Ordinal Categorical Data Analysis

14 ottobre 2024

Indice

1 Chapter 2: Ordinal probabilities, scores and odds ratios

1.1 Notazione

Per una variabile ordinale Y:

- c siano il numero di categorie
- $n_1, \ldots, n_c$ il numero di osservazioni/pazienti di ciascuna categoria (frequenze assolute)
- $n = n_1 + \ldots + n_c$ il numero di pazienti
- $p_j = n_j/n$, per $j = 1, \ldots, c$ la frequenza relativa della categoria j

- π_j sia la probabilità che una osservazione selezionata casualmente dalla popolazione abbia la categoria j con $j = 1, \ldots, c$
- la probabilità cumulata sino alla categoria j, è

$$F_j = \mathbb{P}(Y \leq j) = \pi_1 + \ldots + \pi_j, \qquad j = 1, \ldots, c$$

- la probabilità cumulata stimata dal campione è

$$\widehat{F}_j = p_1 + \ldots + \pi_j, \qquad j = 1, \ldots, c$$

1.2 Sintesi di una variabile ordinale

Vari approcci:

- calcolare la mediana
- assegnare score numerici a ciascuna delle j categorie e poi fare una media; problema è che l'assegnazione degli score è discrezionale
- calcolare il *riddit* come score da assegnare alla categoria j, che corrisponde alla proporzione cumulata "media" sino alla categoria j, così definita

$$a_j = \sum_{k=1}^{j-1}, \qquad j = 1, \ldots, c$$

- *midranks*: calcolare il rango medio della categoria nell'ipotesi che non vi siano ties. Ad esempio il midrank per la prima categoria (con n_1 osservazioni) è $(1 + n_1)/2$. In generale il midrank per la categoria j è:

$$r_j = \frac{\left[\left(\sum_{i=1}^{j-1} n_i\right) + 1\right] + \sum_{i=1}^{j} n_i}{2}$$

- assumere che di fatto la variabile ordinale Y sia generata da una variabile latente quantitativa che per meccanismi di misurazione è discretizzata. Ad esempio se ipotizziamo che la variabile sottostante sia normale si può ottenere uno score numerico da associare a ciascuna categoria trovando il quantile del riddit

$$v_j = \Phi^{-1}(a_j)$$

Example 1.1 (Calcolo di riddits etc (es 2.1 pg 12))**.** Nel GSS una domanda ha ricevuto 1546 risposte "Decisamente" (categoria più bassa), 498 "Probabilmente" , 205 "Probabilmente no", 138 "Decisamente no" (categoria più alta), per un totale di 2387 risposte.

I riddit sono:

$$
\begin{aligned}
a_1 &= \frac{1}{2}\frac{1546}{2387} = 0.32 \\
a_2 &= \frac{1546}{2387} + \frac{1}{2}\frac{498}{2387} = 0.75 \\
a_3 &= \frac{1546}{2387} + \frac{498}{2387} + \frac{1}{2}\frac{205}{2387} = 0.90 \\
a_4 &= \frac{1546}{2387} + \frac{498}{2387} + \frac{205}{2387} + \frac{1}{2}\frac{138}{2387} = 0.97
\end{aligned}
$$

I normal score basati su riddit sono

```
qnorm(c(0.32, 0.75, 0.9, 0.95))

## [1] -0.4676988  0.6744898  1.2815516  1.6448536
```

1.3 Tabelle a due vie

Sia X un'altra variabile categorica (con $i = 1, \ldots, r$ categorie); la incrociamo con Y a $j = 1, \ldots, c$ categorie in una tabella di frequenza dove:

- n_{ij} è la frequenza della cella con $X = i, Y = j$
- $n = \sum_i \sum_j n_{ij}$ è il sample size
- p_{ij} è la proporzione della cella con $X = i, Y = j$
- p_{i+} è la probabilità marginale per la riga i (somma sulle colonne delle p della riga i)
- p_{+j} è la probabilità marginale per la colonna j
- $p_{j|i} = n_{ij}/n_{i+}$ è la probabilità condizionata di avere categoria $Y = j$ dato che si ha $X = i$; ovviamente $\sum_j p_{j|i} = 1$ e i valori $(p_{1|i}, \ldots, p_{c|i})$ formano la distribuzione condizionata
- $\widehat{F}_{j|i} = p_{1|i} + \ldots + p_{c|i}$ è la distribuzione condizionata cumulata

1.3.1 Misure di associazione se X ha due livelli

Nel caso speciale di X a due livelli (es 1=Trattati e 2=Controlli), consideriamo il caso della tabella $2 \times c$. Alcune misure di associazione sono legate alla dominanza stocastica, ossia alla probabilità che una osservazione presa a caso dai Trattati sia maggiore ad un'altra presa a caso dai controlli, indicato con $P(Y_1 > Y_2)$ dove indichiamo con Y_1 la probabilità di Y condizionata ad essere Trattati e con Y_2 quella ad esser controlli

Definition 1.1 (Misura di superiorità stocastica 1)**.**

$$\alpha = \mathbb{P}(Y_1 > Y_2) + \frac{1}{2}\mathbb{P}(Y_1 = Y_2) \tag{1}$$

Remark 1. La versione campionaria di α è

$$\widetilde{\alpha} = \sum_j \sum_{k<j} p_{j|1} p_{k|2} + \frac{1}{2}\sum_j p_{j|1} p_{j|2}$$

dove ad esempio:

- il primo addendo è $\mathbb{P}()$
- al secondo si ha la probabilità che i due gruppi, ottenuta sommando la probabilità che entrambi abbiano 1, entrambi 2, ..., entrambi c.

Remark 2. α ha un range in $[0, 1]$ e per interpretare:

- se $\alpha > 0.5$ l'esito nei trattati Y_1 tende ad essere più largo di quello dei controlli (e viceversa
- Se Y_1 e Y_2 sono distribuite identicamente, o hanno una distribuzione simmetrica su tutte le c categorie, allora $\alpha = 0.5$.

Definition 1.2 (Misura di superiorità stocastica 2)**.** Definita come

$$\Delta = \mathbb{P}(Y_1 > Y_2) - \mathbb{P}(Y_2 > Y_1) \tag{2}$$

Remark 3. Questa invece ha range in $[-1, 1]$ ed è centrata in 0 quando non vi è una dominanza schiacciante

Definition 1.3 (Odds ratio ordinale)**.** Definito come

$$\theta = \frac{\mathbb{P}(Y_1 > Y_2)}{\mathbb{P}(Y_2 > Y_1)} \tag{3}$$

Remark 4. Il suo valore campionario è

$$\widehat{\theta} = \frac{\sum_k \sum_{j>k} p_{j|1} p_{k|2}}{\sum_k \sum_{j<k} p_{j|1} p_{k|2}} = \frac{\sum_k \sum_{j>k} n_{1j} n_{2k}}{\sum_k \sum_{j<k} n_{1j} n_{2k}}$$

dove al numeratore abbiamo la probabilità che i trattati abbiano un valore maggiore dei controlli, ottenuta sommando la probabilità che i trattati abbiano l'ultima categoria e i controlli una delle precedenti, penultima e una delle precedenti, etc.

NB: Questo è molto interessante ma non viene dato un intervallo di confidenza; si va di bootstrap alla peggio.

Example 1.2 (Esempio trial pag 17 agresti)**.** I dati bivariati di tabella 1 Si ha che l'odds ratio ordinale (utilizzando la variante coi conteggi) è

$$\widehat{\theta} = \frac{12(11+8+8) + 10(11+8) + 4(11)}{5(6+4+10) + 8(6+4) + 8(6)} = 2.45$$

Essendo > 1 si vede che il gruppo dei trattati domina stocasticamente i controlli nel senso che tende ad avere punteggi più alti

	peggiorata	guarita $< 2/3$	guarita $\geq 2/3$	guarita	tot
Trattati	6	4	10	12	32
Controlli	11	8	8	5	32
Tot	17	12	18	17	64

Tabella 1: Tab 2.2 pag 17 agresti OrdCDA

1.3.2 Odds ratio

Remark 5. Nella sezione precedente abbiamo visto un odds ratio unico per sintetizzare una intera tabella.
In questa vediamo altre definizioni di odds ratio che a differenza, producono molteplici odds ratio per una singola tabella.

Remark 6 (Ripasso). In una tabella 2×2 un odds ratio è

$$\widehat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \frac{a \cdot d}{b \cdot c}$$

Remark 7. Nel caso di tabelle $r \times c$:

- odds ratio possono essere creati a partire da 4 celle a piacere
- tutti gli odds ratio possibili sono determinati da una base di $(r-1)\cdot(c-1)$ odds ratio
- questa base può essere costruita in vari modi, vediamo i più importanti in seguito

Definition 1.4 (Odds ratio con ultima cella)**.** Per generare la croce di celle considera l'ultima e le altre determinabili sulla base di righe e colonne in comune

$$\widehat{\theta}_{ij} = \frac{n_{ij} \cdot n_{rc}}{n_{rj} \cdot n_{ic}}, \qquad i = 1, \ldots, r-1, \quad j = 1, \ldots, c-1 \tag{4}$$

Remark 8. Stessa cosa si può fare con la prima di base ovviamente

Definition 1.5 (Odds ratio locali)**.** Considera le celle appena a destra e in basso della i, j considerata

$$\widehat{\theta}^L_{ij} = \frac{n_{ij} \cdot n_{i+1,j+1}}{n_{i,j+1} \cdot n_{i+1,j}}, \qquad i = 1, \ldots, r-1, \quad j = 1, \ldots, c-1 \tag{5}$$

Definition 1.6 (Odds ratio globali)**.** Considera la somma delle celle a sx e in alto fino alla i, j inclusa come cella a e determina le altre sommando la tabella nei settori identificati similarmente:

$$\widehat{\theta}^G_{ij} = \frac{\left(\sum_{a \leq i} \sum_{b \leq j} n_{ab}\right) \cdot \left(\sum_{a > i} \sum_{b > j} n_{ab}\right)}{\left(\sum_{a \leq i} \sum_{b > j} n_{ab}\right) \cdot \left(\sum_{a > i} \sum_{b \leq j} n_{ab}\right)}, \qquad i = 1, \ldots, r-1, \quad j = 1, \ldots, c-1 \tag{6}$$

Remark 9. I precedenti due trattano variabili di riga e colonna simmetreicamente e sono utili se entrambe le variabili sono di risposta.
Una famiglia di odds ratio che distingue tra righe e colonne, utile quando le X in riga è una variabile esplicativa e in colonna abbiamo l'outcome

Definition 1.7 (Cumulative odds ratio)**.** Come cella a considera la somma delle celle a sx di i, j e determina le rimanenti 4 similmente.

$$\widehat{\theta}_{ij}^{C} = \frac{\left(\sum_{b\leq j} n_{ib}\right)\cdot\left(\sum_{b>j} n_{i+1,b}\right)}{\left(\sum_{b>j} n_{ib}\right)\cdot\left(\sum_{b\leq j} n_{i+1,b}\right)}, \qquad i=1,\ldots,r-1, \quad j=1,\ldots,c-1 \tag{7}$$

Remark 10. Per tabelle $2\times c$ odds ratio globali e cumulativi sono identici

Important remark 1. Gli odds ratio campionari stimati come in precedenza su un dato campione (con una appropriata randomizzazione nel campionamento) forniscono uno stimatore delle relative grandezze nella popolazione, che riassumiamo di seguito.
Sia $\pi_{i,j}$ la probabilità di appartenere alla cella i, j della popolazione, mentre $p_{j|i}$ la probabilità condizionata di avere j in Y nell'ipotesi che in X si abbia i.
Si ha

$$\theta_{ij}^{L} = \frac{\pi_{ij}\cdot\pi_{i+1,j+1}}{\pi_{i,j+1}\cdot\pi_{i+1,j}} = \frac{\pi_{j|i}/\pi_{j+1|i}}{\pi_{j|i+1}/\pi_{j+1|i+1}} = \text{finire con } \mathbb{P}\,(), \text{ pag } 23 \tag{8}$$

$$\theta_{ij}^{C} = \frac{\left(\sum_{b\leq j} \pi_{ib}\right)\cdot\left(\sum_{b>j} \pi_{i+1,b}\right)}{\left(\sum_{b>j} \pi_{ib}\right)\cdot\left(\sum_{b\leq j} \pi_{i+1,b}\right)} = \text{finire con } \mathbb{P}\,(), \text{ pag } 23 \tag{9}$$

$$\theta_{ij}^{G} = \frac{\left(\sum_{a\leq i}\sum_{b\leq j} \pi_{ab}\right)\cdot\left(\sum_{a>i}\sum_{b>j} \pi_{ab}\right)}{\left(\sum_{a\leq i}\sum_{b>j} \pi_{ab}\right)\cdot\left(\sum_{a>i}\sum_{b\leq j} \pi_{ab}\right)} = \frac{\mathbb{P}\,(X\leq i, Y\leq j)\cdot\mathbb{P}\,(X>i, Y>j)}{\mathbb{P}\,(X\leq i, Y>j)\cdot\mathbb{P}\,(X>i, Y\leq j)} \tag{10}$$

2 Chapter 3: Cumulative logits model

Per c categorie di outcome con probabilità $\pi_1,\ldots,\pi_c$ i logits cumulati sono definiti come

$$\begin{aligned} logit(\mathbb{P}\,(Y\leq j)) &= \log odd\,\mathbb{P}\,(Y\leq j) = \log\frac{\mathbb{P}\,(Y\leq j)}{1-\mathbb{P}\,(Y\leq j)} \\ &= \log\frac{\pi_1+\ldots+\pi_j}{\pi_{j+1}+\ldots+\pi_c} \qquad j=1,\ldots,c-1 \end{aligned}$$

Questa probabilità è quella che si va a modellare con un modello CLM. La costruzione del modello e l'interpretazione dei coefficienti dipende dal software (R, Stata, SPSS si comportano nella stessa maniera, SAS va per i conti suoi).
Il modello CLM è il seguente

$$logit(\mathbb{P}\,(Y\leq j)) = \alpha_j - \widehat{\beta}\mathbf{x}_i$$

alcune note:

- la parametrizzazione con $-$ davanti al beta serve per avere che incrementi delle variabili indipendenti portino una diminuzione della probabilità di ricadere nelle classi inferiori dell'outcome (ossia un beta positivo viene interpretato come aumenta la probabilità di avere un outcome ordinato maggiore). Così si comporta R e tanti altri, mentre SAS pone $+$ davanti ai β.
- nel modello il logit per la probabilità cumulata sino alla classe j ha la propria intercetta, mentre i beta sono unici/comuni (modello succinto). Le a_j aumentano con j poiché anche non considerando le covariate $\mathbb{P}(Y \leq j)$ aumenta se aumenta j
- per ottenere la probabilità cumulata sino alla classe j si ha

$$\mathbb{P}(Y \leq j) = \frac{\exp\left(\alpha_j - \widehat{\beta}\mathbf{x}_i\right)}{1 + \exp\left(\alpha_j - \widehat{\beta}\mathbf{x}_i\right)} \tag{11}$$

- per la probabilità della cella j-esima si ha

$$\mathbb{P}(Y = j) = \mathbb{P}(Y \leq j) - \mathbb{P}(Y \leq j - 1) \tag{12}$$

$$= \frac{\exp\left(\alpha_j - \widehat{\beta}\mathbf{x}_i\right)}{1 + \exp\left(\alpha_j - \widehat{\beta}\mathbf{x}_i\right)} - \frac{\exp\left(\alpha_{j-1} - \widehat{\beta}\mathbf{x}_i\right)}{1 + \exp\left(\alpha_{j-1} - \widehat{\beta}\mathbf{x}_i\right)} \tag{13}$$

2.1 Esempi

2.1.1 DAE UCLA

Fonte qui. Uno studio studia i fattori che influenzano la decisioni di iscriversi all'università:

- outcome: `apply`: `unlikely` $<$ `somewhat likely` $<$ `very likely`
- covariate:
  - `pared`: dummy genitore studiato
  - `public`: dummy proviene da scuola pubblica
  - `gpa`: indicatore di media scolastica

```
# dae_ologit <- foreign::read.dta("https://stats.idre.ucla.edu/stat/data/ologit.dta")
dat <- lbdatasets::dae_ologit
head(dat)
```

```
##             apply pared public  gpa
## 1     very likely     0      0 3.26
## 2 somewhat likely     1      0 3.21
## 3        unlikely     1      1 3.94
## 4 somewhat likely     0      0 2.81
## 5 somewhat likely     0      0 2.53
## 6        unlikely     0      1 2.59
```

```
m <- ordinal::clm(apply ~ pared + public + gpa, data = dat)
summary(m)
```

```
## formula: apply ~ pared + public + gpa
## data:    dat
##
##  link  threshold nobs logLik  AIC    niter max.grad cond.H
##  logit flexible  400  -358.51 727.02 5(0)  1.63e-10 1.3e+03
##
## Coefficients:
##        Estimate Std. Error z value Pr(>|z|)
## pared   1.04766    0.26579   3.942 8.09e-05 ***
## public -0.05868    0.29786  -0.197   0.8438
## gpa     0.61575    0.26063   2.363   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                             Estimate Std. Error z value
## unlikely|somewhat likely      2.2033     0.7795   2.826
## somewhat likely|very likely   4.2988     0.8043   5.345
```

Dalla stima si desume che:

- vi è una relazione positiva tra l'educazione parentale (`pared`) e la media scolastica (`gpa`) e l'ottenere score più alti in probabilita di fare domanda per l'università (`apply`)
- relazione negativa ma largamente non significativa per provenire da una scuola pubblica
- il modello può essere scritto per esteso come

$$logit(\hat{P}(Y \leq 1)) = 2.20 - 1.05 \cdot PARED + 0.06 \cdot PUBLIC - 0.616 \cdot GPA$$
$$logit(\hat{P}(Y \leq 2)) = 4.30 - 1.05 \cdot PARED + 0.06 \cdot PUBLIC - 0.616 \cdot GPA$$

- per pared passare da avere genitori non studiati a genitori studiati diminuisce di 1.05 il log odd di applicarsi in classi inferiori (aumenta la probabilità di applicarsi), a parità di altro, mentre un punto di GPA di 0.616. Per avere gli odds ratio si esponenzia (lo facciamo solo per i beta):

```
exp(cbind(OR = coef(m)[-(1:2)], confint(m)))

##               OR     2.5 %   97.5 %
## pared  2.8509825 1.6958390 4.817088
## public 0.9430059 0.5209033 1.680583
## gpa    1.8510366 1.1136241 3.098472
```

For students whose parents did attend college, the odds of being more likely (i.e., very or somewhat likely versus unlikely) to apply is 2.85 times that of students whose parents did not go to college, holding constant all other variables.
For every one unit increase in student's GPA the odds of being more likely to apply (very or somewhat likely versus unlikely) is multiplied 1.85 times (i.e., increases 85%), holding constant all other variables.

- può essere più comodo presentare (plottandole) le probabilità predette

```
newdat <- expand.grid(pared = 0:1, public = 0:1, gpa = 2:4)
preds <- predict(object = m,
                 newdata = newdat,
                 type = "prob"#,
                 ## interval = TRUE , questo da gli intervalli delle predizioni ma
                 ## fa casino
                 #
                 )$fit # questo serve per pulire nomi, solo estetica
res <- cbind(newdat, preds)
(res)

##    pared public gpa  unlikely somewhat likely very likely
## 1      0      0   2 0.7254844       0.2300381  0.04447746
## 2      1      0   2 0.4810511       0.4017898  0.11715911
## 3      0      1   2 0.7370156       0.2209353  0.04204910
## 4      1      1   2 0.4957128       0.3930628  0.11122441
## 5      0      0   3 0.5880926       0.3325807  0.07932674
## 6      1      0   3 0.3336822       0.4691145  0.19720331
## 7      0      1   3 0.6022307       0.3226239  0.07514532
## 8      1      1   3 0.3468544       0.4650679  0.18807776
## 9      0      0   4 0.4354473       0.4270021  0.13755066
## 10     1      0   4 0.2129350       0.4744926  0.31257239
## 11     0      1   4 0.4499241       0.4193399  0.13073599
## 12     1      1   4 0.2229355       0.4769606  0.30010388
```

- per testare l'ipotesi di odds proporzionali (il poter usare gli stessi beta) si usa il test di Brant (da approfondire), sperando che sia non significativo. Qui è andata bene e non violiamo l'assunto di proportional odds. Altro test è quello fornito dal package `ordinal` che in pratica a turno porta la covariata considerata in non proportional odd e poi confronta i due modelli (in questo caso non funziona nn so perché)

```
gofcat::brant.test(m)

##
## Brant Test:
##            chi-sq   df   pr(>chi)
## Omnibus     4.343    3      0.227
## pared       0.132    1      0.716
## public      3.443    1      0.064 .
## gpa         0.179    1      0.672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## H0: Proportional odds assumption holds

ordinal::nominal_test(m)

## Tests of nominal effects
##
## formula: apply ~ pared + public + gpa
##        Df  logLik    AIC LRT Pr(>Chi)
## <none>    -358.51 727.02
## pared
## public
## gpa
```

2.1.2 Esempio partial prop odd

Utilizziamo l'esempio di wine preso dalla vignetta del pacchetto `ordinal`:

- outcome: rating grado di secchezza (da 1 a 5 del vino)
- covariate: temperature (cold/warm), contact (no/yes)

Dalla seguente stima si nota come la temperatura warm aumenti il grado di secchezza cosi come il contact

```
library(ordinal)
with(wine, table(interaction(temp, contact), rating))[c(1,3,2,4),]

##           rating
##            1 2 3 4 5
##   cold.no  4 9 5 0 0
##   cold.yes 1 7 8 2 0
##   warm.no  0 5 8 3 2
##   warm.yes 0 1 5 7 5

# test prop odds
m <- ordinal::clm(rating ~ temp + contact, data = ordinal::wine)
summary(m)
```

```
## formula: rating ~ temp + contact
## data:    ordinal::wine
##
##  link  threshold nobs logLik AIC    niter max.grad cond.H
##  logit flexible  72   -86.49 184.98 6(0)  4.02e-12 2.7e+01
##
## Coefficients:
##            Estimate Std. Error z value Pr(>|z|)
## tempwarm     2.5031     0.5287   4.735 2.19e-06 ***
## contactyes   1.5278     0.4766   3.205  0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 1|2  -1.3444     0.5171  -2.600
## 2|3   1.2508     0.4379   2.857
## 3|4   3.4669     0.5978   5.800
## 4|5   5.0064     0.7309   6.850

gofcat::brant.test(m)

## Warning:  with 5 zero cell entries in the computation crosstab, test results
may be inaccurate.

##
## Brant Test:
##               chi-sq   df   pr(>chi)
## Omnibus        1.524    6       0.96
## tempwarm       0.934    3       0.82
## contactyes     0.857    3       0.84
##
## H0: Proportional odds assumption holds

ordinal::nominal_test(m)

## Tests of nominal effects
##
## formula: rating ~ temp + contact
##         Df  logLik    AIC    LRT Pr(>Chi)
## <none>     -86.492 184.98
## temp     3 -84.904 187.81 3.1750   0.3654
## contact  3 -86.209 190.42 0.5667   0.9040
```

Qui non ce n'era un bisogno clamoroso ma lui procede per fini esemplificativi. Ipotizziamo che `contact` violi proportional odds; quello che fa la seguente stima è creare un β per ciascuno strato specifico, ossia

$$logit(\mathbb{P}(Y_i \leq j) = \theta_j + \beta_j contact - \beta temp$$

quindi non richiedendo più che il beta sia comune. Occhio che il segno cambia

```
fm.nom <- clm(rating ~ temp, nominal = ~ contact, data = wine)
summary(fm.nom)

## formula: rating ~ temp
## nominal: ~contact
## data:    wine
##
##  link  threshold nobs logLik AIC    niter max.grad cond.H
##  logit flexible  72   -86.21 190.42 6(0)  1.64e-10 4.8e+01
##
## Coefficients:
##          Estimate Std. Error z value Pr(>|z|)
## tempwarm    2.519      0.535   4.708  2.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                 Estimate Std. Error z value
## 1|2.(Intercept)  -1.3230     0.5623  -2.353
## 2|3.(Intercept)   1.2464     0.4748   2.625
## 3|4.(Intercept)   3.5500     0.6560   5.411
## 4|5.(Intercept)   4.6602     0.8604   5.416
## 1|2.contactyes   -1.6151     1.1618  -1.390
## 2|3.contactyes   -1.5116     0.5906  -2.559
## 3|4.contactyes   -1.6748     0.6488  -2.581
## 4|5.contactyes   -1.0506     0.8965  -1.172
```

Quindi:

- il beta comune di contact sparisce rispetto alla stima precedente, mentre si aggiunge un secondo set di coefficienti per cui il β di contact è -1.62 se $j = 1$, sino a -1.051 se $j = 4$; questi quattro si confrontano con il β di 1.52 (di segno opposto) della stima precedente
- i nuovi coefficienti stimati possono essere pensati come un modificatore delle intercette gia presenti, essendo strato specifici. Al fine di semplificare il calcolo delle intercette originali + beta strato specifici viene messa a disposizione

```
fm.nom$Theta

##   contact       1|2        2|3      3|4      4|5
## 1      no -1.323043  1.2464435 3.550044 4.660247
## 2     yes -2.938103 -0.2651238 1.875288 3.609624
```

- filosoficamente assomiglia molto alla stratificazione nel modello di cox dove una variabile che non rispetta prop hazard viene utilizzata non più per il beta ma nel basale

2.1.3 Esempio 3.2.5 pag 51 agresti ordcda

```
# creando la tabella
m <- matrix(c(23, 84, 98,
              50, 286, 574,
              4, 44, 122,
              11, 57, 268,
              1, 23, 148),
            ncol = 3, byrow = TRUE)
rownames(m) <- c("<high school", "high school", "junior college", "bachelor", "graduate")
colnames(m) <- c("very", "sort of", "not at all")
df <- lbmisc::table2df(as.table(m))
names(df) <- c("degree", "astrology_scientific")
table(df)

##                 astrology_scientific
## degree           very sort of not at all
##   <high school     23      84         98
##   high school      50     286        574
##   junior college    4      44        122
##   bachelor         11      57        268
##   graduate          1      23        148

df$degree_n <- as.integer(df$degree) - 1

## stima di table 3.2 pag 52 (parte alta tabella)
m1 <- ordinal::clm(astrology_scientific ~ degree_n, data = df)
summary(m1)

## formula: astrology_scientific ~ degree_n
## data:    df
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1793 -1329.08 2664.16 6(0)  9.61e-12 2.8e+01
##
## Coefficients:
##          Estimate Std. Error z value Pr(>|z|)
## degree_n  0.46137    0.04861   9.492   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                    Estimate Std. Error z value
## very|sort of       -2.31378    0.12461  -18.57
## sort of|not at all -0.02642    0.08535   -0.31

## questa è identica, cambia solo il segno del beta stimato, le
## intercette sono uguali

## stima di table 3.2 pag 52 (parte alta tabella)
```

```r
m2 <- ordinal::clm(astrology_scientific ~ degree, data = df)
summary(m2)
```

```
## formula: astrology_scientific ~ degree
## data:    df
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1793 -1328.20 2668.41 6(0)  5.04e-12 7.1e+01
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## degreehigh school       0.6497     0.1511   4.300 1.71e-05 ***
## degreejunior college    1.0657     0.2164   4.924 8.46e-07 ***
## degreebachelor          1.4746     0.1917   7.692 1.45e-14 ***
## degreegraduate          1.9439     0.2582   7.529 5.11e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                     Estimate Std. Error z value
## very|sort of         -2.1797     0.1603  -13.60
## sort of|not at all    0.1129     0.1360    0.83
```

```r
## questa è differente ma per i differenti contrasti tra R e SAS
```

```r
## # creando la tabella
## m <- matrix(c(350, 307, 345, 481, 67,
##               334, 99, 117, 159, 30),
##             nrow = 2, byrow = TRUE)
## rownames(m) <- c("smoker", "non_smoker")
## colnames(m) <- 0:4
## df <- lbmisc::table2df(as.table(m))
## names(df) <- c("smoke", "chd")
## table(df)

## # test prop odds
## m <- ordinal::clm(chd ~ smoke, data = df)
## gofcat::brant.test(m)

## m2 <- ordinal::clm(chd ~ 1, nominal = ~ smoke, data = df)
## summary(m2)
```

3 Cap 10: modelli mixed

3.1 Esempi

3.1.1 Esempio 10.2.2 pag 290 (arthritis clinical trial)

```r
db <- lbdatasets::arthritis
head(db)
```

```
##   id sex ara age trt baseline extra y1 y2 y3
## 1  1   2   3  54   2        2     3  3  3  3
## 2  2   1   3  41   1        3     4  3  3  3
## 3  3   2   2  48   2        3     3  2  3  3
## 4  4   2   2  40   1        3     4  3  2  3
## 5  5   2   1  29   2        3     4  3  2  3
## 6  6   2   3  43   2        2     3  2  1  2
```

```r
## mettiamo il db in formato long perché l'outcome è in formato wide
names(db)[8:10] <- c("y.1", "y.2", "y.3")
db_long <- reshape(data = db,
                   varying = list(c('y.1', "y.2", "y.3")),
                   idvar = "id",
                   direction = "long")
db_long <- db_long[with(db_long, order(id, time)), ]
names(db_long)[ncol(db_long)] <- 'y'
rownames(db_long) <- NULL
vars <- c("id", "time", "y", "sex", "ara", "age", "trt", "baseline", "extra")
db_long <- db_long[, vars]

## recoding vario per stimare il modello di pg 290
db_long$y <- factor(db_long$y, ordered = TRUE)
db_long$time <- factor(db_long$time)
db_long$s <- db_long$sex -1 # proviamo, le codifiche non coincidono col libro
db_long$tr <- db_long$trt -1 # proviamo..
db_long$a <- db_long$age      # renaming
db_long$b <- db_long$baseline # renaming
db_long$id <- factor(db_long$id)

## stima con tutto dentro
library(ordinal)
m <- ordinal::clmm(y ~ time + b + a + s + tr + (1|id), data = db_long)
summary(m)
```

```
## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: y ~ time + b + a + s + tr + (1 | id)
## data:    db_long
##
##  link  threshold nobs logLik  AIC     niter     max.grad cond.H
##  logit flexible  888  -801.56 1621.13 412(2510) 2.54e-03 2.2e+05
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 3.266    1.807
## Number of groups:  id 301
##
```

```
## Coefficients:
##        Estimate Std. Error z value Pr(>|z|)
## time2 -0.145516   0.181876  -0.800  0.42366
## time3  0.332729   0.183763   1.811  0.07020 .
## b      1.316992   0.160283   8.217  < 2e-16 ***
## a     -0.009494   0.011822  -0.803  0.42192
## s      0.225673   0.293704   0.768  0.44227
## tr     0.837483   0.263932   3.173  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 1|2   1.7093     0.8251   2.072
## 2|3   4.6867     0.8515   5.504
## (18 osservazioni eliminate a causa di valori mancanti)

## bah ci assomiglia al libro a parte il coefficiente di b
## il sas di agresti è incomprensibile

## stima con poca roba (table 10.2 pag291) (lui toglie anche il tempo 2
## lo lasciamo per pigrizia)
m2 <- ordinal::clmm(y ~ time + b + tr + (1|id), data = db_long)
summary(m2)

## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: y ~ time + b + tr + (1 | id)
## data:    db_long
##
##  link  threshold nobs logLik  AIC     niter     max.grad cond.H
##  logit flexible  888  -802.18 1618.36 384(2470) 1.02e-04 3.1e+02
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 3.289    1.814
## Number of groups:  id 301
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
## time2  -0.1461     0.1819  -0.803  0.42180
## time3   0.3333     0.1838   1.813  0.06977 .
## b       1.3247     0.1601   8.276  < 2e-16 ***
## tr      0.8497     0.2643   3.215  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 1|2   2.0512     0.4841   4.237
```

```
## 2|3   5.0299     0.5304   9.484
## (18 osservazioni eliminate a causa di valori mancanti)

## un pochino differente ma il coefficiente del trattamento coincide
## perfettamente, ah i randomizzati <3
```

3.1.2 Esempio wine ordinal

Vedi vignetta