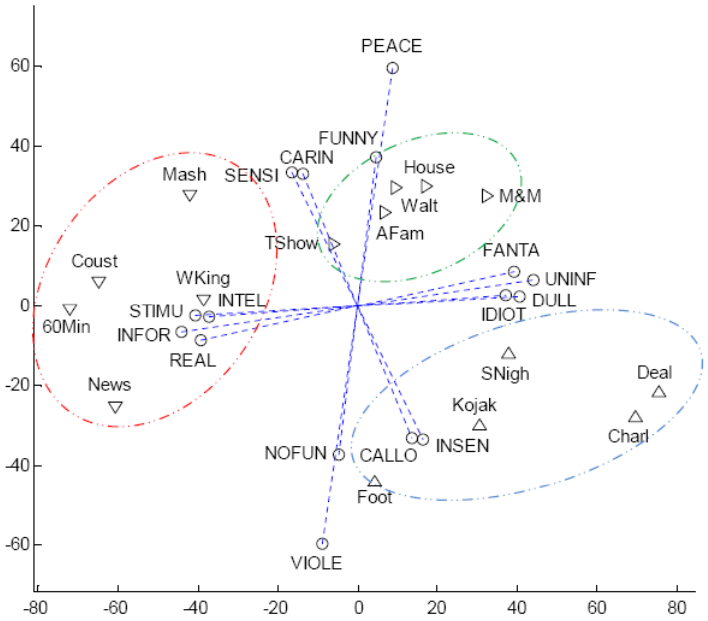


Metodi Statistici per il Management

Statistica Multivariata I

Simone Borra - Roberto Rocci

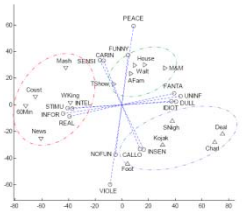
Introduzione e obiettivi

La statistica multivariata si occupa di analizzare e studiare in modo simultaneo un set di *k* **variabili** su un campione di *n* **unità**:

- **simultaneo** indica che lo studio è congiunto, non una variabile alla volta;
- **set di k variabili**: in linea di massima *k* è inteso maggiore o uguale a 3 – nel caso di 2 variabili si utilizzano modelli di correlazione o di regressione semplice.

Obiettivi:

- individuare e misurare i **legami** tra le variabili;
- **segmentare** (raggruppare) le unità in sottoinsiemi simili;
- ricercare **regolarità o tendenze** nei dati;
- definire le **gerarchie** tra variabili.



Indice

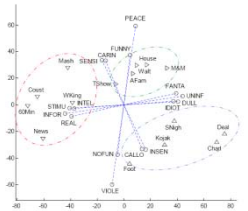
I modelli e le tecniche che tratteremo sono:

Modelli inferenziali per l'analisi della dipendenza (I)

- modelli di regressione multipla;
- modelli di analisi della varianza.

Tecniche esploratorie di riduzione dimensionale (II)

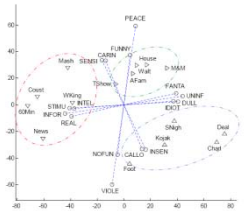
- riduzione variabili: analisi in componenti principali;
- riduzione unità: analisi dei gruppi (*cluster analysis*).



La struttura dati

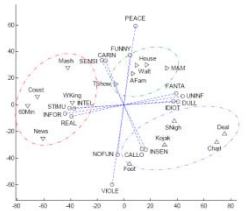
- Il set di dati è solitamente molto “ricco”, caratterizzato da un numero ampio di variabili.

- In linea generale la struttura dati può presentare variabili di diversa natura: quantitative o qualitative.



Esempio dati

UNITA'	PREZZO	CV	SICUREZZA	CONFORT	CONSUMO
MAZDA 3	17.000	110	4	3	8,0
MEGANE	16.500	115	5	4	7,5
Nissan 350 Z	32.000	180	4	3	11,0
Peugeot 307	16.900	105	4	4	7,6
OCTAVIA	21.000	130	4	4	8,4
ALFA 147	16.800	105	4	4	8,1
AUDI A3	20.700	102	4	4	6,9



Regressione lineare multipla

Scopo generale della regressione multipla è quello di studiare la relazione esistente tra una **variabile dipendente** e ***k* variabili indipendenti**, o esplicative. In termini formali:

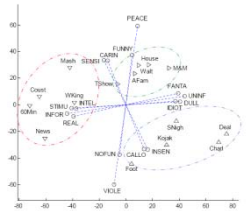
$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}) + w_i$$

Se la regressione è di tipo lineare si ottiene:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + w_i$$

Equazione	*Assunzioni*
$E(y_i \mid x_{i1},\ldots,x_{ik}) = \mu_i = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik}$	$y_i \mid \mathbf{x}_i \sim N(\mu_i, \hat{\sigma}_w^2)$, indip.

L'interpretazione è analoga a quanto visto per la regressione lineare semplice



reg: stima e bontà

- la stima di **b** = $[b_0,b_1,\ldots,b_k]$ è calcolata minimizzando

$$\Sigma_i (y_i - b_0 - b_1 x_{i1} + \ldots - b_k x_{ik})^2$$

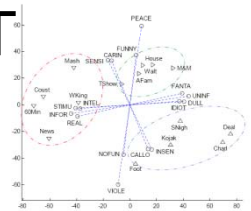
in questo modo otteniamo gli stimatori corretti

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Questo metodo di stima (detto dei minimi quadrati) coincide con quello di massima verosimiglianza nel caso di errori normalmente distribuiti.
- Lo stimatore corretto della varianza dell'errore è

$$\hat{\sigma}_w^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- SST = SSM+SSR $\Rightarrow$ R^2 = SSM/SST = 1 – SSR/SST



reg: esempio

Number of Observations	67

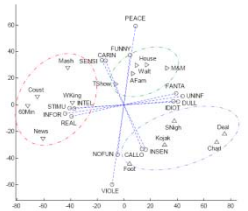
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8478759	1695752	43.54	<.0001
Error	61	2375649	38945		
Corrected Total	66	10854409			

Root MSE	197.345	R-Square	0.781
Dependent Mean	1161.463	Adj R-Sq	0.763
Coeff Var	16.991		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-478.110	162.092	-2.95	0.005	-802.233	-153.988
age	1	-5.712	2.121	-2.69	0.009	-9.954	-1.470
feats	1	4.936	21.394	0.23	0.818	-37.845	47.717
ne	1	146.402	60.796	2.41	0.019	24.832	267.971
cor	1	191.379	61.859	3.09	0.003	67.684	315.073
sqft	1.000	0.987	0.101	9.770	<.0001	0.785	1.189



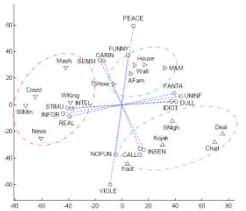
reg: test t e F

Test t (significatività delle singole variabili)

- Ipotesi nulla H_0: $b_j=0$
- Stat. test t_{oss} = Coef./(Std.Err.) ~ T_{n-k-1} se H_0 vera
- regola rifiuto $| t_{oss} | > t_{\alpha/2}$
- p-value $2Pr\{T_{n-k-1} > | t_{oss} |\}$
- **Teorema di inversione del test.** Interessante osservare che un coef. è significativo al 100*alfa% se e solo se lo 0 non è contenuto nell'intervallo di confidenza al 100*(1-alfa)%.

Test F (significatività del modello nel complesso)

- Ipotesi nulla H_0: $b_1=b_2=\ldots=b_k=0$
- Stat. Test F_{oss}=(SSM/dfM)/(SSR/dfR) =MSM/MSR
  $F_{oss} \sim F_{k,n-k-1}$ se H_0 vera
- regola rifiuto $F_{oss} > F_{\alpha}$
- p-value $Pr\{F_{k,n-k-1} > F_{oss}\}$



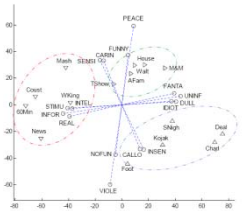
reg: backward elimination

Per selezionare le variabili significative si parte dal modello completo, si elimina la variabile “meno significativa” (i.e. p-value più alto), si stima nuovamente il modello. La procedura termina quando tutte le variabili hanno un p-value inferiore ad una certa soglia.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8042152	4021076	91.51	<.0001
Error	64	2812257	43942		
Corrected Total	66	10854409			

Root MSE	209.6223	R-Square	0.7409
Dependent Mean	1161.463	Adj R-Sq	0.7328
Coeff Var	18.04813		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-596.21613	156.4608	-3.81	0.0003	-908.782	-283.65
sqft	1	1.07648	0.10129	10.63	<.0001	0.87413	1.27883
cor	1	215.05618	64.38887	3.34	0.0014	86.42462	343.6877



reg: modelli lineari e non lineari

Importante precisare che il modello è lineare nei parametri ma non necessariamente nelle variabili.

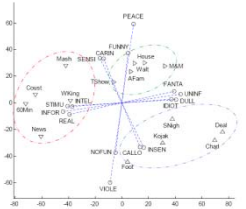
Esempi

$$E(y|x) = b_0 + b_1 x \quad \text{lin. parametri, lin. variabili}$$

$$E(y|x) = b_0 + b_1 x + b_2 x^2 \quad \text{lin. parametri, non lin. variabili}$$

$$E(y|x) = b_0 + b_1 \log(x) \quad \text{lin. parametri, non lin. variabili}$$

$$E(y|x) = b_0 x^{b_1} \quad \text{non lin. parametri, non lin. variabili}$$



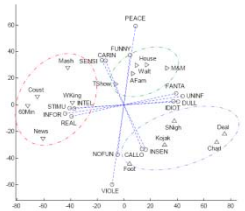
Modelli ANOVA (aov)

- Lo studio della dipendenza tra una variabile risposta quantitativa ed una esplicativa qualitativa si realizza utilizzando il modello

  ANOVA: Analisi della varianza.

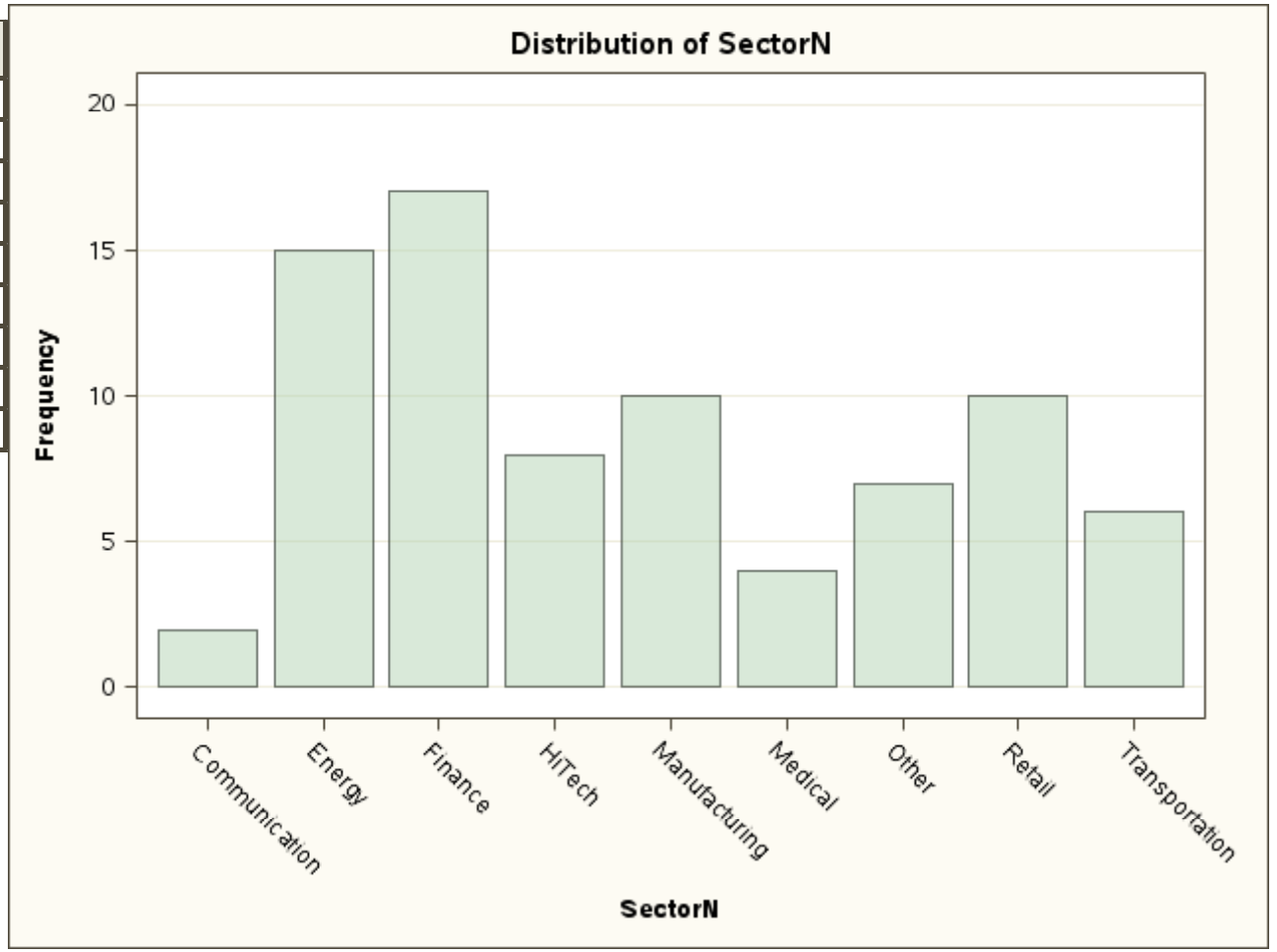
- Il proposito è quello di verificare la presenza di differenze significative tra medie condizionate (i.e. le media della variabile risposta condizionate alle diverse categorie della variabile esplicativa).

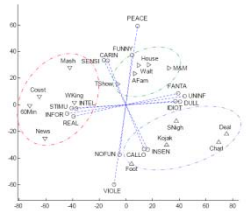
- Il nome deriva dal fatto che per verificare la significatività statistica nella differenza tra medie si devono confrontare varianze.



aov: esempio

SectorN	Frequency	Percent
Communication	2	2.53
Energy	15	18.99
Finance	17	21.52
HiTech	8	10.13
Manufacturing	10	12.66
Medical	4	5.06
Other	7	8.86
Retail	10	12.66
Transportation	6	7.59



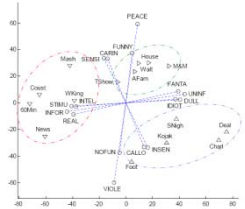


aov: esempio

Analisi della variabile log(fatturato) per settore

Analysis Variable : lSales								
SectorN	N Obs	Mean	Std Error	Minimum	Maximum	N	Lower 95% CL for Mean	Upper 95% CL for Mean
Communication	2	8.480	0.634	7.847	9.114	2	0.427	16.534
Energy	15	7.296	0.257	5.872	9.657	15	6.745	7.846
Finance	17	6.639	0.269	5.170	9.693	17	6.070	7.208
HiTech	8	8.515	0.531	6.510	10.821	8	7.259	9.771
Manufacturing	10	8.250	0.229	7.324	9.615	10	7.732	8.769
Medical	4	6.489	0.666	5.323	8.331	4	4.371	8.607
Other	7	7.456	0.324	6.323	9.118	7	6.665	8.248
Retail	10	8.464	0.218	7.378	9.748	10	7.971	8.957
Transportation	6	7.879	0.238	7.175	8.721	6	7.268	8.489

Sono significativamente diverse?
C’è dipendenza del log(fatturato) dal setttore?

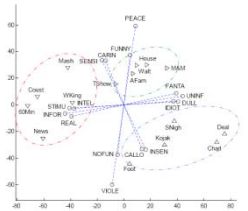


aov: implementazione

- Nell'accezione più semplice si tratta di un modello di regressione lineare multipla avente come regressori *k* variabili dummy, dove *k*+1 è il numero di categorie della variabile esplicativa.

- Ogni dummy rappresenta una categoria della variabile esplicativa. Questa assume il valore 1 se la categoria è presente e 0 altrimenti.

- La categoria esclusa dalla codifica (i.e. senza dummy corrispondente) è detta di riferimento.

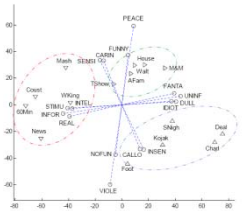


aov: implementazione

Esempio. Consideriamo la variabile titolo di studio con modalità: Analfabeta, Licenza, Diploma, Laurea. Questa può essere codificata in 3 variabili dummy scegliendo come riferimento la categoria Analfabeta.

	d_{i1}	d_{i2}	d_{i3}
Licenza	**1**	**0**	**0**
Diploma	**0**	**1**	**0**
Laurea	**0**	**0**	**1**
Diploma	**0**	**1**	**0**
Laurea	**0**	**0**	**1**
Analfabeta	**0**	**0**	**0**

categorie variabili dummy



aov: modello

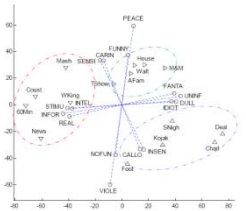
Prendendo come riferimento la categoria k+1, il modello diventa

$$y_i = b_0 + b_1 d_{i1} + b_2 d_{i2} + \ldots + b_k d_{ik} + w_i$$

Ossia

$$y_i = \begin{cases} b_0 + w_i & \text{se } i \text{ presenta la categoria } k+1 \\ b_0 + b_j + w_i & \text{se } i \text{ presenta la categoria } j \neq k+1 \end{cases}$$

- b_0 è la media della variabile risposta condizionata alla categoria k+1 della variabile esplicativa
- b_0+b_j è la media della variabile risposta condizionata alla categoria j della variabile esplicativa
- b_j è la differenza tra le due medie prima citate



aov: esempio

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	42.57515	5.321894	5.29	<.0001
Error	70	70.45086	1.006441		
Corrected Total	78	113.026			

R-Square	Coeff Var	Root MSE	lSales Mean
0.376685	13.21005	1.003215	7.594333

Parameter	Estimate	Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	7.879	0.410	19.24	<.0001	7.062	8.695
SectorN Communication	0.602	0.819	0.73	0.465	-1.032	2.236
SectorN Energy	-0.583	0.485	-1.20	0.233	-1.549	0.384
SectorN Finance	-1.239	0.476	-2.60	0.011	-2.190	-0.289
SectorN HiTech	0.637	0.542	1.17	0.244	-0.444	1.717
SectorN Manufacturing	0.372	0.518	0.72	0.475	-0.661	1.405
SectorN Medical	-1.389	0.648	-2.15	0.035	-2.681	-0.098
SectorN Other	-0.422	0.558	-0.76	0.452	-1.535	0.691
SectorN Retail	0.586	0.518	1.13	0.262	-0.448	1.619
SectorN Transportation	0	.	.	.	.	.