 110
necD4000 750 99 120 14 6300 150
noki3210 500 151 124 35 9400 200
noki7110 850 141 120 38 14400 240
noki8210 1300 79 100 40 9000 200
panaGD30 449 135 135 20 5400 180
panaDG90 590 88 118 25 5700 210
philsavy 390 140 129 12 5220 180
phixeniu 790 95 109 19 7800 210
sagem850 835 137 132 20 8400 240
sams2100 699 125 109 17 7500 240
sams2400 949 90 108 17 7200 210
siemeC25 350 135 117 19 6000 240
siemeS25 690 125 117 40 10800 270
sonyCMD5 449 139 139 30 6900 240
telit410 299 186 130 7 5200 210
telit710 390 105 117 8 4800 240
telit810 530 100 117 9 6000 240
triumAST 350 95 123 7 4800 180
triumARI 720 85 120 9 5200 200
;

proc corr;

proc reg;
model y=x1 x2 x3/influence tol vif;
output out=b
       rstudent=res_stud;

proc gchart data=b;
vbar res_stud;

run;
```

Dai risultati contenuti nella matrice di correlazione, riportata nella tabella seguente, si nota come la variabile prezzo risulti abbastanza correlata con X_1, X_2 e X_3 e poco correlata con le restanti variabili.

Tabella 3.5.1
Matrice di correlazione (output della `proc corr`)

Coefficienti di correlazione di Pearson, N = 30 Prob > |r| con H0: Rho=0						
	y	**x1**	**x2**	**x3**	**x4**	**x5**
y	1.00000	-0.64570 0.0001	-0.77088 <.0001	0.60041 0.0005	0.24093 0.1997	0.15782 0.4049
x1	-0.64570 0.0001	1.00000	0.62492 0.0002	-0.17263 0.3617	0.01259 0.9473	-0.03257 0.8643
x2	-0.77088 <.0001	0.62492 0.0002	1.00000	-0.39181 0.0322	-0.08407 0.6587	-0.21121 0.2626
x3	0.60041 0.0005	-0.17263 0.3617	-0.39181 0.0322	1.00000	0.55888 0.0013	0.38727 0.0345
x4	0.24093 0.1997	0.01259 0.9473	-0.08407 0.6587	0.55888 0.0013	1.00000	0.47575 0.0079
x5	0.15782 0.4049	-0.03257 0.8643	-0.21121 0.2626	0.38727 0.0345	0.47575 0.0079	1.00000

Per questo motivo le variabili X_4 e X_5 non compaiono nel modello di regressione considerato successivamente. Si osservi che, anche effettuando la regressione con tutti i cinque regressori, il test *t* associato a X_4 e X_5 risulterebbe in entrambi i casi non significativo (per cui si concluderebbe che tali variabili non apportano informazioni significativamente rilevanti alla conoscenza di *Y* e si passerebbe quindi a considerare il modello in compaiono solo X_1, X_2 e X_3).

Nelle tabelle seguenti sono riportate le principali informazioni circa la varianza spiegata e residua del modello, il valore del test *F* corrispondente, alcuni indici elementari della variabile *Y* e il valore dell'R^2.

Tabella 3.5.2
Analisi della varianza del modello di regressione considerato nella figura 3.5.1

Analisi della varianza					
Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Modello	3	2278292	759431	27.28	<.0001
Errore	26	723834	27840		
Totale corretto	29	3002126			

Tabella 3.5.3
Statistiche della variabile dipendente e valore dell'R^2 sui dati della figura 3.5.1

Radice dell'MSE	166.85252	**R-quadro**	0.7589
Media dip.	664.73333	**R-quadro corr**	0.7311
Var coeff	25.10067		

Il valore dell'indice R^2 risulta pari a circa 0.76 per cui il modello lineare, nel suo complesso, è altamente significativo. Anche tutte le variabili, singolarmente considerate, risultano significative, come si nota dai risultati delle statistiche test *t* riportate nella tabella successiva. Questa stessa tabella contiene anche i risultati relativi alla tolerance ed alla VIF che, come si vede, non evidenziano una situazione di collinearità fra i regressori.

Tabella 3.5.4
Stime dei parametri della retta di regressione sui dati della figura 3.5.1

Stime dei parametri							
Variabile	**DF**	**Stima dei parametri**	**Errore standard**	**Valore t**	**Pr > |t|**	**Tolleranza**	**Inflazione varianza**
Intercept	**1**	2152.94456	356.43532	6.04	<.0001	.	0
x1	**1**	-3.59270	1.41867	-2.53	0.0177	0.60332	1.65750
x2	**1**	-10.77580	3.35691	-3.21	0.0035	0.52638	1.89976
x3	**1**	11.40962	3.16471	3.61	0.0013	0.83792	1.19343

Dall'analisi dei diagnostici per l'individuazione di eventuali outlier si nota come solo un'osservazione ha 3 diagnostici su 7 al di fuori dal valore soglia, come si nota dai risultati della tabella successiva.

Tabella 3.5.5
Analisi dei diagnostici per l'individuazione degli outliers

Oss	Residuo	RStudent	Diag Hat H	Rapporto cov	DFFITS	DFBETAS			
						Intercept	x1	x2	x3
1	-111.5360	-0.6937	0.0900	1.1911	-0.2181	-0.0559	-0.1589	0.0956	0.0777
2	-35.0083	-0.2121	0.0570	1.2317	-0.0521	-0.0299	-0.0175	0.0265	0.0284
3	-78.1419	-0.4943	0.1285	1.2912	-0.1899	0.1394	-0.0249	-0.1028	-0.1111
4	181.5627	1.1306	0.0638	1.0236	0.2951	0.0400	-0.1336	0.0212	0.1052
5	-92.1916	-0.6172	0.2177	1.4075	-0.3255	-0.1889	-0.2622	0.2669	0.0280
6	262.7217	1.7871	0.1582	0.8592	0.7747	0.3922	-0.0566	-0.3416	0.2706
7	138.9606	0.8835	0.1190	1.1742	0.3247	0.1014	0.1086	-0.0836	-0.2501
8	108.8612	0.7038	0.1573	1.2835	0.3041	-0.0429	0.1833	-0.0126	-0.1083
9	202.7288	1.3152	0.1226	1.0202	0.4916	-0.2795	-0.3470	0.3990	0.1170
10	90.3043	0.6510	0.3242	1.6185	0.4509	0.3495	0.0738	-0.3371	0.0317
11	66.2927	0.4084	0.0838	1.2433	0.1235	-0.0669	-0.0237	0.0774	-0.0102
12	86.0942	0.5312	0.0826	1.2192	0.1594	0.0142	-0.1046	0.0488	-0.0538
13	-173.5842	-1.1402	0.1578	1.1341	-0.4936	0.2047	-0.2038	-0.0543	-0.3559
14	63.1567	0.4067	0.1616	1.3590	0.1786	-0.0636	0.0546	0.0162	0.1431
15	52.0739	0.3430	0.2003	1.4357	0.1717	0.0293	-0.0617	-0.0106	0.1015
16	7.6108	0.0472	0.1007	1.3004	0.0158	-0.0118	-0.0040	0.0121	0.0048
17	-260.4830	-1.7234	0.1173	0.8458	-0.6281	0.1354	0.5105	-0.3344	-0.1946
18	-6.8035	-0.0414	0.0661	1.2523	-0.0110	0.0016	-0.0022	-0.0017	0.0044
19	-63.8585	-0.3910	0.0731	1.2317	-0.1098	-0.0664	0.0390	0.0314	0.0327
20	368.4688	2.5196	0.0738	0.5109	0.7111	-0.4699	-0.0663	0.4523	0.1921
21	-24.2582	-0.1502	0.0987	1.2933	-0.0497	-0.0382	-0.0258	0.0393	0.0228
22	89.2214	0.5540	0.0933	1.2288	0.1777	0.1121	-0.0712	-0.0497	-0.0708
23	-273.9440	-1.7575	0.0571	0.7786	-0.4326	-0.1745	-0.2627	0.2272	0.1001
24	-209.4731	-1.3954	0.1611	1.0330	-0.6115	0.2144	-0.0306	-0.1124	-0.5356
25	-49.0115	-0.3272	0.2216	1.4775	-0.1746	0.1503	0.0423	-0.1384	-0.1162
26	135.2847	0.9588	0.2872	1.4204	0.6086	0.0968	0.5006	-0.2438	-0.2612
27	-216.2192	-1.3937	0.1041	0.9680	-0.4750	-0.2256	0.1310	0.0673	0.3509
28	-105.5924	-0.6616	0.1049	1.2193	-0.2265	-0.0932	0.0933	0.0088	0.1504
29	-216.0818	-1.4328	0.1500	1.0037	-0.6018	-0.0570	0.3828	-0.1869	0.3173
30	62.8445	0.4059	0.1668	1.3676	0.1816	0.0225	-0.1325	0.0561	-0.0800

Considerati i valori h=4 e n=30 i valori critici sono infatti

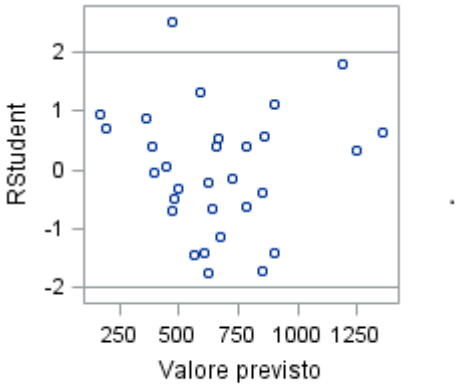
$$|\text{Rapporto cov} - 1| \geq 0.4$$

$$|\text{DFFITS}| \geq 0.7303$$

$$|\text{DFBETAS}| \geq 0.3651$$

Per quanto riguarda le verifiche delle condizioni standard, il grafico 3.5.2 che riporta i residui studentizzati in ordinata ed in ascissa i valori stimati della variabile dipendente non mostra nessuna particolare regolarità, per cui il modello lineare può ritenersi adeguato a descrivere la relazione esistente fra le variabili considerate e non si ha motivo di rifiutare la condizione di omoschedasticità.

Figura 3.5.2
Grafico dei residui studentizzati contro i valori stimati per i dati della tabella 5.8.1



Anche l'ipotesi di normalità non appare violata in maniera evidente, dato che l'istogramma dei residui studentizzati, riportato nella figura 3.5.3, non appare del tutto simmetrico, ma ha una moda che è comunque posizionata in posizione centrale.

Figura 3.5.3
Istogramma dei residui studentizzati