Modelli per dati multinomiali

Risposta multinomiale

- La risposta ha distribuzione **Multinomiale**
- Sia c il numero di categorie possibili per la risposta
- Per il soggetto i-esimo la risposta è $y_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$, con $y_{ij} = 1$ se la risposta è nella categoria j, zero altrimenti.
- Sia $\pi_{ij} = Prob(Y_i = j), j = 1, 2, \ldots, c, \ \sum_j \pi_j = 1$
- $p(y_i) = \prod_{j=1}^{c} \pi_{ij}^{y_{ij}}$
- Due possibili situazioni:

1) Y è una variabile nominale con J categorie e l'ordine con cui sono elencate le categorie non è rilevante
2) Y è una variabile ordinale e l'ordinamento tra le categorie deve essere introdotto nell'analisi

Baseline category logit model

- Variabile risposta nominale che prevede c categorie
- Si formula un modello che descrive in modo simultaneo il logit per tutte le $c(c-1)/2$ **coppie** di categorie
- Si sceglie una categoria di riferimento e si considerano $(J-1)$ confronti a coppie

$$\log \frac{\pi_{i1}}{\pi_{1c}}, \log \frac{\pi_{i2}}{\pi_{1c}}, \ldots, \log \frac{\pi_{i(c-1)}}{\pi_{1c}},$$

- Il *baseline category logit model* consiste in $(J-1)$ regressioni logistiche

$$\log \frac{\pi_j}{\pi_c} = \eta_j = X\beta_j, \; j = 1, 2 \ldots, c-1$$

con X matrice di disegno $n \times p$ e $\beta_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{pj})$

Baseline category logit model

- Si parla anche di modello logit multinomiale
- I logit tra tutte le altre coppie di categorie si ottengono dai β_j di ciascuna regressione. Per due categorie a e b

$$\log \frac{\pi_a}{\pi_b} = \log \frac{\pi_a}{\pi_c} - \log \frac{\pi_b}{\pi_c} = X(\beta_a - \beta_b)$$

- Il confronto non dipende dalla categoria di riferimento
- In termini di probabilità

$$\pi_{ij} = \frac{\exp(x_i \beta_j)}{1 + \sum_{j=1}^{c-1} \exp(x_i \beta_j)}, \quad \sum_{j=1}^{c} \pi_{ij} = 1$$

con $\beta_c = 0$, i.e. $\pi_{ic} = \frac{1}{1 + \sum_{j=1}^{c-1} \exp(x_i \beta_j)}$

Alligator food choice

- Studio condotto dalla *Florida Game and Fresh Water Fish Commision* sui fattori che influenzano la scelta primaria di cibo degli alligatori
- 59 alligatori nel lago Georgia: lunghezza in metri (esplicativa) e il tipo di cibo ritrovato all'interno dello stomaco (risposta). La variabile risposta ha 3 categorie: Pesce, Invertebrati, Altro (che scegliamo come categoria di riferimento)
- Il modello stimato è

	Intercept	size
Pesce	1.618	-0.110
Invert.	5.698	-2.465

$$\log \frac{\hat{\pi}_P}{\hat{\pi}_A} = 1.62 - 0.11\ size$$

$$\log \frac{\hat{\pi}_I}{\hat{\pi}_A} = 5.70 - 2.47\ size$$

Alligator food choice

- Per ciascuna regressione le stime si interpretano come in un modello di regressione logistica per risposta binaria
- La log-quota attesa che nello stomaco di un alligatore si rinvenga pesce piuttosto che invertebrati è

$$\log \frac{\hat{\pi}_P}{\hat{\pi}_I} = \log \frac{\hat{\pi}_P}{\hat{\pi}_A} - \log \frac{\hat{\pi}_I}{\hat{\pi}_A} = -4.08 + 2.36 \; \textit{size}$$

- Per ogni metro in più di lunghezza, la quota attesa di ritrovare pesce nello stomaco piuttosto che invertebrati aumenta in modo proporzionale a $\exp(2.36) = 10.6$

Classificazione

- Il modello logit multinomiale può essere utilizzato per costruire una regola di classificazione, estendendo al caso di più categorie quanto visto nel caso dicotomico
- La capacità predittiva del modello può essere analizzata classificando le unità in base al valore osservato e al valore *previsto* della risposta
- Sulla base del modello stimato, le unità saranno classificate nel gruppo cui corrisponde la probabilità maggiore, i.e. $z_{ij} = 1$ per $j = \text{argmax}\ \hat{\pi}_{ij}, j = 1, 2, \ldots, c$, zero altrimenti
- Leave-one-out cross validation

Regressione logistica ordinale

- L'utilizzo dei modelli logit (e probit) permette di incorporare direttamente l'ordinamento delle categorie in cui si articola la risposta
- Consideriamo la probabilità cumulata che Y ricada nella categoria j o in quelle precedenti (nel senso dell'ordinamento)

$$P(Y_i \leq j) = \pi_{i1} + \pi_{i2} + \ldots + \pi_{ij}, \; j = 1, 2, \ldots, c$$

- Definiamo il logit cumulato

$$logit[P(Y_i \leq j)] = \log \frac{P(Y_{\leq} j)}{1 - P(Y_{\leq} j)} = \log \frac{\pi_{i1} + \pi_{i2} + \ldots + \pi_{ij}}{\pi_{i(j+1)} + \ldots + \pi_{ic}}$$

- Ogni logit utilizza tutte le c categorie della risposta
- Per ciascuna categoria $P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1)$

Proportional odds model

- Un modello per il j-esimo logit cumulato assume la forma di un modello di regressione logistica per risposta binaria

$$Y^* = \begin{cases} 1 & Y \leq j \\ 0 & Y > j \end{cases}$$

- Le prime j categorie costituiscono una nuova categoria e le seguenti ne formano un'altra
- Consideriamo il modello

$$logit[P(Y \leq j)] = \alpha_j + \beta x, \ j = 1, 2, \ldots, c-1$$

che prevede un effetto comune per tutti i logit cumulati ma intercette diverse

- I termini d'intercetta sono crescenti in j per ciascun x fissato
- Richiede un solo vettore di coefficienti β invece che $(c-1)$ vettori distinti

Inferenza

- La log-verosimiglianza multinomiale

$$\begin{aligned}\ell(\alpha,\beta) &= \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log\pi_{ij} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log\left[F(a_j-\eta_i)-F(a_{j-1}-\eta_i)\right]\end{aligned}$$

- Il modello per variabile latente descrive solo effetti di posizione ma assume la stessa variabilità il che implica che le esplicative sono stocasticamente ordinate rispetto alla risposta, i.e.

$$Pr(Y_i \leq j|X=x_1) \leq (\geq) Pr(Y_i \leq j|X=x_2), \forall j$$

- Se il modello stimato fornisce un adattamento non soddisfacente ai dati, spesso questa circostanza dipende dal fatto che la variabilità di Y varia con X
- Si può formulare un modello più complesso che preveda effetti variabili tra i logit cumulati

Proportional odds model

- Per due valori distinti x_1 e x_2 della variabile esplicativa X

$$OR = \frac{P(Y \leq j|X = x_2)/P(Y > j|X = x_2)}{P(Y \leq j|X = x_1)/P(Y > j|X = x_1)} = \exp\left[\beta(x_2 - x_1)\right]$$

- L'OR cumulato è proporzionale a $x_2 - x_1$, $\forall j$, avendo fissato i valori delle altre covariate
- Consideriamo il modello con variabile latente Y^*, tale che

$$y_i = J \Leftrightarrow a_{j-1} < y_i^* \leq a_j$$

con $-\infty = a_0 < a_1 < a_2 < \ldots < a_{c-1} < a_c = +\infty$

$$\begin{aligned} Pr(Y_i = j) &= Pr(a_{j-1} < Y_i^* \leq a_j) = Pr(a_{j-1} < \eta_i + \epsilon_i^* \leq a_j) \\ &= Pr(a_{j-1} - \eta_i < \epsilon_i^* \leq a_j - \eta_i) \\ &= F(a_j - \eta_i) - F(a_{j-1} - \eta_i) \end{aligned}$$

Proportional odds model

- Attenzione alla parametrizzazione adottata nel predittore lineare
- Se si scrive il predittore lineare nella forma $\eta_{ij} = \alpha_j + \eta_i$, con

$$logit[P(Y_i \leq j)] = \eta_{ij} = \alpha_j + \beta x_i$$

  allora $\beta > 0$ indica che la risposta Y tende a collocarsi verso il basso dell'ordinamento al crescere di x
- Se l'ordine delle categorie viene rovesciato cambia solo il segno della stima di β.
- Se $\eta_{ij} = \alpha_j - \eta_i$, allora $\beta > 0$ corrisponde a risposte che con maggiore probabilità si collocano verso l'alto dell'ordinamento al crescere di x
- Cumulative link models: scelte alternative per la funzione legame rispetto al logit

Political Ideology

	Political Ideology					
Party Affiliation	Very liberal	Slightly liberal	Moderate conservative	Slightly conservative	Very conservative	Total
Democratic	80	81	171	41	55	428
Repubican	30	46	148	84	99	407

- Political Social Survey, 1991
- Ideologia politica è la variabile risposta per la quale si assume una scala ordinale che va da *very liberal* a *very conservative*
- Si vuole studiare la relazione tra ideologia politica e partito di appartenenza
- Sia x una variabile dicotomica che denota l'affiliazione politica con $x = 1$ per i democratici

Political Ideology

- Il modello è

$$logit[P(Y \leq j)] = \alpha_j + \beta x, \ j = 1, 2, 3, 4$$

- La SMV di β è $\hat{\beta} = 1.070(0.144)$
- La quota attesa che un Democratico si muova nella direzione liberale piuttosto che conservatrice è $e^{\hat{\beta}} = 2.917$ volte la stessa quota per un Repubblicano (circa il triplo)
- Il modello stima un effetto simile anche se si decide di ridurre il numero di categorie $\rightarrow$ invarianza rispetto alla scelta delle categorie
- Riducendo il numero di categorie si va incontro ad una perdita di efficienza

Mental impairment

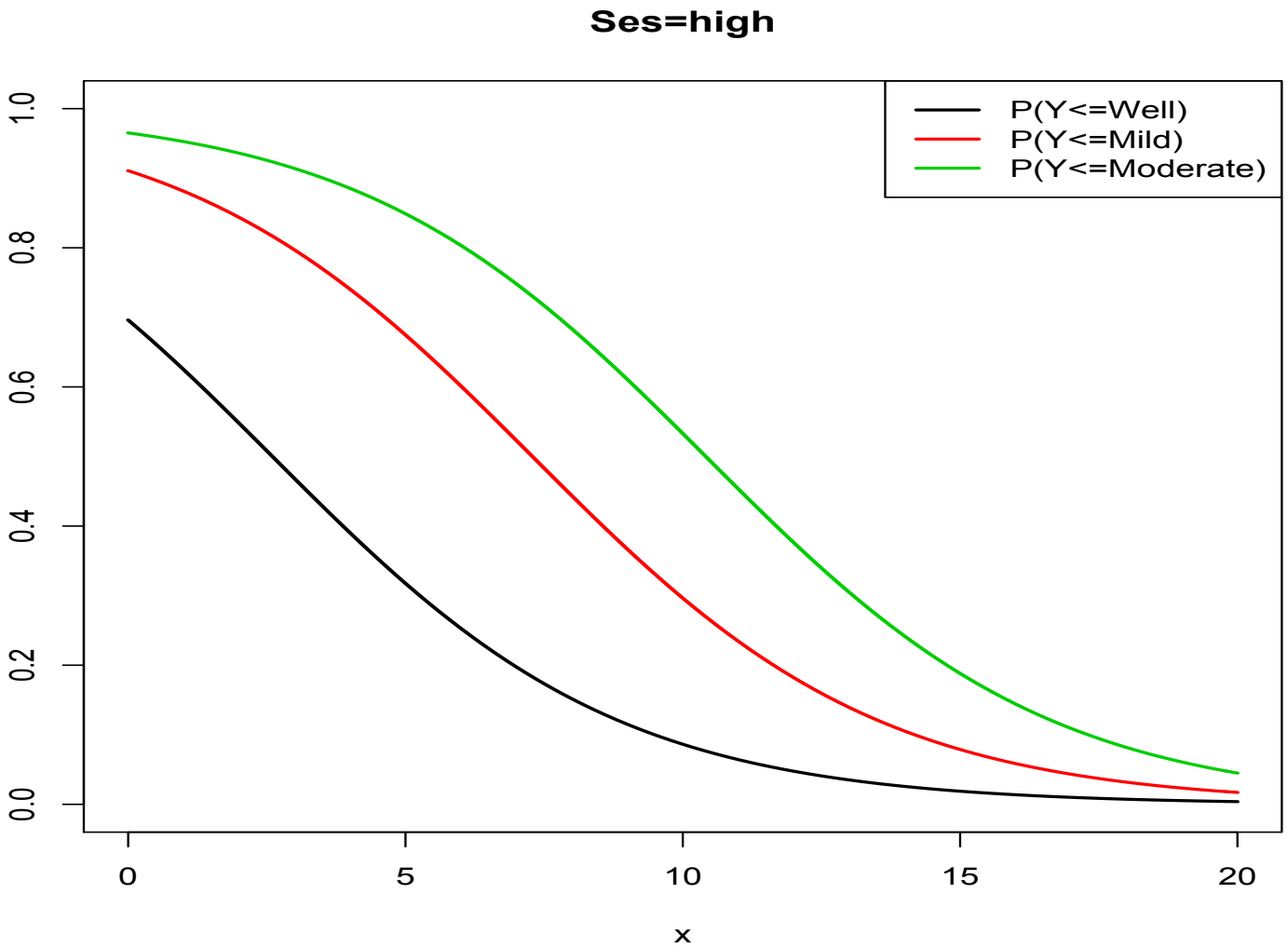
- Risposta ordinale: well, mild, moderate, impaired
- Covariate: life event index (numero di episodi significativi negli ultimi 3 anni, X), condizione socio-economica (binaria, low=0, high=1, S)
- Il modello stimato è

$$logit[\hat{P}(Y_i \leq j)] = \hat{\alpha}_j + 1.11s - 0.32x, \ j = 1, 2, 3$$

- Le probabilità cumulate che iniziano da `well` decrescono all'aumentare del numero di eventi ed aumentano quando la condizione socio-economica è favorevole
- Per un fissato numero di eventi, la quota di individui con salute mentale sotto una certa soglia (quindi tendente alla buona salute) con elevato stato socio-economico è circa il triplo della stessa quota quando lo stato socio-economico è basso
- A parità di condizione socio-economica, la quota di individui con uno stato mentale che si sposta verso la condizione `well` si riduce del 27% circa per ogni evento significativo in più

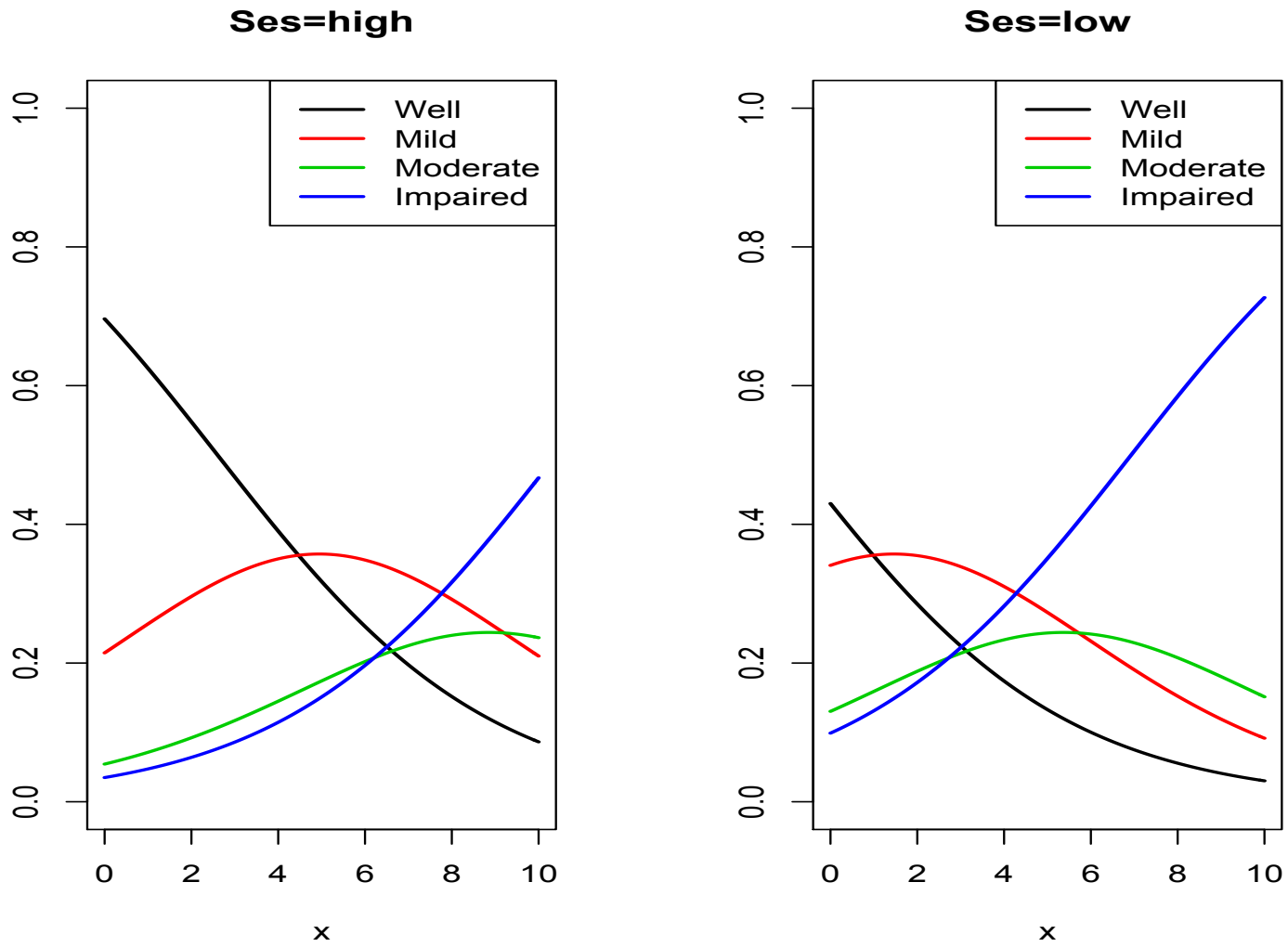
Mental impairment

Le curve non si intersecano



Mental impairment

Andamento delle probabilità stimate in ciascuna categoria



Adjacent Category model

- Si considerano $(c-1)$ logit usando tutte le coppie di categorie adiacenti

$$\log \frac{\mu_{j+1}}{\mu_j},\ j = 1, 2, \ldots, c-1$$

- Nel caso di una sola variabile esplicativa

$$\log \frac{\mu_{j+1}}{\mu_j} = \alpha_j + \beta x$$

come nel *baseline category model*

- Si assume che l'effetto di X sulla quota di ricadere in una categoria più alta invece che più bassa sia lo stesso per tutte le coppie di categorie adiacenti
- si può formulare il modello più complesso che assume variabilità degli effetti tra i logit

Continuation-Ratio model

Nel continuation-ratio model si considera il logit fra ciascuna categoria della variabile dipendente e le precedenti. Dato il valore delle variabili esplicative si ha:

$$r_j(x) = ln\left[\frac{P(Y = j|x)}{P(Y < j|x)}\right] = \beta_{j0} + \beta_{j1}x_1 + \ldots + \beta_{jp}x_p$$

Al variare della modalitá della variabile dipendente variano i parametri del logit, quindi il numero di parametri da stimare sará $p \times m$ e coincide con quello del modello multinomiale.
Questo modello é stimato mediante m regressioni logistiche binarie nelle quali si confronta la j-esima categoria della variabile dipendente rispetto alle precedenti. Il valore della log-verosimiglianza é dato dalla somma dei valori ottenuti nelle singole regressioni.

Modello di Bradley-Terry

- Modello per dati che consistono in confronti a coppie
- Il modello stimato fornisce una classifica

	Perdente				
Vincente	Seles	Graf	Sabatini	Navratilova	Sanchez
Seles	-	2	1	3	2
Graf	3	-	6	3	7
Sabatini	0	4	-	1	3
Navratilova	3	0	2	-	3
Sanchez	0	1	2	1	-

Modello di Bradley-Terry

- π_{ij} è la probabilità che il giocatore i vinca contro il giocatore j
- $\pi_{ji} = 1 - \pi_{ij}$, non possono esserci pareggi
- Consideriamo il modello

$$logit(\pi_{ij}) = \log \frac{\pi_{ij}}{\pi_{ji}} = \beta_i - \beta_j, \;\; i, j = 1, 2, \ldots, I - 1, \;\; i \neq j$$

- Ai fini della stima del modello, ciascuna coppia (y_{ij}, y_{ji}) può considerarsi come il risultato di una prova binomiale con y_{ij} successi su $(y_{ij} + y_{ji})$ prove

Modello di Bradley-Terry

- Il modello prevede un parametro per ciascun giocatore e nessun termine d'intercetta. Per permettere la stima del modello un coefficiente viene posto pari a zero oppure si impone il vincolo $\sum_h e^{\beta_h} = 1$
- Per la coppia di giocatori (i, j) le esplicative sono $(1, -1)$, zero per gli altri.
- Il modello descrive $\begin{pmatrix} N \\ 2 \end{pmatrix}$ probabilità mediante $(N-1)$ parametri, dove N è il numero di giocatori, per cui i gradi di libertà associati alla stima del modello sono $\begin{pmatrix} N \\ 2 \end{pmatrix} - (N-1)$

Modello di Bradley-Terry

- Il modello stimato

Giocatore	Stima	err. st.
Graf	1.81	0.66
Seles	1.49	0.78
Navratilova	1.07	0.72
Sabatini	0.86	0.67
Sanchez	0	-

- Le stime dei coefficienti forniscono una graduatoria tra le giocatrici
- Le probabilità attese

$$\hat{\pi}_{ij} = \frac{\exp(\hat{\beta}_i - \hat{\beta}_j)}{1 + \exp(\hat{\beta}_i - \hat{\beta}_j)}$$

- Un intervallo di confidenza per $(\beta_i - \beta_j)$ si traduce in un intervallo di confidenza per π_{ij} applicando la funzione di distribuzione logistica ai suoi estremi

Modello di Bradley-Terry

- Il modello precedente può essere generalizzato per tener conto nei confronti a coppie di effetti ulteriori
- Ad esempio se i dati si riferiscono a gare di calcio si può considerare l'effetto prodotto dal giocare in casa
- Consideriamo il modello

$$logit(\pi_{ij}) = \log \frac{\pi_{ij}}{\pi_{ji}} = \delta + \beta_i - \beta_j, \;\; i, j = 1, 2, \ldots, N-1, \;\; i \neq j$$

- π_{ij} è la probabilità che la squadra i batta la squadra j quando i gioca in casa
- $\delta > 0$ denota un vantaggio per la squadra che gioca in casa. La probabilità di vittoria della squadra che gioca in casa tra due di pari forza è $\frac{e^{\delta}}{1+e^{\delta}}$

Estensione del Modello di Bradley-Terry

- Il modello precedente può essere generalizzato al caso di confronti su una scala ordinale
- Ad esempio quando si esprime un giudizio nel confronto tra due prodotti su una scala con C livelli
- Consideriamo il modello per i logit cumulati, $c = 1, 2 \ldots, C-1$

$$logit\,[Prob(Y_{ij} \leq c)] = \delta_c + \beta_i - \beta_j,\ i, j = 1, 2, \ldots, N-1,\ i \neq j$$

- Osserviamo che

$$Prob(Y_{ij} \leq c) = Prob(Y_{ji} > C - c) = 1 - Prob(Y_{ji} \leq C - c)$$

da cui

$$logit\,[Prob(Y_{ij} \leq c)] = -logit\,[Prob(Y_{ji} \leq C - c)]$$

che implica $\delta_c = -\delta_{C-c}$

Estensione del Modello di Bradley-Terry

- Se nell'analisi di risultati sportivi si vuole includere la possibilità di pareggiare allora si può considerare una scala con 3 livelli (vittoria, pareggio, sconfitta) e il modello

$$logit\left[Prob(Y_{ij} \leq c)\right] = \delta_c + \beta_i - \beta_j$$

con $\delta_{c_1} = -\delta_{c_2}$

- Il modello può essere arricchito con l'effetto casalingo

$$logit\left[Prob(Y_{ij} \leq c)\right] = \delta + \delta_c + \beta_i - \beta_j$$