

Analisi di regressione con SAS

PROC CORR
PROC GPLOT
PROC REG

Esempio: due test su 31 individui

```
title1 'Risultato di 2 test su 31 soggetti';  
data dati;  
infile 'V:\Didat\Carla\regressione\tab11_1.txt';  
input id test1 test2;  
proc print;  
run;
```

N. di variabili: p = 2

N. di osservazioni: n=31

dati in tab11_1.txt

Risultato di 2 test su 31 soggetti			
Obs	id	test1	test2
1	1	50	69
2	2	66	85
3	3	73	88
4	4	84	70
5	5	57	84
...			
30	30	65	73
31	31	61	52

Ci interessa...

- ◆ Valutare la correlazione tra i due test
- ◆ Noto il risultato del 1° test, possiamo prevedere il risultato del 2° test?
- ◆ La previsione che facciamo è valida?

Matrice di covarianza

$$\mathbf{S} = \{\text{cov}(\mathbf{X}_s, \mathbf{X}_v)\} \quad s, v = 1, 2, \dots, p$$

Matrice di correlazione

$$\mathbf{R} = \{\rho_{sv}\} = (1/n) \mathbf{Z}' \mathbf{Z}$$

Coefficiente di correlazione

$$\rho_{sv} = [\text{cov}(\mathbf{X}_s, \mathbf{X}_v)] / [V(\mathbf{X}_s)V(\mathbf{X}_v)]^{1/2}$$

MEDIE E CORRELAZIONI

```
PROC CORR DATA= dati;  
VAR test1 test2;  
RUN;
```

Calcola la matrice di correlazione e di covarianza (opzione COV) per le variabili indicate nell'istruzione VAR

Output di PROC CORR

Risultato di 2 test su 31 soggetti
The CORR Procedure
2 Variables: test1 test2

Simple Statistics

Variable	N	Mean	StdDev	Sum	Minimum	Maximum
test1	31	70.65	14.83	2190	42.00	95.00
test2	31	75.77	15.92	2349	44.00	98.00

Pearson Correlation Coefficients, N = 31

	Prob > r under H0: Rho=0	
	test1	test2
test1	1.00000	0.70307
		<.0001
test2	0.70307	1.00000
	<.0001	

Correlazione e covarianza

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

Calcolo delle covarianze in SAS

```
proc corr cov;  
var test1 test2;  
run;
```

L'opzione **COV** consente di ottenere sia la matrice di correlazione che la matrice di covarianza

Covariance Matrix, DF = 30

	test1	test2
test1	219.9698925	165.9838710
test2	165.9838710	253.3806452

Scatterplot dei punti



```
TITLE 'SCATTERPLOT di test1*test2';  
PROC Gplot;  
PLOT test2 * test1;  
RUN;
```

Programma SAS per scatterplot: tab11_1.sas

Scatterplot dei punti con retta di regressione

```
symbol1 v=dot c=blue i=rl ci=red;  
proc gplot data=dati;  
plot test2*test1;  
run;
```

istruzione SYMBOL

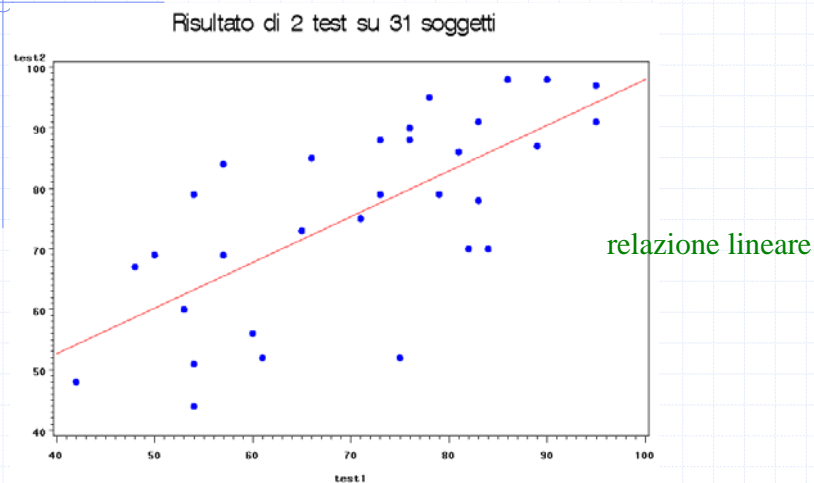
V= definisce simbolo dei punti

C= definisce colore punti

I= definisce tipo di interpolazione (RL: lineare)

CI= definisce colore linea interpolante

OUTPUT DI PROC Gplot



Modello di regressione lineare

```
proc reg data=dati;  
model test2=test1;  
run;
```

$$Y = a + bX + e$$

Stima con metodo dei M.Q.

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{\text{cov}(X, Y)}{\text{var}(X)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Output di PROC REG (1)

Scomposizione della varianza totale:
 $V(\text{total}) = V(\text{Model}) + V(\text{Error})$

Analysis of Variance

devianze

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3757.42041	3757.42041	28.35	<.0001
Error	29	3843.99895	132.55169		
Corrected Total	30	7601.41935			

Output di PROC REG (2)

Root MSE	11.51311	R-Square	0.4943
Dependent Mean	75.77419	Adj R-Sq	0.4769
Coeff Var	15.19397		

Media e coefficiente di variazione di Y

Output di PROC REG (3)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
a Intercept	1	22.46709	10.22358	2.20	0.0361
b test1	1	0.75458	0.14173	5.32	<.0001

parametri stimati

$$Y^* = 22.467 + 0.755 X$$

residui e valori stimati

Opzioni p e r:
calcola e stampa valori stimati e residui

```
proc reg data=dati;  
model test2=test1/p r;  
output out=stime p=stime r=residui;  
run;
```

output out=

crea un sas data set che contiene:
p= i valori stimati per Y
r= i residui

Analisi dei residui

```
PLOT < y*x > < =symbol >
      < ... y*x > < =symbol > < / options >;
```

Grafico residui vs valori stimati

```
plot r.* p.;
```

Grafico residui vs variabile indipendente

```
plot r.*test1;
```

Calcolo di Y stimati

```
DATA new;
test1=70;
DATA new;
SET dati new;
```

Aggiunge una obs per X al data set

```
RUN;
PROC REG DATA=new;
MODEL test2=test1/p;
```

Stima modello di regressione

Calcola valori previsti dal modello

Dependent Variable: test2

Output Statistics

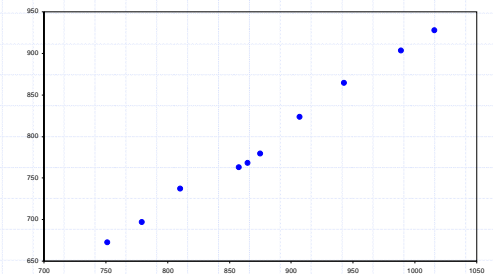
Obs	Dep Var test2	Predicted Value	Residual
1	69.0000	60.1959	8.8041
2	85.0000	72.2691	12.7309
5	84.0000	65.4779	18.5221
...			
29	51.0000	63.2142	-12.2142
30	73.0000	71.5145	1.4855
31	52.0000	68.4962	-16.4962
32	.	75.2874	.

$$Y^* = 22.467 + 0.755 * 70$$

Valore previsto

Esempio: reddito disponibile e consumi (Greene, 2001)

anno	reddito	consumo
1970	751.6	672.1
1971	779.2	696.8
1972	810.3	737.1
1973	864.7	767.9
1974	857.5	762.8
1975	874.9	779.4
1976	906.8	823.1
1977	942.9	864.3
1978	988.8	903.2
1979	1015.7	927.6
media	879.2	793.4
varianza	6719.2	6497.2
ds	82.0	80.6
covarianza		6579.9
rho		0.9959



X: reddito in bilioni di \$ (valuta '72)
Y: consumi in bilioni di \$ (valuta '72)

Modello di regressione lineare semplice

- ◆ Forma funzionale $y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \dots, n$
- ◆ $E(\varepsilon_i) = 0$
- ◆ Omoschedasticità: $Var(\varepsilon_i) = \sigma^2$, costante
- ◆ Residui incorrelati: $cov(\varepsilon_i, \varepsilon_j) = 0$
- ◆ Regressori e errori incorrelati
 $cov(x_i, \varepsilon_i) = 0$
- ◆ $\varepsilon_i \sim N(0, \sigma^2)$

Segue esempio consumo/reddito

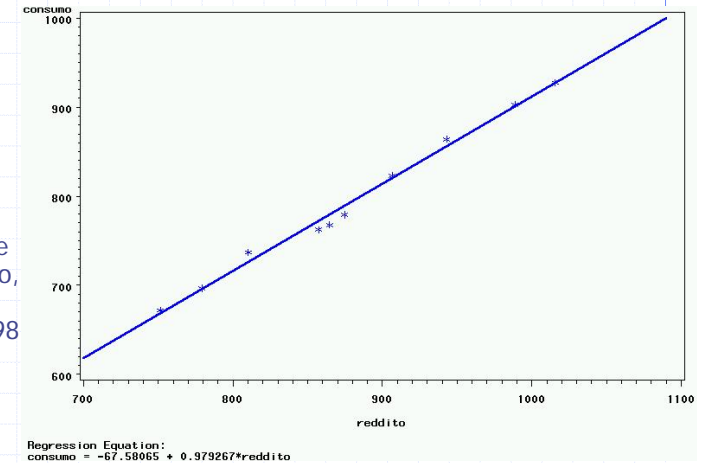
Stima OLS

$a = -67.58$

consumi per reddito=0

$b = 0.979$

per ogni bilione in più di reddito, i consumi crescono di 0.98 bilioni



consumi_greene.sas

Scomposizione della varianza

◆ $SST = SSR + SSE$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2$$

◆ $R^2 = SSR/SST$ indice di determinazione

◆ $R^2 = \rho^2$

$$R^2 = corr(\hat{y}_i, y_i) = \frac{\sum_{i=1}^n [(\hat{y}_i - \bar{y})(y_i - \bar{y})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$

Bontà di adattamento

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	64435	64435	959.92	<.0001
Error	8	537	67.13		
Corrected Total	9	64972			

SSR (Sum of Squares Regression) is 64435.
 SSE (Sum of Squares Error) is 537.
 SST (Sum of Squares Total) is 64972.
 R-Square is 0.9917 (SSR/SST).
 Adjusted R-Square is 0.9907.
 Root MSE is 8.19303.
 Dependent Mean is 793.43000.
 Coeff Var is 1.03261.

TEST F

$$F_{[1, (N-2)]} = \frac{SSR/1}{SSE/(N-2)}$$

$H_0: \beta = 0, H_a: \beta \neq 0$

Test e intervalli di confidenza

◆ Stima di σ^2

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

◆ Statistica per il test $H_0: \beta=0$

$$t = \frac{b}{SE_b} \sim t_{(n-2)}$$

◆ Intervallo di confidenza al livello c per β

$$b \pm t^* SE_b$$

t^* valore di t di Student con $(n-2)$ g.l.: $P(-t^* < t < t^*) = c$

segue esempio consumo/reddito

Parameter Estimates

Variable	Label	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-67.58065	27.91071	-2.42	0.0418
reddito	reddito	1	0.97927	0.03161	30.98	<.0001

$$\text{Var}(b) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$t = \frac{b}{SE_b} \sim t_{(n-2)}$$

Intervallo di confidenza per

$$y_i = a + bx_i$$

