Analisi discriminante e regressione logistica: applicazione al sondaggio sul tema delle aggregazioni comunali per il Comune di Novazzano

Flaminio Cadlini e Roberto Stoppa
www.tiresia.ch*

Giugno 2006

Indice

*L'analisi puó essere scaricata dal sito www.tiresia.ch/

1 Introduzione

L'obiettivo di questo lavoro é di analizzare i dati che sono stati raccolti con un sondaggio (inchiesta), con lo scopo di valutare, fra gli abitanti del Comune di Novazzano che hanno diritto di voto, i favorevoli e i contrari ad una eventuale aggregazione del comune di Novazzano. In totale gli abitanti di Novazzano con diritto di voto (denominati anche elenco elettorale) sono 1'639. Avendo a disposizione delle informazioni su ogni persona abbiamo effettuato un campionamento stratificato scegliendo 526 persone e utilizzando come variabili di stratificazione l'etá dell'individuo e gli anni di residenza nel comune. A queste 526 persone é stato inviato, per posta, un questionario con 13 domande, piú una domanda finale aperta che chiedeva eventuali commenti. In totale sono rientrati 231 questionari (44%) completi dopo diversi richiami. La domanda 9 del questionario: *Lei è favorevole o contrario ad un'eventuale aggregazione del Comune di Novazzano con uno o piú comuni del Mendrisiotto?*, l'abbiamo utilizzata quale variabile da modellizzare, **sì = favorevole** e **no = contrario all'aggregazione del comune di Novazzano**.

I metodi che abbiamo utilizzato sono **l'analisi discriminante** e **la regressione logistica** in un'ottica di score. L'obbiettivo é quello di classificare, utilizzando alcune caratteristiche degli individui ed alcune risposte ottenute nel questionario, i favorevoli all'aggregazione del Comune di Novazzano con uno o più comuni del Mendrisiotto e i contrari. Dei metodi non parametrici saranno pure utilizzati per giudicare la qualità delle analisi.

2 Descrizione dei dati

Come detto i dati analizzati sono stati raccolti con dei questionari inviati direttamente per posta a 526 individui del Comune di Novazzano con diritto di voto. Giá dall'inizio eravamo coscienti che non tutti gli individui avrebbero ritornato il questionario. Dopo alcuni richiami siamo riusciti a ricevere 231 questionari validi. Statisticamente parlando con 231 questionari validi ed un grado di affidabilità del 95% il margine di errore è del +/- 6.0%.
Le variabili che saranno utilizzate per l'analisi sono:

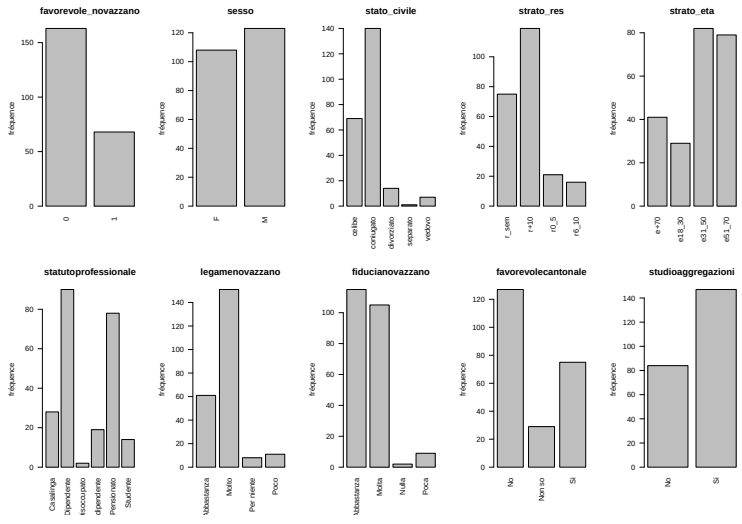
- *favorevole-novazzano*: se l'individuo è favorevole o contrario ad una eventuale aggregazione del Comune di Novazzano con uno o più comuni del Mendrisiotto (variabile da discriminare o da spiegare), **1 = favorevole** e **0 = contrario**;
- *sesso*: sesso dell'individuo (conosscciuto a priori);
- *statocivile*: stato civile dell'individuo (conosciuto a priori);
- *strato-res*: anni di residenza nel comune di Novazzano (conosciuto a priori);
- *strato-eta*: età dell'individuo (conosciuto a priori);
- *statutoprofessionale*: statuto professionale dell'individuo (ottenuto tramite questionario);

- *legamenovazzano*: evidenzia il grado di appartenenza al Comune di Novazzano (ottenuto tramite questionario);
- *fiducianovazzano*: evidenzia il grado di fiducia nell'amministrazione comunale di Novazzano (ottenuto tramite questionario);
- *favorevolecantonale*: se l'individuo è favorevole o contrario alla politica cantonale del Canton Ticino in merito alle aggregazioni comunali (ottenuto tramite questionario);
- *studioaggregazioni*: se l'individuo vorrebbe far effettuare uno studio approfondito dal Comune sul tema delle aggregazioni comunali (ottenuto tramite questionario).

3 Analisi preliminare esploratoria

Qui di seguito presentiamo delle statistiche descrittive delle variabili e dei grafici che evidenziano la loro distribuzione empirica. Per quanto concerne la variabile da discriminare (*favorevolenovazzano*) il campione si divide in 68 individui favorevoli all'aggregazione di Novazzano con uno o piú comuni del Mendrisiotto e 163 individui contrari ad una eventuale aggregazione del Comune di Novazzano.

Figura 1: Barplots di tutte le variabili (frequenze assolute)



Le variabili esplicative sono tutte categoriali e qui di seguito evidenziamo le categorie (modalità) delle variabili:

- sesso (*sesso*):
  - M = maschile;

  - F = femminile.

- stato civile (*stato-civile*):
  - coniugato;
  - celibe;
  - divorziato;
  - vedovo;
  - separato.
- anni di permanenza nel comune (*strato-res*):
  - da 0 a 5 anni;
  - da 6 a 10 anni;
  - da più di 10 anni;
  - da sempre.
- anni (*strato-eta*):
  - da 18 a 30 anni;
  - da 31 a 50 anni;
  - da 51 a 70 anni;
  - più di 70 anni.
- statuto professionale (*statutoprofessionale*):
  - dipendente;
  - indipendente;
  - studente;
  - pensionato;
  - casalinga;
  - altro.
- legame al Comune di Novazzano (*legamenovazzano*):
  - per niente legato al Comune;
  - poco legato al Comune;
  - abbastanza legato al Comune;
  - molto legato al Comune.
- fiducia alle autorità politiche del Comune di Novazzano (*fiducianovazzano*):
  - nessuna fiducia nelle autorità politiche del Comune;
  - poca fiducia nelle autorità politiche del Comune;
  - abbastanza fiducia nelle autorità politiche del Comune;
  - molta fiducia nelle autorità politiche del Comune.

- favorevole alla politica del Governo cantonale del Canton Ticino sulle aggregazioni comunali(*favorevolecantonale*):
  - sì = favorevole;
  - no = contrario;
  - non so = ancora indeciso sulla politica del Governo cantonale.
- favorevole ad uno studio approfondito del Comune di Novazzano sul tema delle agrgegazioni (*studioaggregazioni*):
  - sì = favorevole;
  - no = contrario.

4 Analisi discriminante sull'aggregazione del Comune di Novazzano

Lo scopo dell'analisi discriminante è quello di classificare (attribuire a delle classi preesistenti) gli individui (nel nostro caso gli individui che fanno parte del catalogo elettorale del Comune di Novazzano) caratterizzati per un certo numero di variabili nominali. L'analisi fattoriale discriminante consiste nel ricercare le combinazioni lineari di p variabili esplicative $(x_1, x_2, \ldots, x_p)$, generalmente continue, che permettono di separare al meglio le q classi (nel nostro caso nelle due classi: gli individui favorevoli all'aggregazione del Comune di Novazzano e quelli contrari).

Visto che tutte le variabili che utilizziamo per questa analisi sono di tipo categoriale, un'analisi discriminante classica non sembra essere la più appropriata ma il metodo piú pertinente è il DISQUAL.

4.1 Descrizione del metodo DISQUAL

Come detto, la tabella dei dati è formata da variabili nominali, ed occorre perció procedere ad una codifica disgiuntiva completa delle p variabili eplicative. La codifica disgiuntiva completa consiste nel creare, per ogni variabile, tante colonne quante sono le proprie modalità. Le colonne rappresentano le indicatrici di ogni modalità per ogni variabile. Nel nostro caso abbiamo 9 variabili esplicative, delle quali 2 a due modalità, 1 a tre modalità, 4 a quattro modalità, 1 a cinque modalità e 1 a sei modalità. In totale la matrice disgiuntiva completa è di dimensioni (231, 34): 231 righe che rappresentano gli individui e 34 colonne che sono il totale delle modalità. Le somme per riga di questa matrice ($\mathbf{X}$) sono sempre uguali al numero delle variabili esplicative che nel nostro caso sono 9. La matrice $\mathbf{X}$ non é invertibile in quanto esistono p relazioni lineari fra le colonne della tabella disgiuntiva completa.
In questo caso una possibilità è di realizzare un'analisi discriminante classica sui fattori dell'analisi per corrispondenze multiple (ACM). Questo metodo é conosciuto come DISQUAL.

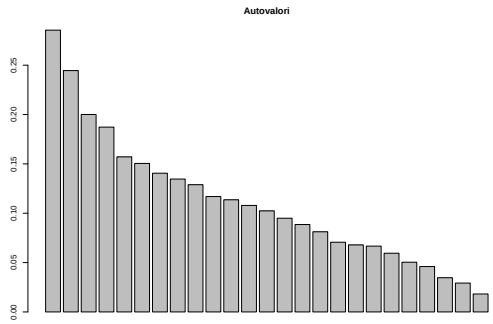
La procedura consiste nell'effettuare un'analisi delle corrispondenze sulla tabella disgiuntiva completa (matrice X) dove le p variabili categoriali sono rimpiazzate per k variabili continue corrispondenti ai fattori dell'analisi delle

corrispondenze multiple. In seguito, a partire da queste k variabili continue, che rappresentano le coordinate sugli assi fattoriali dell'analisi delle corrispondenze multiple, si effettua un'analisi fattoriale discriminante. Delle k variabili numeriche così create, si terranno le coordinate fattoriali piú discriminanti.

4.2 Primi risultati

L'analisi delle corrispondenze multiple ha evidenziato 25 fattori indipendenti dove la ripartizione dell'importanza di ogni fattore è evidenziata nella figura (2) in base agli autovalori (Eigenvalues) della varianza. Gli autovalori, così come la parte di varianza spiegata dai fattori, sono elencati nella tabella 1. Nella figura

Figura 2: Autovalori risultanti dall'analisi delle corrispondenze multiple (ACM)



3 vengono invece rappresentate le coordinate delle 34 modalità possibili sui due primi assi dell'analisi delle corrispondenze multiple.

Anche se i due primi assi spiegano soltatno il 19% della varianza totale, possiamo cercare di interpretare i sensi dei fattori. Per il primo fattore (asse delle ascisse della figura 3) sembra che non ci sia molto da dire sulla discriminazione (i punti sembrano abbastanza distribuiti attorno allo zero). Possiamo peró evidenziare che il primo fattore discriminante (individui favorevoli all'aggregazione e individui contrari all'aggregazione) potrebbe essere visto nell'età così come nello statuto professionale. In particolare notiamo in alto a destra gli individui con piú di 70 ani e pensionati (strato-eta.e+70 e statutoprofessionale.Pensionato) mentre in alto a sinistra notiamo gli individui giovani e ancora studenti (strato-eta.e18-30 e statutoprofessionale.Studente). Il secondo fattore discriminante (asse delle ordinate della figura 3) sembra invece evidenziare una differenza fra gli anni di residenza degli individui cosí come la fiducia nell'Autoritá politica del Comune. In effetti sembra che strato-res.r0-5 e strato-res.r6-10 siano contrapposti allo strato-res.r+10 e strato-res.r-sempre.

Per la ricerca di una funzione discriminante abbiamo scelto di tenere i primi 10 assi (fattori), che assieme spiegano il 63% della varianza. A questo punto applichiamo un'analisi discriminante lineare sui primi 10 fattori che risultano dall'analisi delle corrispondenze multiple con lo scopo di trovare una funzione

Tabella 1: Autovalori dell'analisi delle corrispondenze multiple (ACM)

	Autovalori	Parte di varianza spiegata	Parte cumulata
1	0.29	0.10	0.10
2	0.24	0.09	0.19
3	0.20	0.07	0.26
4	0.19	0.07	0.33
5	0.16	0.06	0.39
6	0.15	0.05	0.44
7	0.14	0.05	0.49
8	0.13	0.05	0.54
9	0.13	0.05	0.59
10	0.12	0.04	0.63
11	0.11	0.04	0.67
12	0.11	0.04	0.71
13	0.10	0.04	0.75
14	0.09	0.03	0.78
15	0.09	0.03	0.81
16	0.08	0.03	0.84
17	0.07	0.03	0.87
18	0.07	0.02	0.89
19	0.07	0.02	0.91
20	0.06	0.02	0.94
21	0.05	0.02	0.95
22	0.05	0.02	0.97
23	0.03	0.01	0.98
24	0.03	0.01	0.99
25	0.02	0.01	1.00

(combinazione lineare dei fattori, che sono delle variabili continue) che possano discriminare al meglio i due gruppi di individui: quelli favorevoli all'aggregazione da quelli contrari all'aggregazione del Comune di Novazzano

La seguente equazione evidenzia i risultati dell'analisi discriminante:

$$\begin{aligned} Z \quad = \quad & -101.041468 \cdot F1 - 128.704190 \cdot F2 + 289.639136 \cdot F3 + 129.822217 \cdot F4 + \\ & -8.416497 \cdot F5 - 241.734277 \cdot F6 + 68.350508 \cdot F7 - 26.099186 \cdot F8 + \\ & -4.913768 \cdot F9 + 121.931964 \cdot F10 \end{aligned} \tag{1}$$

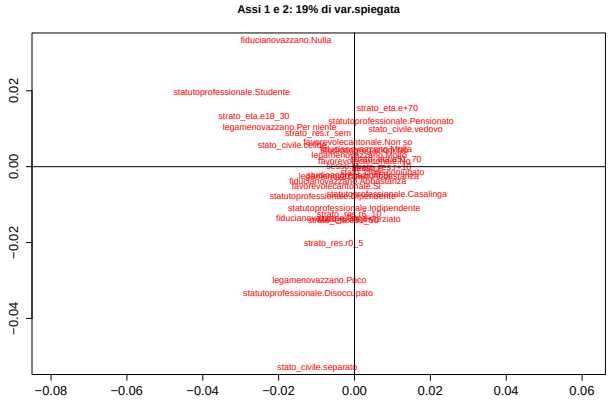
L'equazione (1) definisce uno score Z che "discrimina" la variabile nel seguente modo:

$$\text{se } Z \begin{cases} \geq 0 & \text{l'individuo è favorevole all'aggregazione} \\ < 0 & \text{l'individuo è contrario all'aggregazione} \end{cases}$$

4.3 Valutazione del metodo

Partendo dalla funzione discriminante evidenziata nell'equazione (1), possiamo calcolare gli scores Z per ognuno degli individui e vedere come questa funzione

Figura 3: Coordinate delle modalità sui due primi assi dell'analisi delle corrispondenze multiple (ACM)



li classifichi nei due gruppi di riferimento. I risultati sono evidenziati nella seguente matrice delle confusioni. L'interpretazione della matrice è la seguente:

	0	1	totale
0	153	10	163
1	15	53	68

sulla prima riga osserviamo che fra i 163 individui che fanno parte del gruppo "contrario all'aggregazione", 153 sono classificati correttamente (vale a dire il 93,8%) mentre 10 sono classificati nel gruppo sbagliato; sulla seconda riga osserviamo che fra i 68 individui "favorevoli all'aggregazione" 53 sono classificati in modo corretto (vale a dire il 77,9%) mentre 15 sono classificati nel gruppo sbagliato. Questa matrice, essendo stata calcolata a partire dagli stessi dati sui quali è stata fatta la stima della funzione discriminante non risulta essere una buona misura della qualità del modello. Per questo motivo, per migliorare la misura del modello, vediamo di utilizzare dei **metodi non parmetrici** di ricampionamento.

4.3.1 Validità incrociata (Jackknife)

La validità incrociata consiste nell'effettuare un'analisi discriminante togliendo un'osservazione alla volta (un individuo) per utilizzare in seguito i risultati della stima per classificare l'individuo che è stato tolto nella stima della funzione discriminante. Questa procedura viene effettuata 231 volte (ogni volta si elimina un individuo diverso) ottenenedo così una matrice delle confusioni contenente le previsioni fatte indipendentemente da ogni individuo.

In altre parole, a partire dai fattori ottenuti con l'analisi delle corrisponden-

ze multiple, che costituiscono le nuove variabili continue, ogni volta viene tolto un individuo e viene stimata la funzione discriminante e in seguito, con questa funzione, si classifica l'individuo che è stato tolto. I risultati di questa procedura sono evidenziati nella seguente matrice delle confusioni. Fra i 163 individui con-

	0	1	totale
0	125	38	163
1	7	61	68

trari (prima riga della matrice) 125 sono classificati in maniera corretta $(76,6\%)$ ciò che eivdenzia una buona performance del modello. Per quanto riguarda invece gli individui favorevoli all'aggregazione (seconda riga della matrice) 61 su 68 individui sono stati classificati correttamente $(89,7\%)$. Globalmente, con la validitá incrociata, otteniamo una percentuale di individui classificati in modo corretto di $\frac{125+61}{231} = 80,5\%$.

5 La regressione logistica

In questa sezione effettuiamo una regressione logistica sulla variabile dipendente *favorevolenovazzano*, con l'obbiettivo di arrivare a discriminare gli individui favorevoli da quelli contrari all'aggregazione del Comune di Novazzano.

5.1 Descrizione del metodo e modellizzazione

La regressione logistica è appropriata per la modellizzazione di una variabile dicotomica $Y = 0/1$. L'obbiettivo è di modellizzare la probabilità condizionata di Y conoscendo i valori delle variabili esplicative $X_1, \ldots, X_p$:

$$\pi(x) = \text{Prob}(Y = 1|X = x)$$

Il modello lineare classico del tipo:

$$\pi(x) = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p$$

non é appropriato e il modello logistico é piú naturale. La regressione logistica modellizza la probabilitá condizionata della variabile dicotomica nel seguente modo:

$$\pi(x) = \frac{e^{\beta_0+\beta_1\cdot x_1+\ldots+\beta_p\cdot x_p}}{1 + e^{\beta_0+\beta_1\cdot x_1+\ldots+\beta_p\cdot x_p}}$$

o, in maniera equivalente:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p$$

Il rapporto

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\text{Prob}(Y = 1|X = x)}{1 - \text{Prob}(Y = 1|X = x)} = \frac{\text{Prob}(Y = 1|X = x)}{\text{Prob}(Y = 0|X = x)}$$

é chiamato *odds-ratio*. Il logaritmo naturale di questo rapporto é chiamato *log-odds* o semplicemente *logit*.

5.2 Primi risultati e selezione delle variabili

Nel nostro caso facciamo dapprima una stima logit utilizzando tutte le variabile esplicative. A partire da questa stima con tutte le variabili esplicative, applicheremo la procedura "backward-selection" basata sul criterio di Akaike. Il criterio di Akaike (AIC) è definito nel seguente modo::

$$AIC = -2 \cdot \log(\text{verosimiglianza}) + 2 \cdot \#\text{ dei parametri}$$

e serve per paragonare la qualitá dei modelli con diversi parametri. Lo scopo sarà quello di minimizzare il criterio di Akaike.

I risultati della stima di tutte le variabili (come detto tutte le variabili sono categoriali), sono riassunti nella tabella 2.

Tabella 2: Regressione logistica con tutte le variabili esplicative e categoriali

	Estimate	Std. Error	z value	p-value
(Intercept)	−6.6225	2.1940	−3.02	0.0025
sessoM	−1.3637	0.8363	−1.63	0.1030
stato-civileconiugato	0.1979	0.8962	0.22	0.8252
stato-civiledivorziato	−0.1411	1.5049	−0.09	0.9253
stato-civileseparato	11.6439	2399.5454	0.00	0.9961
stato-civilevedovo	1.5893	1.6693	0.95	0.3410
strato-resr+10	0.6221	0.8089	0.77	0.4418
strato-resr0-5	1.3960	1.2723	1.10	0.2725
strato-resr6-10	−0.2375	1.1434	−0.21	0.8355
strato-etae18-30	0.2686	2.0658	0.13	0.8965
strato-etae31-50	−0.9866	1.3492	−0.73	0.4646
strato-etae51-70	0.1899	0.9181	0.21	0.8361
statutoprofessionaleDipendente	1.7073	1.2303	1.39	0.1653
statutoprofessionaleDisoccupato	2.3959	4.0212	0.60	0.5513
statutoprofessionaleIndipendente	3.0366	1.5388	1.97	0.0484
statutoprofessionalePensionato	0.8147	1.3042	0.62	0.5322
statutoprofessionaleStudente	0.2240	2.0427	0.11	0.9127
legamenovazzanoMolto	−0.6144	0.7347	−0.84	0.4030
legamenovazzanoPer niente	−0.4089	2.2142	−0.18	0.8535
legamenovazzanoPoco	1.8360	1.6101	1.14	0.2542
fiducianovazzanoMolta	−0.0549	0.6202	−0.09	0.9294
fiducianovazzanoNulla	15.5033	1684.9962	0.01	0.9927
fiducianovazzanoPoca	1.8152	1.5476	1.17	0.2408
favorevolecantonaleNon so	2.3180	0.9559	2.43	0.0153
favorevolecantonaleSi	5.9536	0.9606	6.20	0.0000
studioaggregazioniSi	2.4124	0.8519	2.83	0.0046
				AIC= 146.20

Va sottolineato che con tutte le variabili categoriali il numero di coefficienti da stimare diventa maggiore rispetto al numero delle variabili in quanto per ogni variabile viene stimato un numero di coefficienti pari al numero delle

modalità meno uno. Per esempio per la variabile *stro-eta* che ha 4 modalità saranno stimati 3 coefficienti in quanto la modalità non stimata sarà quella di riferimento. I coefficienti stimati vanno perciò interpretati in riferimento alla modalità omessa. Effettuando una procedura "backward-selection" si ottengono i risultati evidenziati nella tabella 3.

Tabella 3: Risultati della "backward-selection"

	Estimate	Std. Error	z value	p-value
(Intercept)	−4.5298	0.9262	−4.89	0.0000
legamenovazzanoMolto	−1.1639	0.6197	−1.88	0.0603
legamenovazzanoPer niente	−0.2954	1.3237	−0.22	0.8234
legamenovazzanoPoco	1.3084	1.2522	1.04	0.2961
favorevolecantonaleNon so	1.7799	0.8157	2.18	0.0291
favorevolecantonaleSi	5.0189	0.7140	7.03	0.0000
studioaggregazioniSi	2.0323	0.6800	2.99	0.0028
				AIC= 122.86

Le variabili tenute dal modello sono 3: *legamenovazzano*, *favorevolecantonale* e *studioaggregazioni*e il criterio di Akaike è passato da 146.20 a 122.86. Le seguenti osservazioni possono essere dedotte:

- per la variabile *legamenovazzano*: sembra che chi si identifica molto con Novazzano (ha un forte legame con il paese) abbia un effetto negativo sulla probabilità di aggregazione (coefficiente significativo);
- per la variabile *favorevolecantonale*: le due modalità sono significative ($p - value < 5\%$) il che evidenzia che chi accetta la politica condotta dal Governo cantonale in materia di aggregazioni comunali ha un effetto positivo sulla probabilità di aggregazione del Comune di Novazzano rispetto alla categoria di riferimento (individui non favorevoli alla politica del Governo cantonale);
- per la variabile *studioaggregazioni*: anche per questa variabile sembra che chi voglia far effettuare uno studio approfondito da parte del Comune di Novazzano sul tema delle aggregazioni abbia una probabilità positiva sull'aggregazione del Comune di Novazzano rispetto ala categoria di riferimento (cioè agli individui che non vogliono uno studio approfondito). Anche in questo caso il coefficiente è significativo ($p - value < 5\%$).

5.3 Valutazione del metodo

In questa sezione utilizziamo la stima ottenuta precedentmente con la procedura "backward-selection". La matrice delle confusioni, basata sui dati utilizzati per la stima dei modelli (detti anche dati di *apprentissage*, è riportata qui di seguito:

Come nel caso dell'analisi discriminante, la matrice delle confusioni basata sui dati di *apprentissage* (cioè quelli utilizzati per la stima del modello) non è un buon indicatore della qualità del modello. Per questo motivo nella prossima

	0	1	totale
0	153	10	163
1	10	58	68

sezione utilizzeremo dei metodi non parametrici di ricampionamento per cercare di migliorare il giudizio sulla qualità del modello.

5.3.1 Validità incrociata (Jackknife)

La procedura é analoga a quella utilizzata nella sezione 4.3.1 a pagina 8 per l'analisi discriminante. La procedura, nel caso in questione è la seguente:

1. effettuare una stima del modello di regressione logistica levando un individuo;
2. utilizzando i parametri stimati, classificare l'individuo che è stato omesso;
3. ripetere la procedura per ognuno dei 231 individui;
4. verificare la qualità della previsione paragonando i risultati con i dati originali della variabile *favorevolenovazzano*.

La procedura è applicata al modello di regressione logistica ridotto (quello ottenuto con la procedura di "backward-selection". La matrice delle confusioni è la seguente:

	0	1	totale
0	153	10	163
1	11	57	68

Per quanto riguarda gli individui "contrari all'aggregazione", il modello classifica correttamente 153 individui su 168 vale a dire il $91,0\%$. Per gli individui "favorevoli all'aggregazione" il modello classifica correttamente 57 individui su 68 vale a dire il $83,8\%$. La percentuale degli individui classificati correttamente dal modello é di $\frac{153+57}{231} = 90,9\%$.

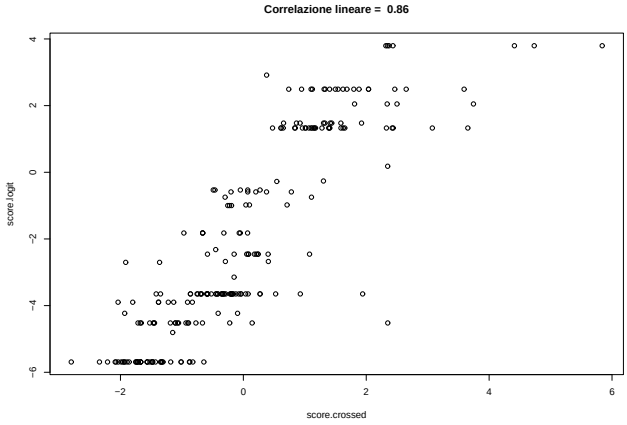
I risultati sembrano leggermente migliori rispetto a quelli ottenuti con l'analisi discriminante ($90,9\%$ contro $80,5\%$)

6 Paragone degli scores ottenuti con i due modelli

In questa sezione paragoniamo graficamente e in termini di correlazione, gli scores degli individui ottenuti con l'analisi discriminante e con la regressione logistica.

La figura 4 evidenzia gli scores per i 231 individui ottenuti con il metodo dell'**analisi discriminante** e quello della **regressione logistic**a, ed evidenzia

Figura 4: Reppresentazione grafica degli scores



una marcata **correlazione positiva**. In effetti il valore della correlazione degli scores é di $0,86$.

7 Conclusioni

In questo nostro lavoro abbiamo analizzato i dati di 231 individui con diritto di voto del Comune di Novazzano (individui iscritti al catalogo elettorale del Comune) che hanno partecipato ad un'inchiesta sul tema delle aggregazioni comunali, tramite questionario postale. Lo scopo è quello di classificare gli individui in gruppi predefiniti come "favorevoli all'aggregazione" e "contrari all'aggregazione" utilizzando due approcci differenti.

Il primo approccio che abbiamo utilizzanto è stato quello dell'analisi discriminante e più precisamente il metodo DISQUAL il quale utilizza i fattori risultanti dall'analisi delle corrispondenze multiple (variabili continue) per calcolare una funzione discriminante. Nel secondo approccio abbiamo invece utilizzato la regressione logistica.

La performance dei due modelli é stata valutata con una procedura denominata "validità incrociata" (Jackknife) che consiste nell'effettuare un numero di stime tante quante sono le osservazioni disponibili, omettendo ogni volta un'osservazione per classificare in seguito l'ossevazione omessa utilizzando i risultati della stima ottenuti senza la specifica osservazione.

Dei due modelli utilizzati quello migliore sembra essere quello ottenuto con il logit rispetto a quello dell'analisi discriminante, anche se i risultati ottenuti sono pressochè simili.

A Allegato: codice R utilizzato

```
#Importazione dei dati originali a partire dal file txt
dati<-read.table(file="C:/Documents and Settings/Roberto/Documenti/Uni_neuchatel/Multivari
attach(dati)

#Caricamento librerie varie
library(MASS)
library(xtable)
library(ade4)


# [1] "favorevole_novazzano" "sesso" "stato_civile" "strato_res" "strato_eta" "statutoprof
# [7] "legamenovazzano" "fiducianovazzano" favorevolecantonale" "studioaggregazioni"

#Grafici barplots delle variabili originali categoriche
postscript(file="./grafici/barplots.eps")
par(mfrow=c(2,5))
nomi=names(dati)
for(i in 1:ncol(dati)){
barplot(table(dati[,i]), main=nomi[i], ylab="frquence", las=2)
}
par(mfrow=c(1,1))
dev.off()

#Creazione della tabella disgiuntiva completa delle variabili esplicative
library(ade4)
disj.complet<-acm.disjonctif(dati[ , -1])
disj.complet
Z<-as.matrix(disj.complet)
B<-t(Z) %*% Z
Dmat<-diag(diag(B))
dim (Z)

#Analisi delle corrispondenze multiple

#Con libreria 'ade4'
dati.acm<-dudi.acm(dati[ , -1], scannf = FALSE, nf=2)
length(dati.acm$eig)
dati.acm<-dudi.acm(dati[ , -1], scannf = FALSE, nf=length(dati.acm$eig))
boxplot.acm(dati.acm)
scatter(dati.acm)
scatter.dudi(dati.acm)
plot(dati.acm$c1[ ,1:2], type="n")
text(x=dati.acm$c1[,1], y=dati.acm$c1[,2], labels=row.names(dati.acm$c1))
abline(v=0,h=0)

#Con libreria 'MASS'
dati.mca<-mca(dati[ , -1], abbrev=FALSE, nf=25)
```

```
#Barplot degli autovalori
dati.eigen<-dati.mca$d^2
postscript(file="./grafici/eigenplot.eps")
barplot(dati.eigen, main="Autovalori")
dev.off()

#Inerzia cumulata dei valori propri (assi fattoriali) - percentuale di varianza che spiega
dati.mca.expl<-data.frame(dati.eigen, dati.eigen/sum(dati.eigen), cumsum(dati.eigen)/sum(d
names(dati.mca.expl)<-c("Autovalori","% di varianza spiegata","% cumulata")
xtable(dati.mca.expl)

postscript(file="./grafici/mca.eps")
plot(dati.mca, rows=F, cex=0.9, cex.axis=1.3, cex.lab=1.3, main="Assi 1 e 2: 19% di var.sp
abline(v=0,h=0)
dev.off()

par(mfrow=c(1,1))

#DISQUAL: analisi discriminante con i fattori ottenuti con la ACM
disqual<-data.frame(dati$favorevole_novazzano, dati.mca$rs)
names(disqual)[1]<-"favorevole_novazzano"

disqual.lda<-lda(favorevole_novazzano ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data=disqual)
plot(disqual.lda, type = "density", dimen = 1)
disqual.pred<-predict(disqual.lda)
confusionmat<-table(disqual$favorevole_novazzano, disqual.pred$class)
confusionmat

#Validazione incrociata
prev<-numeric(nrow(dati))
score.crossed<-numeric(nrow(dati))
n<-length(prev)
for(i in 1:n){
lda.crossed<-lda(favorevole_novazzano ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data=disqual[-i,])
score.crossed[i]<-t(as.matrix(dati.mca$rs[i, 1:10])) %*% as.matrix(lda.crossed$scaling)
if(score.crossed[i]>=0) prev[i]<-1
}
confusionmat1<-table(dati$favorevole_novazzano, prev)
confusionmat1

#Regressione logistica con tutte le variabili esplicative (modello completo)
dati.logit<-glm(favorevole_novazzano ~ . , family=binomial(link="logit"), data=dati)
summary(dati.logit)
xtable(dati.logit)

#Selezione backward
dati.logit.step<-stepAIC(dati.logit, direction="backward")
summary(dati.logit.step)
xtable(dati.logit.step)
```

```
#Score logit
linear.pred<-predict(dati.logit.step)
linear.pred[linear.pred>=0]<-1
linear.pred[linear.pred<0]<-0
linear.pred
confusionmat2<-table(dati$favorevole_novazzano, linear.pred)
confusionmat2

#Validazione incrociata del modello logit ridotto
n<-nrow(dati)
dati.logit.matrix<-model.matrix(dati.logit.step)
score.logit<-numeric(n); prev.logit<-numeric(n);
for(i in 1:n){
stima<-glm(favorevole_novazzano ~ legamenovazzano + favorevolecantonale + studioaggregazio
score.logit[i]<- t(as.matrix(dati.logit.matrix[i, ])) %*% as.matrix(stima$coefficients)
if(score.logit[i]>=0) prev.logit[i]<-1
}
confusionmat3<-table(dati$favorevole_novazzano, prev.logit)
confusionmat3

#Paragone degli scores ottenuti con l'analisi discriminante e la regressione logistica
scores<-data.frame(score.crossed, score.logit)
score.cor<-cor(scores)
xtable(as.data.frame(score.cor))
postscript(file="./grafici/scores.eps")
plot(scores,main=paste("Correlazione lineare = ", round(cor(score.crossed,score.logit),2))
dev.off()
```