1 Normale Multivariata

E' molto simile alla normale univariata ma ha p dimensioni e si indica come N_p ed ha densità $\frac{P}{\sqrt{2\phi\Sigma}} * e^{-(x-\mu)^t\Sigma^{-1}(x-\mu)/2}$, dove μ è il vettore delle medie e Σ è la matrice di varianza covarianza. Essendo una distribuzione normale ha le seguenti proprietà:

- Basta sapere μ e Σ per identificare univocamente la distribuzione.
- Se $Cov(X_i, X_j)$ indica che le X_i sono tra di loro indipendenti.
- La combinazione lineari multivariate da una normale multivariata.
- Se X_1 e X_2 non sono indipendenti e sono rispettivamente N_r e N_p vale che $X_2|X_1$ è N_{p-r}.
- $(x-\mu)^t\Sigma^{-1}(x-\mu)$ può essere vista come una misura di distanza dei dati dalla media, che vengono poi pesati inversamente rispetto alla loro varianza. Se ne prendo la radice ottengo la distanza di Mahalonobis.
  Notiamo che $\Sigma^{-1} = I_p$ è identica alla distanza euclidea.
  Se i valori di Σ sono alti, avrò una alta dispersione dei valori dalla media.
- Sia a vettore e X distribuzione Normale multivarita alla se di corrette dimensioni a'X da come risultato una normale univariata $N(a'\mu, a' * \Sigma * a)$.
- Sia A (p*k) matrice di rango pieno (k) allora $A'X$ è $N_k(A'\mu, A' * \Sigma * A)$.
- Le marginali di X sono tutte normali univariate, però non è detto che una combinazione di normali dia una normale multivariata.
- Se X_1 e X_2 sono indipendenti e entrambi N_p allora $X_1 + X_2$ è $N_p(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$ e $X_1 - X_2$ è $N_p(\mu_x - \mu_y, \Sigma_x + \Sigma_y)$.

1.1 Stime di massima verosimiglianza

Valgono le stesse stime di massima verosimiglianza di una normale univariata infatti ho che $\mu_j = \sum x_j/n$ e $\Sigma = \sum(X_i - \hat{\mu})(X_i - \hat{\mu})^t/n = S * n/(n-1)$.
Di conseguenza avrà anche la stessa funzione di log-verosimiglianza $n*log(P) - n/2log(\Sigma) - \sum(x-\mu)^t\Sigma^{-1}(x-\mu)/2$.
Da qua è facile ricavare come $\mu_j = \sum x_j/n$ sia la stima di massima verosimiglianza, infatti devo solo massimizzare $-\sum(x-\mu)^t\Sigma^{-1}(x-\mu)/2$, essendo un quadrato negativo ha come massimo possibile 0. quindi se opto la trasformazione $\sum((x-\hat{\mu}) + (\hat{\mu}-\mu))^t\Sigma^{-1}((x-\hat{\mu}) + (\hat{\mu}-\mu)) = 0$ ricordando che $\sum(x-\hat{\mu}) = 0$ ottengo che $(\hat{\mu}-\mu)^t\Sigma^{-1}(\hat{\mu}-\mu) = 0$ quindi $\mu = \hat{\mu}$. Ovviamente essendo stima di max verosimiglianza $\hat{\mu}$ si dispone come $N_p(\mu, \Sigma/n)$.

1.2 Distribuzione di Wishart

Se definiamo $W = \sum X_i^t * X_i$ dove X_i è $N_p(0, \Sigma)$ otteniamo che W si dispone come una $Wistart_p(n, \Sigma)$, sarebbe un estensione al multi dimensionale di una chi quadro.
Noto quindi che se $Y_i \to N_p(\mu, \Sigma)$ allora $W = \sum(Y_i - \mu)^t * (Y_i - \mu)$ si dispone come $Wistart_p(n, \Sigma)$. Se uso l'approssimazione che $(n-1) * S$ allora scopra che $(n-1) * S \to Whistart_p(n-1, \Sigma)$.

1.3 Condizioni di normalità

La normalità è centrale per i metodi che vedremo successivamente quindi bisogna controllare se i dati sono normali oppure no, ciò è facile se ho una normale univariata, infatti posso confrontare la differenza con con i quantili teorici meno semplice è per una normale multivariata. Gli approcci sono svariati un elenco potrebbe essere:

- Controllare ogni componente che sia una normale univariata attraverso un QQPLOT e un QQLINE, cioè confronto i quantili teorici con quelli empirici. Questa risoluzione è teoricamente sbagliata ma operativamente accettata, infatti spesso basta normalità tra i componenti.
- Se voglio controllare i dati congiunti devo stare attento alla sparsità dei dati ecco perchè i confronti non vanno bene allora posso adoperare il seguente algoritmo:
  - Calcolo la distanza di Mahalanobis tra i dati ottenendo $D^2 = (x - \hat{\mu})^t \Sigma^{-1} (x - \hat{\mu})$.
  - Creo $U_i = n * D_i^2/(n-1)^2$ esso si dispone come una $Beta(P/2, (n-p-1)/2)$
  - Confronto i quantili empirici con quelli teorici di una beta se coincidono i dati sono normali.
- Controllo lo scatterplot dei dati combinati due a due se sono tutti con andamenti lineari ho una normale multivariata.

1.3.1 Trasformazioni Box-Cox

Se i dati non sono normali posso cercare di renderli tali usando le trasformazioni Box-Cox, tali si possono applicare sono se tutti gli $X > 0$. Si tratta di applicare una trasformazione $h(x)$ monotona tale che $\begin{cases} (X^\lambda - 1)/\lambda \textbf{ se } \lambda \neq 0 \\ log(X) \textbf{ se } \lambda = 0 \end{cases}$.
Dove il parametro λ viene trovato massimizzando la log-verosimiglianza $log(f(x)) = (\lambda - 1) \sum_i log(Y_i) - n/slog(S_x^2)$.
Infatti sia $f(x) \neq N$ ma $Y = h(x) \to N$ solo se $\prod_i h_\lambda(x_i) = (2\pi * \sigma^2)^{-n/2} * e^{-\sum (x_i - \mu)^2/2\sigma^2}$, inoltre so che $l(Y_i) = log(f(Y_i))$ quindi posso dire che $log(f(x)) = log(f(Y_i) * |J|)$ dove $|J| = \prod_i Y_i^{\lambda - 1}$.
Tale ragionamento si può applicare anche al multidimensionale però operativamente sappiamo che fare la trasformazione sull'intero dataset o componente per componente non cambia molto.

2 Classificazione secondo Bayes

2.1 Classificazione generale

Se voglio predire una variabile G, preferibilmente dicotomica, da un insieme di dati X devo creare una funzione di perdita, questa può associare ad ogni errore un peso diverso oppure usare la matrice formata da 1 ovunque e 0 (essendo il giusto) sulla diagonale e la chiameremo L_1.

Da questa funzione possiamo dedurre l'errore di predizione atteso cioè $E(L(G, G(x)))$ cioè il valore atteso di $P(G = g|X = x)$ sulla funzione di perdita L. Se lo chiamiamo $E|X$ io vorrò minimizzare tale errore con i dati a disposizione, e da questa minimizzazione trovare $G = argmin(E|X) = argmin_{g\in G} \sum_j L(G_j, g) * P(G_j|X)$, se intendiamo la nostra funzione di perdita come L_1 allora $G = argmax_{g\in G}(P(g|X))$, in particolare questa regola è detto **classificatore di Bayes**, infatti se $P(G_1|X) > P(G_2|X)$ scelgo come categoria G_1.

2.1.1 Approccio Naif

Se ho una variabile dicotomica $G_0 = 0$ e $G_1 = 1$, posso usare una regressione lineare multivarita per predire G, in particolare $G = X * B + \epsilon$ e se $G > 0.5$ la classifico come G_1 se no come G_0, dove $B = (X^tX)^{-1}X^t * G$.
Se non avessi una variabile dicotomica agirei sempre nello stesso modo però classificando K dove $K = MAX(G_i)$.
Come vedremo anche più avanti questi tipi di classificazione creano una frontiera decisionale formata da iperpiani che dividono le osservazioni in categorie.

2.2 Analisi discriminante lineare

Se $P(X|G_i)$ si dispone come $N_p(\mu_i, \Sigma)$ dove Σ è comune per tutte le distribuzioni, allora posso applicare l'analisi discriminante lineare.
Partendo da $P(G_1|X) > P(G_2|X)$ posso trovare che $P(G_1|X) = \frac{P(X|G_1)P(G_1)}{P(X=x)}$, dove conosciamo tutte queste probabilità infatti:

- $P(X|G_1)$ è la probabilità tra le categorie che è data, ed è una normale, dove $\mu = \sum X_i|G_i/n$ e $\Sigma = \sum_j \sum_{i:G=G_J}(x_i - \mu_j)(x_i - \mu_j)^t/(n - k)$.
- $P(X = x)$ sempre uguale quindi non viene mai contata
- $P(G_1) = \Phi_1$ probabilità a priori, cioè la frequenze dei gruppi, viene calcolata come $\sum 1_{G_i}/n$

Sapendo ciò sviluppo $P(G_1|X)P(G_1) > P(G_2|X)P(G_2) =$
$\frac{\Phi_1}{\Phi_2} * \frac{P}{\sqrt{2*\pi*\Sigma}} * e^{-(x-\hat{\mu_1})^t\Sigma^{-1}(x-\hat{\mu_1})/2} > \frac{P}{\sqrt{2*\pi*\Sigma}} * e^{-(x-\hat{\mu_2})^t\Sigma^{-1}(x-\hat{\mu_2})/2}$, se trasformo con il logaritmo
$2X^t\Sigma^{-1}(\mu_1 - \mu_2) - (\mu_1 + \mu_2)^t\Sigma^{-1}(\mu_1 - \mu_2) > log(\Phi_2/\Phi_1)$, se è così classifico come G_1.

Tale analisi è detta discriminante lineare visto che divide le osservazioni in base ad una frontiera lineare, infatti tale metodo può essere anche visto come $log(\frac{P(G=G_1|X}{P(G=G_2|X)}) = a_1^t * X$.
Inoltre tale metodo ha come funzione di verosimiglianza $L(\theta) = \prod_i^n \prod_j^k P(X = x, G = G_j)^{Y_{i,j}}$. Visto che G è una variabile dicotomica posso scriverlo come $L(\theta) = \prod_i^n P(X = x, G = G_1)^{Y_{i,j}} * (1 - P(X = x, G = G_1))^{1-Y_{i,j}}$

2.3 Analisi discriminante quadratica

Se $P(X|G_i)$ si dispone come $N_p(\mu_i, \Sigma_j)$ dove Σ_j è diversa tra i gruppi e si stima come $\Sigma_j = \sum_{i:G=G_j}(x_i - \mu_j)^t(x_i - \mu_j)/(n_j - 1)$ Quindi ho $\frac{\Phi_1}{\Phi_2} * \frac{P}{\sqrt{2*\pi*\Sigma_1}} * e^{-(x-\hat{\mu_1})^t\Sigma_1^{-1}(x-\hat{\mu_1})/2}$, non posso semplificare molto ed ottengo $DF = log(\Phi_j) - P/2 * log(|\Sigma|) - (x - \hat{\mu_1})^t\Sigma^{-1}(x - \hat{\mu_1})/2$.

Tale analisi è detta discriminante quadratica visto che divide le osservazioni in base ad una frontiera curva, infatti tale metodo può essere anche visto come $log(\frac{P(G=G_1|X}{P(G=G_2|X)}) = a_1^t * X + a_p * X_1^2 + a_{p+1} * X_1 * X_2$

2.4 Approccio secondo fisher

Per classificare una variabile dicotomica le G, posso ridurre la dimensionalità di X con un opportuna trasformazione, in questo modo si può proiettarlo su una retta, su questa retta posso disegnare istogrammi di densità dei due gruppi, ed il nostro obbiettivo e massimizzare la distanza tra le medie è minimizzare la varianza dentro i gruppi.

- Avendo X applico la trasformazione V, ottengo $V^t * X$ la sua forma unidimensionale.
- Voglio massimizzare $(V^t(\mu_1 - \mu_2))^2$ se sviluppo, ottengo $V^t * \Sigma_b * V$ dove Σ_b varianza tra i gruppi.
- Minimizzare $S_j^2 = \sum_{i:G=G_j}(V^t X_i - V^t\mu_j)^2$ ottenendo così $V^t * \Sigma_j * V$ dove Σ_j varianza intragruppo di un singolo gruppo, ora li unisco ed attuo una trasformazione $\sum S_j^2 * (n_j - 1)/(n - k) = V^t\Sigma_w * V^t$ dove Σ_w varianza intragruppi totale.
- Voglio $MAX\frac{V^t*\Sigma_b*V}{V^t*\Sigma_w*V}$, ottenendo così $V = \Sigma_w^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$.

Se $P(Y|G_i)$ si dispone come $N_p(\mu_i, \Sigma)$, allora la versione di fisher è identica al discriminante lineare, visto che $2X^t\Sigma_w^{-1}(\mu_1 - \mu_2) - (\mu_1 + \mu_2)^t\Sigma_w^{-1}(\mu_1 - \mu_2) > log(\Phi_2/\Phi_1)$ è uguale $2V^tX > V^t(\mu_2 + \mu_1) + log(\Phi_2/\Phi_1)$

3 Regressione logistica

Nella regressione naif otteniamo che $G = X * B + \epsilon$,invece con questo modello vediamo che $log(\frac{P(G=G_1|X=x)}{P(G=G_2|X=x)}) \rightarrow B^tX + \epsilon$ quindi stimiamo la frontiera attraverso una regressione lineare, se $P(X|G_i)$ si dispone come $N_p(\mu_i, \Sigma)$ diventa simile alla ADL o se ci si allarga al caso eteroschedastico se aggiungiamo i coefficienti quadrati otteniamo un ADQ.
Nell'analisi prendiamo omoschedasticità, se ho solo due categorie otteniamo $log(\frac{P(G=G_1|X=x)}{1-P(G=G_1|X=x)}) = B^tX + \epsilon$, così che $P(G = G_1|X = x) = \frac{e^{B^tX}}{1+e^{B^tX}}$ e $P(G = G_2|X = x) = \frac{1}{1+e^{B^tX}}$.
Se invece ho più di due categorie, devo impostare una categoria di riferimento k, avrò $log(\frac{P(G=G_j|X=x)}{P(G=G_k|X=x)}) = B_j^t * X$, quindi per ogni coppia devo stimare B_j diversi.
Se trasformo ottengo che $P(G = G_k|X = x) = \frac{1}{1+\sum_l^{k-1} e^{B_l^tX}}$ e $P(G = G_j|X = x) = \frac{e^{B_j^tX}}{1+\sum_{l=1}^{k-1} e^{B_l^tX}}$.

Ora per trovare B voglio il valore che massimizzi la massima verosimiglianza.
La verosimiglianza in un modello logistico con una variabile dicotomica come risposta è $\prod_i^n P(G = G_1|X = x)^{y_i} * (1 - P(G = G_1|X = x))^{1-y_i}$ e y_i è generato da una bernulliana.
Quindi trasformo con il logaritmo da ottenere la log-verosimiglianza e derivo per B_j, ottengo $\sum_i x_i(Y_i - P_i) = 0$, quindi $\sum_i Y_i = \sum_i P_i$, vedo che il valore atteso delle risposte è lo stesso delle risposte teoriche, notiamo inoltre che non ho una forma chiusa per stimare B e dovrò usare un algoritmo.
Bisogna verificare anche le condizioni del secondo ordine controlliamo quindi che sia concava, dalla matrice hessiana $H = -X^tP(1 - P)X$, vedo che è semi definitiva negativa sempre quindi B stimato sarà massimo.

Per stimare B usiamo l'algoritmo Newthon Rapson, cioè $B^n = B^o - H^{-1}(l) * \Delta_b L(B^o)$.
Sapendo che W matrice diagonale di $P(1 - P)$, posso scrivere $H^{-1} = -(X^t * W * X)^{-1}$ ho quindi $B^n = B^o + (X^t*W*X)^{-1}*X^t(Y-P)$ e se raccolgo $(X^t*W*X)^{-1}$, ottengo $B^n = (X^t*W*X)^{-1}X^t*W(X*B^o+W^{-1}(Y-P))$ deduco quindi che $B^n = (X^t * W * X)^{-1} * X^t * W * z$. Tale algoritmo si ferma quando $|B^n - B^o| < tol$.

Se invece guardiamo la log-verosimiglianza di una regressione logistica con $K > 2$ ottengo che $l(B) = \sum_i^n \sum_j^k y_{i,j} * log(P(G = G_j|X = x))$, dove y_i, j è generato da una multinomiale $(P_1(X_i, B_1), ..., P_n(X_i, B_n))$, con tutti 0 e un uno dove la probabilità è più alta. Spesso nei modelli più complessi per valutarne la bontà si valuta AIC.

Il rapporto $\frac{P(G=G_1|X=x)}{P(G=G_2|X=x)}$ è chiamato **ODDS**. Dato un modello di regressione logistico con i regressori B_i se ho un $B_1 * X_1$ ed X_1 varia di un unità, allora posso dire che a parità di tutte le altre variabili ho una variazione degli ODDS di essere G_1 del $e_1^B - 1$ percento.

3.1 Differenza analisi lineare e regressione logistica

I B stimati sembrano simili ma la differenza sta nelle assunzioni fatte sulla distribuzione.
Infatti la $P(G = G_1|X = x)$ è modellata nello stesso modo ma $P(X = x, G = G_1)$ no infatti dato $P(X = x, G = G_1) = P(X = x)P(G = G_1|X = x)$ la $P(X = x)$ viene modellata in maniera diversa.
In ADL abbiamo che $P(X = x) = \sum_i^k P(X = X|G = G_i) * P(G = G_i) = \sum_i^k f_n(x) * \Phi_j$ cioè una mistura di normali pesata per la loro frequenza; invece la regressione logistica non sceglie alcun modello per $P(X = x)$.
Se poi calcolo la funzione di verosimiglianza ottengo che per la ADL ho $L = \prod_i^n \prod_j^k P(X = x, G = G_j)^{y_{i,j}}$ invece per la pressione logistica ho $L = \prod_i^n \prod_j^k P(G = G_j|X = x)^{y_{i,j}}$.
Quindi la sicurezza che si dispongano come un normale i dati porta ad un miglioramento della previsione, se invece l'assunzione non è rispettata avremmo che è meglio la regressione logistica.

3.2 Altri tipi di link

Abbiamo visto la regressione logistica che è caratterizzata dal link che modifica i dati in questo caso il **logit** $\frac{e^{Bx}}{1+e^{Bx}}$, se cambiamo tale link otteniamo altri tipi di regressioni, sempre del tipo $g(P(G = G_1|X = x))$.
Per scegliere quali tra questi modelli sono migliori si può usare il criterio dell'Accuracy o del AIC.

3.2.1 Regressione probit

Una di queste alternative può essere il modello probit dove $g = \Phi^{-1}(x)$, cioè una normale inversa.

Può essere estratta anche usando una variabile aleatoria non osservata dove la risposta G è correlata alla variabile. In particolare posso dire che la variabile è $Y = B * x + \epsilon$, dove $\epsilon \to N(0, \sigma^2)$ se dico che dopo un valore τ di Y ho $G = G_1$ posso dire che $P(G = G_1|X = x) = P(y > \tau|X = x) = P(B * x + \epsilon > \tau) = P(\epsilon < B * x - \tau)$ è possibile visto che ϵ è simmetrica, da qua se moltiplico per σ ottengo $P(z < (B * x - \tau)\sigma)$ noto quindi che si può scrivere come $\Phi((B * x - \tau)\sigma) = P(G = G_1|X = x)$, invertendo trovo il risultato di prima.
Se scegliamo come $\epsilon \to t$ una T di student, riusciamo a gestire meglio gli outlier.

In questo modo inoltre diciamo che la funzione $P(G = G_1|X = x)$ si dispone come $N(-\frac{B_0}{B_1}, 1/B_1^2)$ se $B_1 > 0$, se invece $B_1 < 0$ allora è $P(G = G_2|X = x)$ a disporsi in quel modo. Inoltre sappiamo che una deviazione standard corrisponde al 68% dei dati centrali, quindi X da $P(G = G_1|X = x) = 0.16$ e $P(G = G_1|X = x) = 0.84$ hanno distanza $1/B_1$ da valore di $P(G = G_1|X = x) = 0.5$.
Posso individuare anche il punto x di pendenza massima vedendo dove la derivata seconda è $= 0$, ottengo che $X = -B_0/B_1$ ed ha derivata che vale $0.4 * B_1$. La derivata massima ha punto $X = -B_0/B_1$ anche nella regressione logistica solamente che in quel punto la derivata vale $0.25 * B_1$.
Se invece voglio avere la stessa varianza per i modelli devo imporre che $1/B_{1,p}^2 = \pi^2/3B_{1,l}^2$. Il massimo si ha in entrambi i casi quando $P = 0.5$.
Entrambi i link sono simmetrici, quindi avendo massimo in P=0.5 saranno simmetrici rispetto a dove hanno il massimo. Vale che $F(x_0 + x) = F(x_0 - x)$, i dati però non sempre sono simmetrici, quindi i modelli considerati possono produrre errori.

3.2.2 Link C-loglog

Un link asimmetrico è quello **C-loglog** cioè $P(G = G_1|X = x) = 1 - exp(-exp(B * x))$, da qua ricaviamo che $B_0 + B_1X = log(-log(1 - P(G = G_1|X = x)))$.

Se prendiamo due dataset diversi X_1 e X_2 se per vedere quali dei due è più adatto al gruppo meno adatto a $G = G_1$ vedo la differenza tra $B_0 + B_1 * X_2 - B_0 + B_1 * X_1 = B_1(X_2 - X_1)$ se elevo ottengo $\frac{log(1-P(G=G_1|X=x_2))}{log(1-P(G=G_1|X=x_1))} = e^{B_1(X_2-X_1)}$.
Inoltre si nota che cloglog va a 1 velocemente e ad 0 lentamente quindi è sbilanciato verso i numeri alti.

3.2.3 Link loglog

In questo caso $P(G = G_1|X = x) = exp(-exp(B * x))$ notiamo che va a 0 velocemente e ad 1 lentamente quindi è sbilanciato verso lo 0.

4 Logit cumulato

Ora vedremo un modo per trattenere un esempio multi categoriale ma con le categorie ordinabili secondo un certo criterio, infatti se diciamo che $G_1 < .. < G_k$ e voglio $P(G \leq G_j|X = x) = \sum_i^j P(G = G_i|X = x)$, esso si chiamano probabilità comulate e se voglio il loro logit noto che si scrive come $log(\frac{P(G \leq G_j|X=x)}{P(G>G_j|X=x)}) = log(\frac{\sum_{i=1}^{j} P(G=G_i|X=x)}{\sum_{i=j}^{k} P(G=G_i|X=x)})$, quindi la probabilità dipende da tutte le categorie, come nel caso della regressione logistica anche in questo caso $log(\frac{P(G \leq G_j|X=x)}{P(G>G_j|X=x)}) = B_{0,j} + B^t * X$ notiamo che l'unico elemento che cambia è $B_{0,j}$, si prende così per vantaggi teorici e computazionali.
Essendo G_j crescenti anche le loro probabilità lo devono essere quindi se non prendo $B^t * X$ dovrei ristimarli in maniera tale da rendere le regressione continua e crescente. Invece mantenendo fisso B e variando solo $B_{0,j}$ rispetto queste condizioni molti più facilmente.
Anche in questo caso le variabili $B_{0,j}$ e B vengono attraverso un algoritmo che massimizza la verosimiglianza tipo Newthon-Rapson, la verosimiglianza in questo caso è $\prod_i P(G = G_1|X = x)^{Y_{i,1}} \prod_{j=2}(P(G \leq G_j|X = x) - P(G \leq G_{j-1}|X = x))^{Y_{i,j}}$, dove $Y_{i,j}$ è una multinomiale del tipo $P(G = G_1|X = x)...P(G = G_k|X = x)$.

Quindi quando vado a stimare la regressione di $log(\frac{P(G \leq G_1|X=x)}{P(G>G_1|X=x)})$ e $log(\frac{P(G \leq G_2|X=x)}{P(G>G_2|X=x)})$ l'unica cosa che cambia è la costante nei regressori che sposterà la curva a sinistra da G_1 a G_2.

Se prendo due differenti dataset X_1 e X_2, e voglio fare $log(\frac{P(G_j|X=x_1)}{P(G>G_j|X=x_1)}) - log(\frac{P(G_j|X=x_2)}{P(G>G_j|X=x_2)})$ ottengo che è scrivibile come $\frac{P(G_j|X=x_1)}{P(G>G_j|X=x_1)} = e^{B(X_1-X_2)} \frac{P(G_j|X=x_2)}{P(G>G_j|X=x_2)}$, quindi gli ODDS che $Y < G_j$ quando $X = X_1$ sono uguali a quelli se $X = X_2$ moltiplicati per la costante $e^{B(X_1-X_2)}$.

Come nel probit possiamo usare una variabile latente per descrivere il logit cumulato, se definiamo $Y = -Bx + \epsilon$ dove ϵ e una variabile casuale che si dispone come una logistica standard, se introduco i valori soglia $B_{0,1} < ... < B_{0,k}$ e considero $G = G_j$ solo se $B_{0,j-1} < Y < B_{0,j}$, consideriamo quindi $P(G \leq G_j|X = x) = P(y < B_{0,j}) = P(\epsilon < B_{0,j} + Bx) = F(B_{0,j} + Bx) = \frac{e^{B_{0,j}+Bx}}{1+e^{B_{0,j}+Bx}}$.

L'interpretazione del modello è simile a quella logistica, infatti se da $log(\frac{P(G_j|X=x)}{P(G>G_j|X=x)}) = B_{0,j} + B^t * X$, prendiamo $\beta_{0,1}$ si può interpretare come: se tutti i predittori sono base line allora l'odds stimato della baseline rispetto alle altre categorie è $e^{\beta_{0,1}}$, quindi $\frac{P(G=G_1|X)}{P(G>G_1|X)} = e^{\beta_{0,1}}$, stessa interpretazione avrebbe se abbiamo $\beta_{0,2}$.
Invece se prendiamo β_1 quindi uno dei predittori fissi, vale che per un aumento di un unità di X_1, se tutti gli altri predittori sono rimasti fissi in baseline, gli ODDS che y sia baseline contro che non lo sia è pari cambia del e^{β_1} rispetto agli ODDS baseline, cioè $\frac{P(G=G_1|X=1)}{P(G>G_1|X=1)} = e^{\beta_1} * \frac{P(G=G_1|X=0)}{P(G>G_1|X=0)}$.
Ovviamente se il coefficiente $\beta_1 = 0$ allora non è influente sulle probabilità.

4.1 Altri tipi di link

Come in precedenza posso cambiare il tipo di link infatti il modello è $g(P(G < G_j|x)) = \beta_{0,j} + \beta * x$, quindi se si cambia il tipo di link avrò un tipo diverso di variabili comulate in particolare:

- Se $g(x)$=logit ho un modello logistico cumalato.
- Se $g(x)$=probit ho un modello probit cumalato.
- Se $g(x) = log(-log(1-x))$ ho un modello log-log complementare cumulato.

Lo si può notare anche se valutiamo la variabile latente $Y = -Bx + \epsilon$ ed ϵ si dispone come una normale ho un modello probit.
I modelli anche se con link diversi hanno spesso stime molto simili tra di loro, come accadeva nel caso non cumulato.

4.1.1 Modello hazard proporzionali o log-log complementare cumulato

E' un modello molto usato per stimare l'analisi di sopravvivenza o per vedere l'incidenza di un evento.
Sia T variabili continua su $(0, \infty)$, detto tempo di sopravvivenza. Abbiamo poi f(t) e F(t), funzione di densità e di distribuzione di t. Creiamo $S(t) = 1 - F(t) = P(T > t)$ funzione di sopravvivenza.

Voglio capire la probabilità che l'evento succeda in t sapendo che non è successo fin'ora, però essendo una variabile continua tale probabilità è 0, allora aggiriamo il problema costruendo $h(t) = lim_{\Delta t \to 0} \frac{P(T \in (t, t\Delta t)|T \geq t)}{\Delta t}$ ricordandoci il teorema di bayes troviamo che è uguale a $lim_{\Delta t \to 0} \frac{P(T \in (t, t\Delta t), T \geq t)}{\Delta t * P(T \geq t)} = lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t * S(t)}$ vediamo che c'è una derivata ed otteniamo che $h(t) = f(t)/S(t)$.
Ora creiamo la funzione hazard cumulata $H(t) = \int_0^t h(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds$ so che $f(s) = -(1-F(s))'$ ottengo quindi $H(t) = \int_0^t \frac{1}{1-F(s)} d(1-F(s))$ quindi $H(t) = -\log(1-F(t))$ trovo quindi che $S(t) = e^{-\int_0^t h(s)ds}$.

Se x sono i predittori ho che $h(t|x) = h_0(t) * e^{\beta * x}$, quindi $\frac{h(t|x_1)}{h(t|x_2)} = e^{\beta(x_1 - x_2)}$ quindi la differenza è una costante che non dipende da t, ed è equivalente a scrivere che $S(t|x) = S_0(t)^{e^{\beta * x}}$, con questa riscrittura notiamo che è un modello proportional odds con link log-log complementare, infatti esso è scritto come $1 - P(G < G_j|X) = S(G_j|x) = e^{-e^{\beta_{0,j} + \beta * x}} = (e^{-e^{\beta_{0,j}}})^{e^{\beta * x}}$ che è la forma scritta di prima dove $S_0(t) = e^{-e^{\beta_{0,j}}}$.

5 Logit a categorie attigue

Il modello si definisce come $logit(P(G = G_j|G \in (G_j, G_{j+1}), X = x)) = B_{0,j} + B^t * x$, dove le categorie sono ordinate come nel modello logit cumulato. Possiamo vedere che una forma equivalente della scrittura di prima è $logit(P(G = G_j|G \in (G_j, G_{j+1}), X = x)) = B_{0,j} + B^t * x = logit(\frac{P(G=G_j, G\in(G_j,G_{j+1})|X=x)}{P(G\in(G_j,G_{j+1})|X=x)})$ semplificando ho $logit(\frac{P(G=G_j|X=x)}{P(G\in(G_j,G_{j+1})|X=x)})$ se sviluppo il logit e semplifico ottengo $log(\frac{P(G=G_j|X=x)}{P(G=G_{j+1}|X=x)}) = B_{0,j} + B^t * x$, notiamo dalla formulazione che è molto simile al modello di regressione logistico multinomiale, solamente con due differenze:

- Non c'è alcuna categoria di riferimento, infatti ogni modello ha la sua categoria j-esima, a differenza di quanto accadeva nel modello multinomiale.
- Il vettore di regressione B è comune come nel modello cumulato invece la multinomiale lo ha specifico.

Anche se con queste differenze il modello multinomiale è modellabile come quello attiguo e vice versa, infatti:

- Se $\pi_j = P(G = G_j|x)$ allora posso scrivere il modello multinomiale $log(\frac{\pi_j}{\pi_k})$, ed è uguale a $log(\frac{\pi_j}{\pi_k} * \frac{\pi_{j+1}}{\pi_{j+1}} * ... * \frac{\pi_{k-1}}{\pi_{k-1}})$ ordinando ottengo che $\sum_{l=j}^{k-1} log(\frac{\pi_l}{\pi_{l+1}})$, vediamo che il modello multinomiale è visto come sommatoria di probabilità di modello attiguo.
- Il passaggio tra modello multinomiale e modello attiguo è possibile se conosco $log(\frac{\pi_l}{\pi_k})$ per $Vl \in (j, .., K-1)$. Da prima so che $log(\frac{P(G=G_j|x)}{P(G=G_k|x)}) = \sum_{l=j}^{k-1} log(\frac{P(G=G_j|x)}{P(G=G_{j+1}|x)}) = \sum_{l=j}^{k-1} B_{0,l} + B^t * x = \sum_{l=j}^{k-1} B_{0,l} + B^t(k-j) * x$, tale modello può essere riscritto come $B^*_{0,j} + B^t * U_j$, vedo che corrisponde ad un modello attiguo ma con vettore delle covariate riscalato per ogni J.

Il modello attiguo e quello cumulato hanno un fit simile dei dati, però hanno una differente interpretazione di B, infatti nel modello attiguo sono specifiche per ogni categoria, ciè moltiplico $B^t(k-j)$ con j mobile.

Inoltre a differenza del modello cumulato, posso scegliere anche B_j specifico come nel modello multinomiale, facendo così ignoro la natura ordinale di G_j ma il modello creato non è sbagliato, in particolare se $log(\frac{P(G=G_j|X=x)}{P(G=G_{j+1}|X=x)}) = B_{0,j} + B^t_j * x$ otteniamo esattamente un modello multinomiale.

Le considerazione fatte con il modello logit attiguo possono essere estese cambiando il tipo di link come abbiamo visto in precedenza.

6 Classificazione per coppie o cluster

E' un metodo molto utile in particolare per le serie storiche, infatti se abbiamo una risposta categoriale $G_{i,1}, G_{i,2}$ dove $t = 1, 2$ è il tempo, per tali categoria l'unica risposta è G_1 o G_2.
Con questo modello vogliamo modellare la dipendenza tra $G_{i,1}, G_{i,2}$, infatti è ragionevole pensare ci sia essendo una serie storica, invece i modelli trattati in precedenza non vanno bene infatti assumono l'indipendenza dei predittori.

Il primo caso che trattiamo non ha predittori, quindi lo creiamo noi in particolare ho $X_i = 0$ se ho effettuato la misurazione al tempo 1 e $X_i = 1$ se fatta al tempo 2. Possiamo sviluppare 2 modelli:

- Modello Marginale, detto così perchè modella le distribuzioni marginali di G_1 e G_2 come le misurazione effettuate nei due tempi 1 e 2. Il modello è $logit(P(G_{i,t} = G_1)) = B_0 + B_1 * x_{i,t}$. In questo modello stimo 2 parametri
- Modello condizionale, è specificato condizionatamente alle scelta fatte al tempo t-esimo. E' scritto come $logit(P(G_{i,t} = G_1)) = B_{0,i} + B_1 * x_{i,t}$ notiamo che qua stimiamo n+1 parametri, visto che l'intercetta è diversa per tutti.

Se ci concentriamo sul secondo metodo vedo che $P(G_{i,1} = G_1) = \frac{e^{B_{0,i}}}{1+e^{B_{0,i}}}$ e $P(G_{i,2} = G_1) = \frac{e^{B_{0,i}+B_1}}{1+e^{B_{0,i}+B_1}}$ noto quindi che $logit(P(G_{i,1} = G_1)) - logit(P(G_{i,2} = G_1)) = -B_1$, da qua vediamo che gli ODDS di assumere G_1 al tempo 2 sono proporzionali agli ODDs di assumere G_1 al tempo 1 infatti vale $\frac{P(G_{i,1}=G_1)}{1-P(G_{i,1}=G_1)} = e^{-B_1}\frac{P(G_{i,2}=G_1)}{1-P(G_{i,2}=G_1)}$.

Altra cosa che notiamo nel modello condizionale sempre che $G_{i,1}, G_{i,2}$ sembra vengano trattate in maniera indipendente in contraddizione in quanto detto prima, invece vediamo che $B_{0,i}$ modellano la dipendenza.
Se $|B_{0,i}| >> |B_1|$ mi si aprono due scenari:

- Se $B_{0,i} >> 0$ ottengo che $P(G_{i,1} = G_1) = \frac{e^{B_{0,i}}}{1+e^{B_{0,i}}} = 1$ visto che divisione tra numeri grandi simili, e vale per lo stesso motivo $P(G_{i,2} = G_1) = \frac{e^{B_{0,i}+B_1}}{1+e^{B_{0,i}+B_1}} = 1$ vediamo che anche se in due tempi diversi ho risultati simili.
- Se $B_{0,i} << 0$ ottengo che $P(G_{i,1} = G_1) = \frac{e^{B_{0,i}}}{1+e^{B_{0,i}}} = 0$ visto che divisione tra un numero piccolissimo e un numero vicino ad 1, e vale per lo stesso motivo $P(G_{i,2} = G_1) = \frac{e^{B_{0,i}+B_1}}{1+e^{B_{0,i}+B_1}} = 0$ anche qua nonostante i due tempi diversi ho risultati simili.

Confermato la tesi iniziale che $B_{0,i}$ modellano la dipendenza, inoltre so che più alta è la variabilità dei $B_{0,i}$ più la dipendenza è alta, se invece $B_{0,i} = B_0$ non ho dipendenza ed ottengo il modello 1.

Per stimare i coefficienti della regressione di solito uso gli stimatori di massima verosimiglianza, in questo caso però non posso applicarli visto che i parametri crescono con l'aumentare delle osservazioni, inoltre i coefficienti trovati per lo stesso motivo non hanno alcuna proprietà asintotica.
Quindi posso trattarlo in 2 diversi modi.

6.1 Verosimiglianza condizionata

Stimarli attraverso la verosimiglianza, in particolare condizioniamo rispetto alle statistiche sufficienti dei parametri $B_{0,i}$ così da escluderli, per poi trovare B_1 che è il parametro di nostro interesse.

Dobbiamo supporre che le risposte per soggetti diversi e per gli stessi soggetti siano indipendenti quindi $Y_{i,t} = 1$ se $G_{i,t} = G_1$ e $Y_{i,t} = 0$ se $G_{i,t} = G_2$.
Da qua possiamo calcolarne la verosimiglianza. Abbiamo sapendo che ci sono solo 2 tempi: $L(\beta_0, \beta_1) = \prod_i^n P(G_{i,1} = G_1)^{y_{i,1}}(1 - P(G_{i,1} = G_1))^{1-y_{i,1}} * P(G_{i,2} = G_1)^{y_{i,2}}(1 - P(G_{i,2} = G_1))^{1-y_{i,2}}$ sapendo che $P(G_{i,1} = G_1) = \frac{e^{B_{0,i}}}{1+e^{B_{0,i}}}$ e che $P(G_{i,2} = G_1) = \frac{e^{B_{0,i}+B_1}}{1+e^{B_{0,i}+B_1}}$ ottengo che $L(\beta_0, \beta_1) = \prod_i^n \frac{e^{B_{0,i}*y_{i,1}*(B_{0,i}+B_1)*y_{i,2}}}{(1+e^{B_{0,i}})(1+e^{B_{0,i}+B_1})}$, che può essere scritta come $A * e^{\sum_i^n B_{0,i}(y_{i,1}+y_{i,2})+B_2*y_{i,2}}$, da qua noto che la statistica sufficiente per $B_{0,i}$ è $S_i = y_{i,1} + y_{i,2}$ quindi $Y_i|S_i$ ha una distribuzione che non dipende dai $B_{0,i}$. Se provo a modellarla vedo che

- $P(y_{i,1} = 0, y_{i,2} = 0|S_i = 0) = 1$
- $P(y_{i,1} = 1, y_{i,2} = 1|S_i = 2) = 1$
- $P(Y_{i,1} = y_{i,1}, Y_{i,2} = y_{i,2}|S_i = 1)$ è l'unico evento non degenere, scrivibile con bayes come $\frac{P(Y_{i,1}=y_{i,1},Y_{i,2}=y_{i,2},S_i=1)}{P(S_i=1)}$ sapendo che $S_i = y_{i,1} + y_{i,2} = 1$ allora $y_{i,2} = 1 - y_{i,1}$ quindi $\frac{P(Y_{i,1}=y_{i,1},Y_{i,2}=1-y_{i,1})}{P(Y_{i,1}=1,Y_{i,2}=0)+P(Y_{i,1}=0,Y_{i,2}=1)}$ ma so anche che $Y_{i,1}$ è indipendente a $Y_{i,2}$ quindi si può scrivere come $\frac{P(Y_{i,1}=y_{i,1})P(Y_{i,2}=1-y_{i,1})}{P(Y_{i,1}=1,Y_{i,2}=0)+P(Y_{i,1}=0,Y_{i,2}=1)}$.
  Tale equazione può essere esplicitata sapendo che $P(Y_{i,1} = y_{i,1}) = \frac{e^{B_{0,i}}}{1+e^{B_{0,i}}}^{y_{i,1}} * \frac{1}{1+e^{B_{0,i}}}^{1-y_{i,1}}$ e $P(Y_{i,2} = 1 - y_{i,1}) = \frac{e^{B_{0,i}+B_1}}{1+e^{B_{0,i}*B_1}}^{1-y_{i,1}} * \frac{1}{1+e^{B_{0,i}+B_1}}^{y_{i,1}}$ con le dovute semplificazioni il numeratore è uguale a $\frac{e^{B_1-B_1*Y_{i,1}}}{1+e^{B_1}}$.
  Da qua capisco che $P(Y_{i,1} = 1, Y_{i,2} = 0|S_i = 1) = \frac{1}{1+e^{B_1}}$ e $P(Y_{i,1} = 0, Y_{i,2} = 1|S_i = 1) = \frac{e^{B_1}}{1+e^{B_1}}$

Da qua posso scrivere che la verosimiglianza di $Y_i|S_i$ è $L(B_1) = \prod_{i:S_i=1}^n (\frac{1}{1+e^{B_1}})^{Y_{i,1}} * (\frac{e^{B_1}}{1+e^{B_1}})^{Y_{i,2}}$, se n^* = numero di volte in cui $S_i = 1$, $n_{2,1}$ = numero di volte in cui $Y_{i,2} = 1$ e $n_{1,2}$ = numero di volte in cui $Y_{i,1} = 1$, dove $n_{1,2} + n_{2,1} = n^*$, sapendo ciò la verosimiglianza si può scrivere come $\frac{e^{B_1*n_{2,1}}}{(1+e^{B_1})^{n^*}}$. Da qua ricavo la logverosimiglianza $l(B_1) = B_1 * n_{2,1} - n^* log(1 + e^{B_1})$ e se derivo trovo che la stima di massima verosimiglianza è $B_1 = log(\frac{n_{2,1}}{n_{1,2}})$.

6.2 Trattare i $B_{0,i}$ come parametri di disturbo

Immaginiamo ora che $u_i = B_{0,i} - B_0$ ottengo quindi che $logit(P(G_{i,1} = G_1|u_i)) = B_0 + u_i$ e $logit(P(G_{i,2} = G_1|u_i)) = B_0 + B_1 + u_i$.
Trattiamo quindi u_i come una variabile aleatoria in particolare nel sistema logistico-normale $u_i \to N(0, \sigma^2)$ dove σ^2 non nota e u_i sono indipendenti.

Tale visione può essere estesa attraverso il modello di Rash cioè se $T > 2$ ho che $logit(P(G_{i,t} = G_1|u_i) = B_0 + B_t + u_i$, se identifichiamo $G_{i,t}$ risposta dall'i-esima persona alla t-esima domanda e se G_1 e risposta giusta e G_2 risposta sbagliata, allora avrò che la preparazione del candidato sarà data da u_i e invece la difficoltà di quella domanda sarà B_t inoltre maggiore è la preparazione maggiore che la probabilità di fare giusta quella domanda.

Il modello di Rasch può essere generalizzato e scritto come $logit(P(G_{i,t} = G_1|u_i) = B'X_{i,t} + u_i$ e potendo cambiare link posso scrivere che $g(P(G_{i,t} = G_1|u_i) = B'X_{i,t} + u_i$.

Inoltre con tale modello so che $G_{i,1}, G_{i,2}$ condizionato a u_i sono indipendenti, però se marginalizzo la distribuzione rispetto a u_i vedo che $G_{i,1}, G_{i,2}$ non sono più indipendenti, anzi ci sarà un correlazione positiva tra i due.
Ciò è dimostrabile infatti dato $Y_{i,t} = 1$ se $G_{i,t} = G_1$ e $Y_{i,t} = 0$ se $G_{i,t} = G_2$. Allora $Cov(Y_{i,1}, Y_{i,2}) = E(Cov(Y_{i,1}, Y_{i,2}|U_i)) + Cov(E(Y_{i,1}|U_i), E(Y_{i,2}|U_i))$, $Cov(Y_{i,1}, Y_{i,2}) = Cov(E(Y_{i,1}|U_i), E(Y_{i,2}|U_i))$ visto che le Y_i se condizionate sono indipendenti, inoltre so che $E(Y_{i,2}|U_i) = g^{-1}(B' * X_{i,2} + U_i)$ dove la funzione è monotona crescente, quindi la covarianza tra loro è per forza positiva.
Questo di può vedere anche attraverso una simulazione montecarlo, posso generare i valori U_i invertirli trovando così $g^{-1}(B_0 + U_i)$ e $g^{-1}(B_0 + B_1 + U_i)$, e poi ne valuto la correlazione tra questi due valori, ottengo una funzione che decresce fino a $\sigma = 5$ e poi cresce.

Questi due modelli hanno idee simili cioè ridurre l'importanza dei dati specifici cosi da stimare più facilmente i coefficienti. Il modello a parametri di disturbo è più malneabile infatti permette l'estensione di altri link tale modello è detto item-response, l'altro va solo con il logistico normale, ed inoltre se ho tanti predottori è più efficiente. Di contro richiede assunzioni forti e non verificabili su U_i.

6.2.1 Esempio applicativo

Sia $G_{i,1}...G_{i,T_i}$ osservazioni prodotte del soggetto i-esimo dove per ogni soggetti ho un diverso numero di domanda T_i e le risposte possibili sono solo G_1, G_2, se indico come Y_i il numero di risposte G_1 del soggetto i-esimo.
Voglio calcolare $\Pi_i = P(G_{i,t} = G_1)$ ho due metodi alternativi:

- Stimo $\hat{\Pi}_i = Y_i/T_i$ cioè la proporzione di risposte G_1 nel campione, ed effettivamente se abbiamo T_i grande è sensato, se piccolo porta a diversi errori.
- Stimo $\hat{\Pi}_i$ attraverso lo "schrinkage", cioè creo un modello $logit(\Pi_i) = B_0 + u_i$, sostituisco B_i con U_i, e ipotizzo che tutti gli U_i abbiano la stessa distribuzione da una $N(0, \sigma^2)$. Quindi per stimare $\hat{\Pi}_i$ userò i dati relativi a tutti gli stati. In particolare calcolato $\hat{\Pi} = \sum_i^k Y_i/n$, ho che se $\hat{\Pi}_i$ stimato con il metodo delle frequenze è $> \hat{\Pi}$ allora quello stimato con il metodo "schrinkage" è $< \hat{\Pi}_i$, vice versa se $\hat{\Pi}_i < \hat{\Pi}$.

Per stimare il secondo modello seguiamo la seguente strategia:

- Scrivo la distribuzione congiunta delle osservazioni$G_{i,1}...G_{i,T}$, condizionato agli effetti causali non osservabili gli U_i. Se ho 2 categorie ottengo $L(B, \sigma^2) = \prod_i^n \prod_t^{T_i} (\frac{e^{B' X_{i,t}+U_i}}{1+e^{B' X_{i,t}+U_i}})^{y_{i,t}} * (\frac{1}{1+e^{B' X_{i,t}+U_i}})^{1-y_{i,t}}$.
- Scrivo la distribuzione marginale delle osservazioni marginalizzando per gli effetti causali la distribuzione trovata al punto 1. Ottengo $L(B, \sigma^2)|u_i = \prod_i^n (\int_R \prod_t^{T_i} (\frac{e^{B' X_{i,t}+U_i}}{1+e^{B' X_{i,t}+U_i}})^{y_{i,t}} * (\frac{1}{1+e^{B' X_{i,t}+U_i}})^{1-y_{i,t}} * \frac{e^{-X' \Sigma^{-1} X/2}}{(2*\pi*\Sigma)^{1/2}} du_i)$
- Massimizzo la distribuzione 2 rispetto a B, σ^2.

Questo metodo quindi ha due problemi l'integrale da risolvere in maniera numerica e la condizione di normalità della variabile non osservabile che è una assunzione molto forte.

6.2.2 Soluzione non parametrica

Per ovviare ai problemi di prima decido di non fare assunzioni su U_i, la modellerò in maniera non parametrica come una distribuzione discreta che assume i valori $P(U_i = w_j) = P_j \; j \in 1...K$ dove K non è specificato.
K viene deciso in base al valore che massimizza la verosimiglianza del modello e in questo avremo una verosimiglianza più semplice infatti la distribuzione 2 può essere scritta come
$L(B, \sigma^2)|u_i = \prod_i^n \sum_j^k \prod^t {}^{T_i} (\frac{e^{B' X_{i,t}+U_i}}{1+e^{B' X_{i,t}+U_i}})^{y_{i,t}} * (\frac{1}{1+e^{B' X_{i,t}+U_i}})^{1-y_{i,t}} P_j$, quindi riusciamo a stimare in maniera più semplice il parametro.