Benassi 28 marzo

Ripasso

Analisi discriminante: tecnica descrittiva il cui utilizzo è legato alla suddivisione del campione in gruppi. Questa avviene in due fasi: una di addestramento, che usa una variabile che indica il gruppo di appartenenza, e una di analisi, in cui la funzione discriminante precedentemente costruita viene applicata anche ai soggetti che prima non erano stati presi in considerazione. Due risultati: classificare i sogg in sottogruppi ed evidenziare quali parametri sono legati alla classificazione di gruppo. La classif viene data nella prima fase da una var che conosciamo, e quindi anche il numero di gruppi viene conosciuto prima dallo sperimentatore. Gli assunti di questa analisi sono che le var indipendenti che servono per creare la funzione discriminante devono essere a distribuzione gaussiana, mentre la var di gruppo è qualitativa, di tipo categorico. Un altro assunto è che le var ind siano poco correlate tra loro, è importante perché ciascun parametro apporti l'info riguardante il più possibile solo quella var. Se fossero correlate l'info di un parametro sarebbe correlata a quella di un'altra. E' importante anche il tipo di funzione che utilizza (quali sono le caratteristiche della funz discriminante? E' lineare, associa alla var di gruppo una serie di parametri che vengono sommati tra loro, e ognuno è moltiplicato per un coefficiente. In seguito all'attribuzione dei sogg ai gruppi si può valutare quanto sia buono il modello di riferimento dell'analisi discriminante, ovvero confrontare le frequenze (i gruppi) creati dall'analisi discriminante con quelli della var di partenza, e vedere se ci sono delle discrepanze grandi o no. Se sono piccole, vuol dire che la funz è adeguata rispetto alla classificazione in gruppi che mi sono posta, se sono grandi non sono utili alla discriminazione tra i gruppi, quindi si deve scegliere altre var o escluderne alcune. C'è anche un coefficiente che ci dice quanto ciascuna var è legata alla funzione. L'ultima fase dell'analisi discriminante è la verifica del modello, perché è una tecnica descrittiva; è possibile riflettere sui risultati ed effettuare ulteriori misure. Non è possibile calcolare delle sig, i risultati dell'analisi sono legati al campione e al tipo di var che consideriamo. Sono risultati descrittivi, non inferenziali, ci offrono una prospettiva del fenomeno che studiamo.

Un'altra tecnica utile per la sudd in gruppi è la cluster analysis. Mentre la prima fa riferimento a valori gaussiani, nella seconda i fattori che suddividono i gruppi possono essere categorici. Non serve una var di raggruppamento per distinguere i gruppi, si parte da zero, senza la conoscenza a priori di alcuni soggetti della loro appartenenza al gruppo di riferimento. Un altro aspetto è che in un tipo di cluster analysis io posso indicare il tipo di gruppi, posso dire quanti gruppi voglio che la cluster mi crei.
Ci sono due tipi di cluster:

- gerarchica: fa riferimento ad un modello gerarchico, in cui i cluster creati sono ordinati gerarchicamente, ovvero si parte da un cluster che contiene un solo sogg, poi c'è un secondo con due sogg, poi il terzo con i due di prima e uno nuovo, ecc ecc. Si parte da un cluster che è il livello più basso, è contenuto in un cluster successivo che comprende quello di partenza più un ulteriore caso. I cluster possibili sono dati da n-1, dove n è il numero di sogg, perché i cluster partono dalla relazione a due a due, quindi a coppie di soggetti. Es. il terzo cluster è considerato superiore agli altri due perché li contiene.
- Non gerarchica: la suddivisione è creata a priori dallo sperimentatore, perché lui indica il numero di gruppi (e quindi di cluster). In questo caso l'analisi dei cluster è riservata ai dati quantitativi (var qtt). Solo nell'analisi gerarchica si possono

analizzare dati categoriali.

I passaggi che caratterizzano la cluster gerarchica sono:

1. identificare le var che servono per formare i gruppi, devo utilizzare delle var che penso siano influenti nella discriminazione dei gruppi; a seconda delle var che scelgo posso avere un num di gruppi diverso;
2. scegliere il tipo di distanza fra i cluster che voglio utilizzare. La distanza è la funzione matematica che permette di creare i gruppi massimamente eterogenei tra loro. A seconda del tipo di var devo scegliere il calcolo (funzione) opportuno per creare i gruppi, per calcolare la max distanza tra i gruppi che creo. Si parla di distanze euclidee quando parliamo di distanze utilizzate per var gaussiane.
3. Valutare il numero di gruppi;
4. valutare quanto il modello è adeguato rispetto ai risultati che ci si aspettava di trovare;
5. si interpreta la soluzione ottenuta.

Identificare le var e creare le distanze: distanze create da un algoritmo, che prevede tanti gruppi quanti sono i casi, e poi unisce i casi a due a due fino ad ottenere un unico cluster. Quindi si parte da una condizione iniziale in cui ogni caso rappresenta un gruppo e poi si uniscono a due a due fino a comprendere tutti i casi possibili. Alla fine in ciascun cluster non è detto che saranno compresi tutti i gruppi. Il calcolo delle distanze si basa sulle correlazioni, generate da una matrice di correlazione, che individua le prossimità tra i casi. In una matrice è possibile sapere per ogni caso quali relazioni ha con gli altri casi. Questa relazione si misura attraverso la loro correlazione, quanto i casi sono correlati tra loro tramite il coefficiente di correlazione r, come nell'analisi fattoriale. Io individuo le prossimità tra i casi. Da questa matrice io identifico i casi più vicini tra loro e li raggruppo in modo da formare dei casi che abbiano medie più distanti possibili e più omogenei possibile al loro interno. Tutti i casi sono correlati, il coefficiente mi dice quanto, e se lo sono molto si mettono tutti nello stesso cluster. La distanza tra le medie è tra i cluster, non all'interno di uno stesso. Scelgo i casi maggiormente correlati tra loro, che vanno a costituire un cluster, che è il più possibile distante dagli altri, e quindi questa distanza può essere data ad esempio dalla media di un gruppo, ma non solo da questa. Il modello crea un numero di cluster dato da n-1 casi. Quale suddivisione scegliere? Quanti cluster per avere il modello più adeguato? Il sistema arriva ad individuare un numero molto elevato di cluster, ma sarebbe meglio scegliere il num più basso di cluster che mi consente di avere la stessa info rispetto al campione di riferimento. La soluzione finale della cluster analysis gerarchica è quella in cui c'è il max di complessità, in cui tutti i cluster sono riuniti sotto un unico cluster. Allora, come si sceglie? Ci sono diverse strategie e non c'è accordo nel definire la soluzione ottimale. Alcuni autori scelgono quel cluster che ha i coefficienti di agglomerazione (coefficienti che possiamo associare per somiglianza a quelli di correlazione più alti), altrimenti si può scegliere un numero di cluster dal tipo di relazioni che sono state considerate, e quindi in questo caso è una scelta più arbitraria, perché per ogni cluster guardo quali sono i soggetti, gli item legati tra loro e in base alle conoscenze che ho del fenomeno che studio, posso scegliere un numero di casi arbitrario, e quindi dal tipo di relazione che caratterizzano i miei cluster scelgo quello che penso caratterizzi meglio il mio fenomeno, in base alle relazioni individuate. In questo caso, nell'analisi gerarchica, scegliere il numero di gruppi resta un problema, perché in ogni caso si sceglie in modo arbitrario, cioè in base alle conoscenze del fenomeno. Effettivamente scegliere il coefficiente di

correlazione più alto significa scegliere arbitrariamente la condizione che meglio è in grado di rappresentare i miei dati. Ancora una volta l'aspetto descrittivo è il punto debole di questa analisi. A questo problema si ovvia con l'analisi dei cluster non gerarchica, anche chiamata k-medie. Si chiama così perché k mi indica il fatto che è possibile individuare attraverso questa analisi non gerarchica un numero k di gruppi in modo scelto dallo sperimentatore, ovvero è lui che individua il num di gruppi ottimale per quel tipo di fenomeno che si sta studiando. In questo caso il procedimento fa seguire al primo punto (identificazione var) l'identificazione dei gruppi, fatta dallo sperimentatore. Una volta individuato quanti gruppi devono essere creati, si calcola per ciascun gruppo il valore medio, e quindi si caratterizza il gruppo attorno ad una media. Le medie di gruppo devono essere il più possibile lontane tra loro, e devono avere attorno ad esse la minima variabilità possibile, quindi una varianza più piccola possibile. Si parla di media e varianza perché siamo nell'ambito delle distanze euclidee (valori gaussiani). Una volta valutate le medie, si passa alla valutazione ed interpretazione della soluzione, si valuta quanto è buona rispetto a ciò che ci si aspettava di trovare.
Passaggi:

1. calcolo della distanza dei gruppi sulle medie; si creano i gruppi in modo che le medie siano più diverse possibile tra loro.
2. Individuazione dei centroidi: i centroidi sono i punti che hanno come coordinate le variabili prese in considerazione.
3. Assegnazione dei soggetti ai gruppi in base alla distanza che hanno rispetto ai centroidi (in modo da minimizzare la distanza rispetto al centroide di riferimento). Più i soggetti sono vicini al centroide, più la varianza sarà ridotta rispetto al centroide di riferimento.
4. Calcolate le distanze, si cerca di individuare dei nuovi centroidi in modo da rendere minima la varianza interna e max quella tra i cluster; una volta individuato il primo step dei cluster, il modello non si ferma lì, si cerca il modello migliore rispetto al primo, attraverso un metodo si producono tanti modelli fino a quando non si ottiene il migliore. Il migliore è quello che minimizza le distanze dei soggetti rispetto al loro gruppo e massimizza le distanze tra i gruppi. Le distanze sono misurate dalla variabilità, il processo termina quando il rapporto tra la variabilità interna e quella tra i cluster è a favore della variabilità tra i cluster (esterna). Quando non c'è più un decremento significativo di questa funzione da minimizzare, il processo si ferma. [i centroidi identificano il punto di incontro di due variabili nello spazio, si calcola in modo geometrico. Il centroide è il punto in comune che hanno le variabili, riesce ad individuare e a rappresentare graficamente l'unione di queste due var in modo da prevedere quali sono i punti che alla var x corrispondono alla var y. E' il punto di incontro tra due variabili]. Tutte le var considerate devono essere trasformate in punti z, var gaussiane a media=0 e varianza=1 (unitaria). Si trasformano in punti z perché spesso le var che consideriamo hanno scale di misura diverse, ed inevitabilmente è difficile confrontarle; trasformandole in punti z è possibile in modo diretto confrontarle. Sui punti z si calcolano poi le distanze euclidee (date dalla somma dei quadrati delle differenze tra tutte le var studiate). L'importante è aver capito l'utilizzo e il procedimento, e il fatto che le var sono a distribuzione gaussiana, quindi tutte le trasformazioni avvengono in base ai valori di media e varianza, e il metodo delle distanze è utile per le gaussiane. Questo ci permette in SPSS di scegliere le opzioni del metodo da scegliere. In base alla distanza tra le var, io individuo i loro punti in comune, e in base a questi io faccio in modo che ciascun soggetto appartenga ad un

gruppo piuttosto che ad un altro.
[Esempio in SPSS rimandato perché il proiettore non va...]

Analisi della correlazione

La cluster, come molte tecniche descrittive, parte dalla matrice di correlazione delle variabili. In questa è possibile individuare tutti i coefficienti che caratterizzano le relazioni tra le variabili, è possibile dire la forza delle relazioni tra le variabili in studio. Le correlazioni possibili sono di due tipi: correlazioni parametriche o non parametriche. Tra le parametriche, il coefficiente che è più usato è l'r di Pearson, mentre se non c'è distribuzione gaussiana (le var non sono a distribuzione gaussiana) si usa il tao di Kendal (?) o rho di Spearman. Nel caso dell'analisi fattoriale, come assunto, le var sono a distribuzione gaussiana, quindi ci si basa sull'r di Pearson; esso è assimilabile al coefficiente r dell'analisi della regressione. Nell'analisi fattoriale c'è una matrice di correlazione che indica quanto ciascuna var è legata alle altre, e questa misura è data dall'r di Pearson, che misura l'associazione lineare tra due var. Può andare da -1 a +1, quando vale 0 vuol dire che non c'è relazione, se vale 1 la relazione è direttamente proporzionale, c'è perfetta corrispondenza, quando è -1 la relazione è perfettamente inversa. Il suo valore assoluto indica la forza della relazione: più il coefficiente si avvicina agli 1, più la relazione tra le var è forte. Quale è il limite di questo coefficiente preso da solo? Sta nella sua forte dipendenza dalla numerosità campionaria, quindi io posso trovare i coefficienti di correlazione molto alti, e quindi mi aspetto che indichino una buona relazione tra le var, ma questo può essere prodotto da un campione molto grande di soggetti. Per questo insieme a r di Pearson si deve sempre calcolare anche la significatività. La misura di questa tiene conto anche dell'errore di misura, della possibilità di sbagliare considerando le due var come fortemente correlate. Nella matrice di correlazione utile alla formazione dei fattori nell'analisi fattoriale non abbiamo alcuna misura della significatività; abbiamo solo indici r che indicano la forza della relazione tra le var. Anche nel tao di Kendal e rho di Spearman, la relazione è data da coefficienti con valori compresi tra -1 e +1. I dati devono essere sempre di tipo quantitativo, anche se la distribuzione è non gaussiana, devono essere quantitativi, o ordinari o a ranghi. L'analisi della correlazione viene utilizzata molto per la validazione dei questionari, perché è molto utile nella stima dell'attendibilità e della validità interna dei questionari. La stessa analisi fattoriale viene utilizzata per validare i questionari, perché è in grado di indicare qual è la struttura che sta alla base di un insieme di variabili messe in relazione tra loro. Attraverso l'analisi fattoriale è possibile vedere come alcune var siano raggruppabili tra loro e individuino dei fattori, ovvero delle ulteriori var che sono in grado di prevedere la maggior parte dell'informazione dovuta alla relazione delle var studiate. L'analisi fattoriale, come la discriminante e la cluster ha lo scopo di sintetizzare l'info e individuare dei fattori che siano riassuntivi, in grado di contenere al loro interno la max info relativa a quelle var che andiamo a considerare. Anche l'analisi fattoriale ha il suo punto di forza negli aspetti esplorativi; ci sono diversi metodi di applicazioni di questa analisi, ognuno dei quali produce risultati diversi, e a seconda della scelta dello sperimentatore è possibile evidenziare delle relazioni piuttosto che altre. Questo determina una forte arbitrarietà nell'interpretazione dei risultati. Attraverso l'analisi fattoriale vengono individuati dei fattori, che possono essere associati alle var e nella fase finale gli si può dare un nome, a seconda della relazione che ha il fattore con le var che sono state considerate. Assegnare il nome al fattore è una procedura molto arbitraria. Quello che viene fatto quindi è un'operazione che ha un valore solo arbitrario, anche perché il num di fattori viene dato senza essere rapportato

ad un errore di misura, quindi non solo si può dare un nome arbitrario, ma anche una volta individuato il numero di fattori, questo non prevede un'oscillazione intorno a se stesso. Scegliendo la tecnica posso scegliere un numero di fattori che meglio corrisponde alla teoria di riferimento. La discrezionalità dello sperimentatore influenza il risultato dell'analisi fattoriale. Esempio: nel test del QI le sottoscale a cui viene date un nome sono sottoscale create in modo arbitrario in base all'analisi fattoriale da chi ha costruito la scala. E' corrispondente agli item, alle var che sono state usate per la misura e sono fortemente correlate al fattore.

Un'altra analisi, bassata sulla forza dell'effetto, è in grado di dare una misura della relazione tra i risultati di varie ricerche che riguardano uno stesso argomento. Metanalisi: è in grado di fornire una misura dei risultati ottenuti da ricerche che riguardano uno stesso argomento. La sig da sola non ci dà un indice di quanto il fattore è legato alla var dip, ma solo un indice di errore. Una misura della forza di questo legame è data dall'effect size. La metanalisi è basata tutta sullo studio della forza dell'effetto, ed è in grado di confrontare la variabilità dovuta ai fattori considerati nelle varie ricerche rispetto alla variabilità totale. E' in grado di ovviare al fatto che le diverse ricerche sono state fatte su campioni di numerosità diversa, anche utilizzando strumenti di misura diversi, o diversi test statistici. La finalità della metanalisi è di comprendere il funzionamento del fenomeno che si sta studiando e avere un'idea di come i diversi risultati, che a volte sono discordanti, abbiano misurato in modo diverso il fenomeno, e quindi di unire i risultati delle ricerche circa lo stesso fenomeno. Il lavoro di metanalisi è molto dispendioso perché occorre raccogliere tutte le ricerche fatte su quell'argomento, quindi è un lavoro molto dispendioso che potrebbe portare a risultati difficili da interpretare. Quello che occorre fare è delimitare bene il campo di studio da analizzare, anche per avere più facilità nell'interpretazione dei risultati. Se rispetto a quel fenomeno si ritiene che comunque ci siano molte variabili interessanti che lo riguardano, si possono fare più metanalisi, dividere il lavoro di metanalisi in più parti. In una metanalisi unica è meglio considerare il minor numero di variabili. Un esempio è quello di andare a valutare i diversi studi che si sono occupati di valutare la relazione tra l'ansia e le prestazioni cognitive. Per limitare il numero di variabili si possono limitare i costrutti, dividendo ad esempio il costrutto di ansia di tratto e di stato. Il procedimento della metanalisi è:

1. raccogliere gli studii e codificarli, quindi ricavare da ognuno le var e l'analisi statistica su di esse effettuata
2. sulla base delle var, calcolare gli indici di confronto, che si basano sul calcolo della forza dell'effetto, quindi per ogni gruppo di ricerche che riguardano le stesse var, viene identificato un indice di forza dell'effetto.
3. Poi si calcola la forza dell'effetto medio, si mettono insieme tutti gli indici per verificare mediamente la forza dell'effetto
4. si interpretano i risultati della metanalisi.

Gli studi raccolti devono essere adeguati, dev'essere chiara la relazione tra lo studio e il fenomeno; la numerosità degli studi dev'essere ampia, perché c'è un errore dovuto al fatto che gli studii pubblicati a cui si fa riferimento non sono tutti quelli circa un argomento, ma