22.03.07
In alcuni casi è possibile applicare sia l'analisi log lineare che la regressione logistica.

**Analisi log lineare e regressione logistica: differenze**

Nella regressione logistica le variabili vengono distinte in fattori e variabile dipendente, nell'altra in realtà il modello ci dà una misura dell'associazione tra le variabili, tant'è che abbiamo detto che fattori possono essere considerate tutte le variabili che sono praticamente delle covariate. Nella regressione logistica abbiamo 1 risultato di una misura dell'effetto di quel fattore sulla variabile dipendente, nella analisi log lineare il risultato ci dà la presenza o l'assenza di una relazione fra le variabili, e quindi è una misura di quanto queste variabili sono legate fra loro, misura la relazione.

**Punti in comune:**

Entrambe si basano sulla trasformazione logaritmica della variabile e si pongono come modelli a struttura predeterminata, quindi modelli sensibili alla differenza statistica.
Abbiamo quindi una procedura in comune: la definizione del modello viene seguita dalla stima dei parametri e dalla possibile significatività dei parametri e dalla valutazione della bontà del modello.
Abbiamo detto k l'importante è valutare la bontà del modello perchè il modello viene pensato a priori rispetto ai dati, ma mi aspetto che questo modello statistico sia poi corrispondente, sia adeguato a rappresentare i miei dati; inevitabilmente ci sarà una discrepanza tra i dati e il modello, ma voglio verificare k essa sia però limitata e ci sia una buona rispondenza tra il modello e i dati.
Effettivamente questa analisi della bontà del modello viene fatta per ogni modello a struttura predeterminata. Abbiamo visto k ci sono diverse tecniche o metodi per valutare la bontà del modello e sono abbastanza comuni l'un all'altro, e dunque è possibile valutarli o considerarli analoghi, quelli k vengono fatti x l'analisi log lineare e la regressione logistica.

La maggior parte della valutazione si basa su test statistici all'inverso, basati sul $\chi^2$, che è costruito per valutare delle differenze, che vogliono dimostrare che il modello sia diverso dai dati, in realtà mi aspetto k tali differenze non ci siano, altrimenti significherebbe k il modello non è adeguato ai dati sperimentali. Non trovando 1 differenza, sono autorizzato a considerare il modello adeguato. Quindi, in sostanza, questa valutazione attaraverso il $\chi^2$ si porta dietro 1 errore metodologico, però effettivamente è l'unico metodo per valutare la bontà del modello.
In realtà questo è l'unico modo diretto, ma la bontà del modello viene anche valutata dalla stima di tutti i parametri: se essi risultano significativi, si può ritenere il modello adeguato. Se xò alcuni parametri non fossero significativi, si avrebbe una bontà parziale, una risposta spezzettata, ecco perchè si utilizzano anche le analisi goodness of feet, ecc.
Ritorniamo al discorso dell'analisi log lineare confrontata con la regressione logistica: ci sono alcuni punti in comune ed alcuni punti che le diversificano. Sicuramente quello che le diversifica sono le variabili perchè l'analisi log lineare è in grado di valutare delle variabili categoriche a più di 2 valori, mentre la regressione logistica quando la variabile ha + di 2 valori, difficilmente riesce a costruire un modello adeguato e anche l'interpretazione dei risultati che produce è difficile, e quindi se si ha una variabile con + di 2 valori si può utilizz la procedura attraverso l'analisi log lineare per analizzare le variabili.
La tavola di contigenza e la misura k si viene a creare è di tipo multidimensionale, perchè avendo un numero *n* di incroci, un numero *n* di interazioni tra le variabili maggiore di 2, corrisponde ad 1 tavola con + di 2 dimensioni.
Attraverso l'analisi log lineare è possibile studiare tutte le combinazioni possibili e tutte le interazioni possibili e si parla di modello saturo oppure solo alcune combinazioni e si parla di modello non saturato. Se voi studiate la relazione tra una variabile A e una variabile B e una variabile C, si parla di modello saturo quando avete la possibilità di analizzare tutte le singole variabili più tutte le loro interazioni: quindi il fattore legato alla variabile A, il fattore legato alla variabile B, il fattore legato alla variabile C e, dopo, tutte le interazioni legate tra A e B, tra A e C e tra B e C. Chiaram per la ricerca e per la logica inferenziale difficilm ci si interessa al modello saturo perchè, chiaramente, esso è perfettamente adeguato rispetto ai dati. Frequentemente si è interessati a rappresentare adeguatamente i dati con il minimo dei parametri. Questo è legato al principio della parsimonia del modello. Infatti: meno parametri mi servono per rappresentare adeguatamente i miei dati, e più avrò un risultato sintetico e quindi più chiaro anche da descrivere. Chiaramente mentre nel modello saturo non c'è alcuna perdita di informazione, nel modello non saturato bisogna tenere presente k si perde un po' di informazione e bisogna valutare quanta ne posso perdere.
Si parla di modello gerarchico perchè vengono definiti i fattori legati all'interazione come fattori di ordine superiore, mentre invece i fattori legati alle singole variabili, alle singole covariate, vengono chiamate fattori di ordine inferiore. Viene dunque stabilita una gerarchia tra le interazioni e i fattori singoli e si può, in un'analisi, decidere di analizz o solo le interaz o solo i fattori. Questi sono più aspetti descrittivi dell'analisi log lineare.
Effettivam spesso ci si trova nella situazione dove è + utile analizzare solo l'effetto interazione rispetto invece ai singoli fattori.
Esempio tavola di contingenza

| | | Tipo di personalità | | |
|---|---|---|---|---|
| | Terapia | A | B | C |
| Esito negativo | farmacologica | 120 | 46 | 38 |
| | integrata | 14 | 7 | 11 |
| Esito positivo | farmacologica | 28 | 64 | 147 |
| | integrata | 17 | 22 | 80 |

Vengono messe in relazione 3 variabili, tipi di personalità (A, B, C), con la terapia scelta (farmacologica o psicologica) e all'esito della terapia stessa k può essere positiva o negativa. Queste variabili sono considerate fattori del disegno sperimentale e le frequenze di cella sono considerate la variabile dipendente, perchè è sulla frequenza di cella che io vado ad applicare la funzione logaritmica dell'analisi log lineare. Quindi le frequenze di cella ci danno 1 idea dell'interazione tra le variabili, perchè io voglio vedere le frequenze di cella in rapporto al totale e quindi questo rapporto ci dice come stanno in relazione tra loro le variabili che producono quell'incrocio della tabella. Per la valutare questa relazione è necessario una trasformazione logaritmica k permette di trasformare queste frequenze in una funzione particolare che è in grado di vedere la relazione come 1 relazione lineare. Quindi come nel modello della regressione logistica, attraverso la trasformazione logaritmica la relaz tra le var viene linearizzata.
Sarebbe sbagliato applicare più test $\chi^2$ perchè un modo alternativo all'analisi log lineare sarebbe quello di vedere l'associazione a 2 a 2 delle variabili che stiamo studiando, ma ciò produrrebbe sia un aumento dell'errore $\alpha$ di una quantità non misurabile e sia una difficoltà nell'interpretazione dei risultati. Quindi la log lineare attraverso un'unica procedura riesce a darmi 1 stima dell'associazione tra le variabili. Da questa stima, da questa misura della log lineare, si hanno 2 risposte 1 riguardante il modello globale che si è costruito e 1 altra che invece riguarda ogni relazione singola e quindi ogni parametro preso singolarmente. I parametri creati chiaramente sono la misura dell'associazione tra almeno 2 variabili.
Questa è la formula della funzione log lineare k è in grado di trasformare il vostro valore, la vostra frequenza di cella nel logaritmo della frequenza, dati quei parametri che volete studiare.
Quindi il valore di ogni cella viene trasformato in 1 valore atteso ovvero in 1 valore prodotto da quella funzione logaritmica nella quale vengono considerati solo i parametri k mi interessa studiare in quel modello. In questo caso la funzione logaritmica della cella ij, quindi della cella prodotta dall'incrocio tra la var i e la var j, (in questo caso saremmo nell'ambito + semplice di una tavola, di una tabella di contingenza a 2 dimensioni, perchè abbiamo solo la dimensione i e la dimensione j, quindi l'incrocio prodotto dalla variabile i con la variabile j) questa freq viene trasformata attraverso 1 funzione logaritmica k tiene in considerazione 1 media generale $\mu$ (mu) che è il 1° termine che vedete comparire nella formula, sommato a un fattore lambda dovuto al 1° fattore considerato (i) sommato ad 1 altro fattore lambda (corrispondente ad 1 simbolo per identificare 1 parametro, nella regressione logistica venivano chiamati b, quindi è la funzione analoga a quella logistica). Nel modello saturo vengono considerati oltre ai singoli fattori anche 1 parametro lambda dovuto all'interazione tra le 2 variabili k stiamo studiando. Chiaramente il modello si complica quando la tabella, la tavola di contingenza, è multidimensionale, quando c'è anche un fattore K, lambda relativo a K e così via per le successive interazioni dovute a questo fattore. Quindi questa è la formulazione più semplice e direi che non c'è niente di nuovo rispetto a questa formula, rispetto alla regressione logistica, a parte l'interazione, ma anche rispetto al modello lineare generale.
Ancora una volta i parametri che mi danno una stima dei fattori che penso siano influenti sulla mia variabile sono sommati fra loro e sommati ad una media comune, quindi un valore basale comune a tutti i soggetti (in questo caso i soggetti corrispondono alle frequenze di cella).
C'è anche 1 esempio di modello multidimensionale a 3 dimensioni.

Il modello saturo non è xò quello preferibile per la ricerca scientifica, ma occorre identificare un modello con il minor numero di relazioni tra le variabili, di parametri, e che sia comunque in grado di rappresentare bene i dati.
Già ieri abbiamo visto quali sono i passi per la stima dei parametri. Si parte dalla tabella di contingenza dove ci sono tutte le frequenze di cella su cui devo lavorare. Ogni frequenza fa riferimento ai totali di riga e di colonna. Su queste celle, sulle frequenze osservate viene calcolato il loro logaritmo e poi vengono calcolate, in base al modello scelto, le frequenze attese, ovvero attraverso quella formula del logaritmo, del modello di riferimento, vengono ricalcolate tutte le frequenze di cella, viene calcolata 1 tabella diversa da quella di partenza e totalmente basata sul modello che abbiamo pensato. Una volta calcolate le frequenze attese occorrerà quantificare la differenza tra le 2 tabelle (quella di partenza e quella prodotta dal modello) e la differenza produce quelli che in statistica vengono chiamati residui e ci danno 1 idea di quanto il mio modello è adeguato rispetto ai dati sperimentali e hanno 1 distribuzione k può essere considerata gaussiana e quindi è possibile trasformare in punti z. (la variabile z è quella variabile gaussiana che ha media 0 e varianza 1; la trasformazione in punti z serve perchè è + facile andare ad individuare quali sono i limiti di falsificazione dell'H0, ma anche perchè così le variabili vengono ad avere la stessa scala di riferimento ed è possibile confrontarle direttamente). Attraverso la trasformazione in punti z dei residui riusciamo a vedere se quei residui sono troppo elevati (modello non adeguato) o troppo contenuti (se il modello e i dati si discostano poco tra loro, se sono abbastanza simili) 1 volta calcolati i residui, vengono stimati i parametri. Nell'esempio precedente ci sono 11 parametri; ad ogni parametro corrisponde 1 relazione tra le variabili, di tipo binario. Il 1° param viene misurato per la costante, nel 2° parametro vado a considerare la relazione tra il fatto di aver avuto esito 0 e terapia 1, il 3° parametro indica la relazione tra esito 0 e terapia 2, ecc. vedete che nella regressione log lineare ciascun parametro identifica una relazione tra le variabili di tipo binario cioè di 2 variabili. Alcuni parametri (segnalati con l'asterisco) vengono considerati ridondanti e quindi non vengono considerati, non vengono calcolati e vengono esclusi dall'analisi.
Gli stessi parametri calcolati attraverso la funzione del modello vengono anch'essi trasformati in punti z (semplicemente dividendoli per il loro errore standard) sappiamoche i limiti di falsificazione di 1 variabile z sono: + o - 1,96 e quindi se il valore dei parametri è inferiore a 1,96 sono in grado di falsificare H0.
Come viene calcolato il parametro? Occorre togliere dalla media generale gli effetti dovuti alla terapia, al tipo di personalità e all'esito. Quindi come nel caso del GML ogni parametro viene reso indipendente dagli altri perchè viene tolto l'effetto dovuto ai parametri diversi da lui rispetto alla media generale.
Dopo aver stimato i parametri e valutato la loro significatività si valuta la bontà del modello attraverso il goodness of fit test o likelyhood ratio test. Sono entrambi basati sul $\chi^2$, quindi funzionano attraverso una non falsificazione dell'ipotesi nulla, sono entrambi 2 misure di probabilità: 1 è riferita alla probab del modello, quindi è relativa alla probabilità che quel modello rappresenti bene i dati sperimentali. In questo caso la formula del $\chi^2$ è data dalla sommatoria della freq delle celle osservate meno le frequenze attese, tutte elevate al quadrato diviso le frequenze attese. Questa è la classica fo

$$\chi^2 = \sum_i \sum_j \frac{(Fij - \hat{F}ij)^2}{\hat{F}_{ij}}$$

Nel caso del likelyhood ratio test la probabilità viene calcolata sui dati sperimentali e quindi il likelyhood ratio è la probabilità che i dati sperimentali siano stati generati dal modello. La sua formula è data dal logaritmo del rapporto tra frequenze osservate e frequenze attese. Anche in questo caso la formula è data dal rapporto delle frequenze osservate con quelle attese. Chiaramente deve riferirsi alla sommatoria di tutte le possibili condizioni.

$$L^2 = 2\sum_i \sum_j F \ln \frac{F_{ij}}{\hat{F}_{ij}}$$

DOMANDE D'ESAME

1. quali sono le caratteristiche delle variabili analizzate dalla regressione logistica?
   Sia quantit che qualitative. La variabile dipendente può essere qualitativa, preferibilmente dicotomica perchè in caso contrario sono + difficili da analizzare.
2. qual è la funzione che rappresenta la regressione logistica?
   È il logit (viene chiamata così) che trasforma la variabile dicotomica nel suo logaritmo in funzione di alcuni parametri, moltiplicando ciascuno per il suo coefficiente e sommati ad un parametro basale o costante, nel caso in cui si consideri la costante un fattore determinante nella relazione tra le variabili che si stanno studiando. Quindi la costante può essere considerata oppure no, come anche, effettivamente, l'effetto interazione difficilmente viene ad essere considerato. Solitamente i fattori che si pensa siano influenti sulla variabile dipendente vengono sommati alla costante e vengono analizzati senza considerare l'effetto interazione.
3. quali trasformazioni sono attuate attraverso il logit?
   la trasformazione logaritmica delle variabili permette di rappresentare in modo lineare le variabili e di trasformare la variabile dipendente dicotomica in 1 variabile continua perchè il logaritmo, viene considerata come il logaritmo dell'odds della variabile. Abbiamo detto che la variabile attraverso il logit prima viene trasformata nella sua probabilità ovvero in valori che sono compresi da 0 a 1, poi questa probabilità viene trasformata nell'odds, ovvero nel rapporto tra la probabilità di una variabile, quindi la probabilità di un evento, fratto la probabilità di non evento. Quindi praticamente di 2 valori di probabilità, praticamente la probabilità dei 2 valori possibili della variabile vengono confrontati l'1 con l'altro attraverso l'odds. Cioè considerando invece che l'odds, il logaritmo dell'odds è possibile considerare questa variabile come una variabile continua perchè si trasforma una variabile che può avere valori che vanno da 0 a + infinito, nel caso dell'odds, in una variabile k ha da – infinito a + infinito. Questo per le proprietà dei logaritmi. Quindi la relazione tra i fattori viene resa lineare attraverso il logit e la variabile dipendente viene ad essere considerata come 1 variabile continua con valori che vanno da – infinito a + infinito perchè è trasformata nel logaritmo dell'odds
4. cosa è l'odds ratio?
   è la misura, il rapporto tra odds. Al numeratore abbiamo l'odds relativo alla condizione sperimentale avendo 1 determinato valore del fattore e sotto l'odds, relativo sempre alla condizione sperimentale, avendo però il complementare al valore di quel fattore. Per fare un esempio sopra abbiamo l'odds di avere una malattia avendo una forte esposizione al rischio per esempio l'odds del rischio di contrarre un tumore essendo fumatori, sotto abbiamo l'odds di contrarre un tumore non essendo fumatori. Quindi mi rappresenta la probabilità di contrarre la malattia rispetto alla probabilità di non contrarla essendo fumatori, quindi data quella condizione del fattore, il tutto diviso dalla probabilità di contrarre la malattia rispetto al non contrarla non avendo quella condizione, quindi non essendo fumatori. A cosa serve? Serve per interpretare i risultati significativi della regressione logistica e quindi per valutare quanto quella condizione che abbiamo posto al numeratore sia determinante sulla variabile dipendente rispetto alla condizione che abbiamo posto al denominatore. Mi dà 1 idea ad es di quanto il fatto di essere fumatori determini una modifica della probabilità di contrarre il tumore rispetto al non esserlo. L'odds ratio è possibile, proprio per la sua formulaz, tradurlo in una vera e propria probabilità perchè per ciascun soggetto è possibile dall'odds ratio ricavare la probabilità della variabile dipendente quindi ad es in questo caso è possibile calcolare la probabilità di contrarre 1 tumore date quelle condizi sperimentali. Quindi è sempre possibile passare dall'odds ratio alla probabilità e viceversa. È possibile avere 1 stima dell'odds ratio ed anche della probabilità.

5. come viene trasformata la covariata quantitativa?
   le covariate o i fattori dell'analisi della regressione logistica possono essere anche delle variabili quantitative come per es l'età. Tutte le variabili siano esse sia quantitative che qualitative vengono dicotomizzate, cioè rese dicotomiche. Quindi anche la variabile quantitativa viene resa dicotomica. Come si fa? Viene presa la condizione dell'età media come un valore e il 2° valore invece è dato dall'incremento di 1 unità rispetto all'età media. In pratica rispetto al valore medio, l'altro possibile valore che può avere la variabile quantitativa è dato dall'età media +1.
   In questo modo i risultati relativi alla covariata quantitativa come vengono tradotti? Se io trovo che l'età è significativa, cosa vuol dire? Significa che l'incremento di 1 anno di età rispetto all'età media produce delle differenze significative nella probabilità dell'evento legato a quella variabile e quindi anche l'odds ratio si riferisce all'incremento di un anno di età. Quindi ci da l'idea di quanto, aumentando l'età media di 1 anno, questo comporti una variazione nella probabilità di, ad es., avere una certa patologia o la probabilità comunque dell'evento che andiamo a studiare. Quindi l'odds ratio nel caso di una variabile quantitativa è quel coefficiente che andrebbe moltiplicato alla differenza fra l'età media e l'età reale del soggetto. Se un soggetto, abbiamo detto ieri, ha 2 o 3 anni in più rispetto all'età media, significa che, nel caso in cui questa è significativa, devo moltiplicare per quei 3 anni il coefficiente trovato tramite l'odds ratio per avere un'idea di come questo comporti una variazione rispetto alla probabilità dell'evento, della variabile dipendente.
   Che significato ha l'odds ratio di 1 variabile quantitativa?
   l'odds ratio è inferiore a 1 e significa anche 0.5, i valori, perchè per le proprietà dei logaritmi i valori inferiori a 1 non sono negativi ma hanno il valore dato dal rapporto. Perchè la probabilità non può essere negativa. Nel caso in cui l'odds ratio è = 1 siamo nel caso di assenza dell'effetto, nell'esempio fatto ieri, la posizione dell'età aveva un odds ratio pari a 1, significa che la probabilità di contrarre la malattia avendo l'età media rispetto alla probabilità di contrarla avendo l'età media +1, non aveva nessuna differenza, hanno lo stesso valore, quindi la probabilità di avere la malattia a qualsiasi età, non cambia. Se l'odds ratio è uguale a 1 significa che non c'è effetto. È il caso dell'H0 vera, anche se abbiamo detto che non si può mai verificare H0. Lo stesso vale quando tutti i parametri b hanno valore 0.
6. metodi per verificare la bontà del modello logistico
   Ce ne sono diversi. Sicuramente già dalla stima dei parametri è possibile vedere se il modello è adeguato o no. Poi anche la statistica vald può essere considerato un metodo per verificare la bontà del modello, ma ha un difetto: bisogna tener presente che quando ci sono i valori molto grandi la statistica vald anche il suo valore standard è molto grande e quindi si può falsificare facilmente H0 anche quando questa è vera. Meglio dire quando questa non è falsa.
   Poi ci sono alcuni metodi basati sul $\chi^2$ (goodness of feet o il likehood) basati sul likeed... che quindi valutano la probabilità che il modello generi i dati oppure la probabilità che i dati siano stati generati dal modello.
   Ci sono anche dei coefficienti chiamati pseudo r squared che sono dei coefficienti di relazione tra i dati ed il modello. Che sono assimilabili all'$r^2$ che è il coefficiente di determinazione nell'analisi della regressione lineare.
   Quindi ci sono diversi modi. Il + usato è la stima dei parametri e la valutazione della loro significatività. Poi ci sono metodi che sono più descrittivi: $r^2$ e quelli basati sul $\chi^2$**.** Tali metodi sono più descrittivi perchè nel caso dell' $r^2$ in realtà abbiamo una stima del rapporto fra il modello e dati, quindi non si ha una vera e propria stima anche dell'errore casuale perchè, come nel caso della regressione lineare, il coefficiente di determinazione si ottiene dal rapporto fra i dati del modello e i dati totali, quindi non abbiamo la possibilità attraverso

l'$r^2$ di conoscere anche l'errore casuale. Allo stesso modo i test basati sul $\chi^2$ sono creati sul non falsificare l'ipotesi nulla con tutte le conseguenze ed i difetti che questo comporta.

7. come vengono stimati i parametri b nella regressione logistica?
   con il modello di riferimento vengono stimati i parametri b con dei metodi di approssimazione successiva. Si creano dei parametri in base a quel modello in modo k essi siano il più rappresentativi possibili dei dati sperimentali, dopodichè si verifica la bontà del modello, poi si cerca di approssimare ancora meglio, quindi si effettua un'altra approssimazione e si ritorna a verificare la bontà del modello e si valuta: se tale modifica ha apportato delle modifiche alla bontà del modello, una stima migliore, si va avanti; se invece non c'è nessuna modifica ci si ferma lì. Quindi ci si ferma quando la modifica apportata al modello da quell'approssimazione è più piccola di un valore fisso determinato proprio dal programma. Quindi i parametri vengono stimati per successiva approssimazione. Si cerca l'approssimazione che meglio rappresenta i dati. Quindi ad ogni approssimazione viene valutata la bontà del modello. È possibile farsi dare da Spss la storia di tali approssimazioni e quindi quanto queste approssimazioni hanno migliorato o no il modello. In Spss vengono chiamate blocchi o step o passi. Ci sono 2 possibili vie: o l'approssimazione viene fatta partendo dalla costante poi considerando tutti i parametri assieme →metodo enter o a blocchi. Oppure si attua 1 metodo a passi per esclusione nel quale vengono o tolti o aggiunti dei parametri sulla base dell'apporto che essi danno alla significatività del modello. Backwards (ne toglie 1) e in avanti o forward (ne aggiunge 1).
8. quando è utile usare il modello log lineare?
   quando volete avere 1 idea della misura di associazione fra 2 o + variabili categoriche. Tutti i possibili incroci vengono dicotomizzati e vengono viste le relazioni sempre a 2 a 2 (come nel caso della correlazione) e quindi quando ad es non si può usare il $\chi^2$ perchè le variabili sono + di 2 il metodo da usare è quello dell'analisi log lineare. La correlazione è 1 misura meno informativa (ci dà solo la stima di r e la significatività di r) dell'analisi della regressione lineare (ci dà anche il valore del parametro b, l'intervallo di confidenza e questo comporta anche la stima dell'errore di b e quindi è molto più informativa. Sono molto simili come procedure, però l'analisi della regressione ha in più delle informazioni aggiuntive perchè ci da una stima dell'errore casuale e quindi anche della bontà del modello che si va ad usare). In effetti qui abbiamo parlato di test non parametrici. Il corrispettivo dell'analisi log lineare nei test parametrici è la regressione lineare multinomiale cioè nella quale ci sono più variabili, più fattori considerati.
9. Perchè il modello saturo è poco informativo?
   perchè in realtà riproduce esattamente, senza perdita di informazione, i dati osservati, le frequenze osservate invece noi siamo interessati a riprodurre quelle frequenze, anche con 1 minimo errore, con l'impiego minimo di fattori, con un numero di parametri inferiore ai parametri totali e quindi con un numero di interazioni più piccolo rispetto alle interazioni possibili, alle variabili.
10. quali sono i fattori e la variabili dipendenti nell'analisi log lineare?
    i fattori sono tutte le variabili, invece la variabili dipendenti sono le frequenze di cella
11. come vengono calcolate le frequenze attese nel modello log lineare?
    attraverso la formula del modello, quindi in riferimenti ai parametri considerati nel modello e quindi trasformando le frequenze osservate nel loro logaritmo e successivamente togliendo dalla media generale quei fattori che non interessano quella particolare combinazione che invece corrispondono all'incrocio di cella
12. Perchè i parametri vengono trasformati in punti z?
    perchè essi sono utili per la falsificazione immediata di H0 si conoscono già i limiti di falsificazione e quindi si riesce ad avere già 1 idea sull'effetto significativo o meno di quel parametro

13. come viene stimata la bontà del modello log lineare?
    attraverso il goodness of fit o il likehood ratio che stimano quanto il modello e i dati sono uguali, quindi si basano sul confronto fra le frequenze attese e le frequenze osservate e nel momento in cui non trovano una differenza tramite il test del $\chi^2$ e sperano di trovare un $\chi^2$ non significativo ovvero sperano di non trovare differenze tra le frequenze attese e le frequenze osservate. Il goodness of fit valuta la probabilità che il modello generi i dati mentre il likehood ratio valuta la probabilità che i dati siano stati generati dal modello La stessa cosa da 2 angolature diverse.