Benassi 20 marzo

Giovedì verrà ripreso l'argomento di giovedì scorso, oggi ne iniziamo uno nuovo, ovvero:

La regressione logistica

Essa è uno dei più importanti modelli non parametrici, molto avanzato.
La variabile dipendente dev'essere qualitativa; siamo nell'ambito dei modelli non parametrici, quindi si usano variabili non gaussiane. In questo caso, la var dip non ha distribuzione gaussiana ed è di tipo qualitativo; in particolare, si occupa di analizzare delle var dip qualitative dicotomiche (quelle var che possono avere solo due valori, uno escludente l'altro). Un esempio è: avere una patologia o non averla (presenza/assenza di una patologia). Queste var dip vengono studiate sia in funzione di altre var dip qualitative, come var di gruppo, che in funzione di var di tipo quantitativo. Esempio: potrei dover studiare la presenza/assenza di una malattia in funzione del genere di appartenenza e dell'età; qui si applica la regressione logistica nella sua formulazione 'migliore'. Si sta studiando la applicabilità della regressione logistica in variabili a più livelli, quindi non solo dicotomiche. Nel caso di prima, potrebbe essere utile studiare anche i diversi livelli di gravità di una patologia. Per questo tipo di analisi però l'applicazione della regressione logistica dev'essere ancora studiata. In SPSS c'è già, ma dà dei problemi, perché il meccanismo funziona per successive applicazioni, e se in queste non trova il miglior modello, il meccanismo matematico si blocca e dice di non aver trovato il modello. La regressione logistica funziona normalmente con var dicotomiche, ma nel caso della regressione logistica multinomiale ci possono essere dei problemi perché è ancora in studio.
La regressione logistica è un modello a struttura predeterminata, viene costruito a priori in base alle conoscenze del ricercatore, si stabilisce a priori var ind, dip ecc e si stimano i parametri in funzione dei valori. Procedura comune per analizzare i modelli in statistica parte dalla definizione del modello, nel quale si definiscono le relazioni tra le variabili, poi c'è la stima dei parametri e in seguito si ha la valutazione della bontà del modello e per alcuni il calcolo della significatività.

1. **definizione del modello**: come si può studiare una var dip in funzione di fattori diversi quando è qualitativa? Tramite una funzione logaritmica, siamo nell'ambito dei logaritmi, non più delle funzioni. La funzione logaritmica ci permette di trasformare la var dicotomica in var continua, e ci permette di analizzare dal punto di vista matematico questo tipo di var che si presterebbe poco ad uno studio matematico. Questa funzione logaritmica che trasforma la variabili si chiama 'logit'. Il logit della var viene visto in funzione di una somma di parametri b moltiplicati ciascuno ad un suo coefficiente – assomiglia molto al modello lineare, ma al posto della y della var abbiamo il logaritmo della y- (logit= b0+b1*x1+b2*x2...). Esempio: il logit di avere la dislessia può essere determinato da b0 che è comune a tutti i soggetti, valore di base ovvero di alpha nel modello lineare generale, corrisponde alla costante, o il valore di a nella regressione lineare semplice. A questo valore basale si sommano ad esempio b1 relativo all'avere un deficit a livello fonologico; si moltiplica per x1 calcolato per quella situazione e sommato a b2, valore relativo all'età del soggetto, che viene poi moltiplicato al valore di x2 (ci possono essere anche valori quantitativi). Ciascun parametri in questa funzione ha un suo peso nella determinazione del valore della var dip. Nello studio dell'esempio di prima mi aspetto che ogni parametro abbia un peso, e la loro forza non sta solo nel valore assoluto di b, ma anche nella significatività che accompagna ciascun b. Vedremo come questa viene calcolata.
   Come si arriva da un valore a quello di una variabile continua? Prima trasformiamo la variabile nella sua possibilità, così facendo possiamo avere infiniti valori tra 0 e 1. Invece di considerare la var come solo qualitativa la consideriamo come la prob di avere o non avere la patologia, può assumere valori tra 0 e 1. Poi possiamo trasformare questa in odds, cioè il rapporto tra la probabilità dell'evento e quella del non evento. Ci dà un'idea immediata del

rapporto tra queste due. In questo modo i valori che può assumere la var non sono più 0-1 ma vanno da 0 a infinito. Quindi io posso considerare al posto della var, l'odds della var. Esempio: la prob di avere la dislessia rispetto alla prob di non averla. L'odds si usa nelle scommesse. Se io poi trasformo l'odds nel logaritmo dell'odds, per le proprietà dei log io ottengo una var che va da – infinito a + infinito, perché il logaritmo è la funzione inversa dell'esponente. Parliamo in questo caso di log naturali, e il log naturale di x è l'esponente da dare ad e per ottenere x; e è un numero naturale, fisso, che ha valore di 2.718. Il Ln5=1.6, perché 2.718 alla 1.6 fa 5. Nell'ambito dei logaritmi le operazioni matematiche sono diverse rispetto al solito. Il logit della variabile è il logaritmo della var intesa come log dell'odds dell'evento e quindi il log della prob dell'evento diviso il non evento. Attraverso la regressione logistica io non studio solo la prob di avere una patologia, quanto la prob di avere una patologia rispetto alla prob di non averla. E' possibile studiarlo attraverso la trasformazione logaritmica della mia variabile, che mi permette di vedere i fattori che influenzano la mia var in una funzione lineare, e quindi di vedere come una somma di parametri b moltiplicati ciascuno per il coefficiente x.

Proprietà dei logaritmi:

Ln 1 = 0

Ln 0 = - infinito

Ln + infinito= (vedi lucidi)

Il logit della var può essere visto come il rapporto tra odds della var nella quale non ci sono più i parametri b sommati tra loro e moltiplicati ciascuno per il suo coefficiente, ma c'è il prodotto degli esponenti dare a ciascun parametro b. L'odds della var è dato dal prodotto tra l'esponente di ciascun parametro moltiplicato per ciascun coefficiente. Questi due modi producono lo stesso risultato, è la stessa funzione rappresentata in modo diverso (vedi lucidi per rappresentazione).
[l'importante è cogliere che la regressione logistica funziona col calcolo logaritmico, che permette di trasformare la var in continua che va da + a – infinito. Si parla di odds. Il log viene visto in funzione di parametri relativi al fenomeno che mi interessa, ciascuno moltiplicato per il coefficiente. L'importante è capire come funziona la cosa dal punto di vista logico, non serve capire come funzionano i logaritmi.]
Dev'essere chiaro che quando analizzo i risultati, questi non si riferiscono alla prob dell'evento che sto studiando, quanto al rapporto tra prob dell'evento su prob del non evento. La funzione logaritmica mi permette di pesare la forza di ciascun fattore che io penso determini la mia probabilità.

2. una volta definito il modello, composto da logit della var in funzione di parametri b sommati e moltiplicati per il coefficiente, andiamo a stimare i parametri b, quindi valutiamo il peso. Attraverso la stima di essi è già possibile avere una stima della bontà del modello, perché nel momento in cui questi sono significativamente diversi da 0, allora il modello che io ho costruito è appropriato rispetto ai dati che ho raccolto, quindi è giusto studiare quella patologia attraverso questa funzione. Come si misura il valore dei parametri b? Attraverso diversi metodi alternativi, che sono metodi di approssimazione. Posso stimare i parametri b tutti nello stesso momento nel mio modello, tramite un criterio chiamato 'di tolleranza', ovvero vengono esclusi dal modello quei parametri che apportano poca informazione (cioè cambiano poco al modello stesso, una loro modificazione non apporta una modificazione significativa, quindi è possibile escluderli dal modello in quanto irrilevanti). In SPSS si parla di var nell'equazione, mettendo nella tabella solo le var che sono determinanti nella definizione del modello, e poi dice che ce ne sono altre che però non ha considerato. L'equazione è la funzione logaritmica di prima. Noi a priori partiamo con funzione logaritmica in tutte le variabili, poi con questo metodo di SPSS riduciamo l'equazione ed escludiamo le var poco influenti, che apportano poche info e sono poco determinanti nei cambiamenti di valore di var dip. Questo metodo è detto 'a blocchi'. Alla fine abbiamo una stima per ogni parametro b che abbiamo misurato. Allo stesso modo, c'è un altro metodo (la scelta tra l'uno o l'altro si fa a priori), detto 'a passo' o 'per esclusione', in cui si tolgono o

aggiungono a seconda del livello che si sceglie i parametri in funzione dell'apporto di ciascun parametro alla sig del modello. Quindi se aggiungendo un parametro cambio qualcosa del modello, allora lo considero nell'equazione, altrimenti no e lo metto nelle var non considerate. Parto da b0, provo ad aggiungere b1, questo b1 cambia qualcosa? Sì, allora lo inserisco nell'equazione. Poi provo ad aggiungere b2 rispetto a b0 e b1, cambia qualcosa? Se sì lo aggiungo, sennò va nei fattori esclusi. E avanti così per tutti i b. E' possibile specificare anche l'effetto di due b in relazione. Queste approssimazioni vengono fatte per stimare il valore dei parametri b. Nei modelli a struttura stimata, insieme al parametro b, ho anche l'oscillazione casuale di b, cioè l'intervallo di confidenza attorno al parametro b. Questo mi permette di calcolare la sig dei b, quindi di valutare ciascun fattore rispetto alla propria sig. La direzione dell'effetto ci viene detta dall'**odds ratio**, il rapporto tra gli odds del fattore che stiamo studiando rispetto alla var dipendente. Ad es., se voglio stabilire se essere m rispetto ad essere f produce più prob di avere una certa patologia, e in che senso questo produce differenze, ce lo dice il rapporto tra l'odds di avere la patologia essendo m rispetto all'odds di avere quella patologia essendo f. Questo comporta una maggiore difficoltà nell'interpretazione dei risultati, che sono meno immediati rispetto a quelli dei test parametrici, perché siamo nell'ambito della probabilità. L'odds m è dato da prob di avere la patologia essendo m fratto la prob di non averla essendo m, e poi si mette in rapporto con l'altro odds, dato da prob di avere la patologia essendo f fratto quella di non averla essendo f.Il senso dell'effetto che studiamo è dato dal rapporto di queste probabilità. A seconda che io considero i m con le f come il fattore determinante li metterò o sopra o sotto nella mia formula. Se penso che i m avranno valori più grandi, farò in modo che il computer li consideri con valore 1 e li metterò sopra. I parametri vengono scelti attraverso un principio detto 'maximum likelyhood' (massima verosimiglianza). Si ritrova anche nel modello loglineare, ed è la probabilità che i dati sperimentali siano stati generati dal modello. Mi dice con che probabilità quei dati sperimentali sono stati generati dal modello, quindi con che prob il modello è adeguato rispetto ai dati sperimentali. Serve a stimare la bontà del modello, ovvero quanto il modello che ho creato sia aderente ai dati sperimentali ce ho raccolto. La valutazione della bontà del modello è anche data dalla 'statistica di Wald', che valuta il rapporto tra il valore di b e il suo standard error al quadrato, e attraverso questa valutazione io ho un'idea della bontà del mio modello. In realtà questo tipo di statistica ha un limite: quando il valore assoluto di b diventa molto grande, anche l'errore standard diventa molto grande per cui la statistica Wald assume valori piccoli, che facilmente falsificheranno H0 anche quando non è da falsificare. Noi difficilmente ci occuperemo di questo aspetto della regressione logistica, ovvero della bontà del modello. L'esponenziale di b ha un proprio intervallo di confidenza per ogni b considerato, e questo serve per calcolare la sig di ogni b. Come ogni modello, la sig nella regressione logistica si calcola sia per il modello globale, sia per i singoli parametri. Difficilmente siamo interessati alla sig del modello, più spesso a quella dei singoli parametri, ovvero a sapere quali dei param che consideriamo siano influenti sulla var dip. Odds ratio è esponenziale di b (vedi lucidi per formula). Il fattore **quantitativo** come viene trattato? Se noi consideriamo b1 come sex di appartenenza come b=1 è m, e b=0 è femmine. La prob di avere un comportamento aggressivo essendo f viene considerato nell'insieme. Se la var non ha solo due valori, ma è quantitativa, cosa succede? Il mio modello considera l'età media, quindi quel valore basale che accomuna tutti i soggetti con la prob di avere quella patologia. Poi b2 è quanto cambia rispetto all'aumento di 1 anno di età la prob di avere la patologia. Il valore di b è sempre relativo all'aumento di 1 anno, non è arbitrario. Nel caso in cui ci sia una var qualitativa con più di due valori il modello trasforma questa var in tante sottovariabili a ciascuna delle quali attribuisce i valori di 0 o 1. La mappa di queste sottovariabili viene data assieme ai risultati, per interpretarli (alla prima var si dà 1 al primo valore, 0 agli altri tre; nella seconda si dà 0 al primo, 1 al secondo, 0 agli altri ecc. Si trasformano in sottovariabili dicotomizzate).
Esempio (c'è anche sul libro 'tecniche psicometriche'): analisi del comportamento aggressivo

di un soggetto in funzione di un insieme di parametri. Si pensa che la prob di rispondere in modo aggressivo sia determinato da: b1= essere m, b2 = età, b3 = lavoratore dipendente. [fa vedere l'esempio in SPSS, 4 colonne con m o f, età, dipendente o superiore, aggressivo o non aggressivo. Condizione 1 è essere aggressivi, la mettiamo al numeratore. Prima usiamo metodo 'enter', a blocchi. Dice che ha considerato tutti i casi, quindi si tratta di un disegno sperimentale fattoriale, a nessun caso mancavano delle relazioni].