Generalized Estimating Equations part duex

Alexander McLain

Contents

1 Multinomial GEE example

A two-year study was conducted to assess the impact of access to Section 8 housing as a means of providing independent housing to the severely mentally ill homeless (Hurlbut, Wood, and Hough 1996). In this study, half of the 362 clients received Section 8 housing certificates. The assignment of Section 8 housing certificates is recorded in the variable Sec; 0 indicates clients who did not receive a certificate, and 1 indicates clients who received a certificate.

Every six months during the study, research staff interviewed all 362 clients, who provided data about their living arrangements in the previous 60 days. Clients' living arrangements were also recorded during a baseline interview. The time of interviews is recorded in the variable Time, whose value is 0, 6, 12, or 24 (for the number of months since the study began). There were a total of 159 missed interviews. The variable Housing records the living arrangement of a client and is coded as 0 (street living), 1 (community living), or 2 (independent living).

```
library(readxl)
house <- read_excel("Housing.xlsx", na = ".")
## First checked to see if "." was read as NA
str(house)
```

```
## tibble [1,448 x 4] (S3: tbl_df/tbl/data.frame)
##  $ ID     : num [1:1448] 1 1 1 1 2 2 2 2 3 3 ...
##  $ Housing: num [1:1448] 1 2 2 2 1 2 2 1 0 2 ...
##  $ Time   : num [1:1448] 0 6 12 24 0 6 12 24 0 6 ...
##  $ Sec    : num [1:1448] 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(house)
```

ID	Housing	Time	Sec
1	1	0	1
1	2	6	1
1	2	12	1
1	2	24	1
2	1	0	1
2	2	6	1

To fit a multinomial model we'll use the `multgee` package. This package contains two functions that fit GEE models for correlated multinomial responses; `ordLORgee` for an ordinal response scale and `nomLORgee` for a

nominal response scale.

The main arguments in both functions are:

1. an optional data frame (data),
2. a model formula (formula),
3. a cluster identifier variable (id) and
4. an optional vector that identifies the order of the observations within each cluster (repeated).

2 Different multinomial models

```
library(multgee)
```

```
## Loading required package: gnm
```

```
table( house$Housing, house$Time)
```

/	0	6	12	24
0	180	45	32	37
1	136	138	108	102
2	45	139	163	164

```
table( house$Housing, house$Sec)
```

/	0	1
0	161	133
1	305	179
2	166	345

```
table( house$Time, house$Sec, house$Housing)
```

Var1	Var2	Var3	Freq
0	0	0	100
		1	61
		2	19
	1	0	80
		1	75
		2	26
12	0	0	13
		1	85
		2	48
	1	0	19
		1	23
		2	115
24	0	0	18
		1	66
		2	61
	1	0	19
		1	36

Var1	Var2	Var3	Freq
		2	103
6	0	0	30
		1	93
		2	38
	1	0	15
		1	45
		2	101

A marginal baseline category logit model is offered in the function `nomLORgee`.

2.0.1 Baseline-Category Logit Model

- This model is the most popular form of $\pi_j(X)$ when the responses are nominal (i.e., there is no order).
- This model compares $r-1$ categories to a baseline or referent category, denoted by j^* (commonly $j^*=1$).
- We model $\pi_j(X) = \Pr(Y=j|X)$ as

$$\log\left\{\frac{\Pr(Y=j|X)}{\Pr(Y=j^*|X)}\right\} = \log\left\{\frac{\pi_j(X)}{\pi_{j^*}(X)}\right\} = X\beta_j = \beta_{0j} + \beta_{1j}X_1 + \ldots$$

  If X has length p, then this model has $(r-1)\times p$ free parameters.

- If the response is ordinal, usually the highest or the lowest category in ordinal scale is chosen.

```
form_gee <- Housing ~ factor(Sec) + factor(Time) + factor(Sec)*factor(Time)
nom_log_2 <- nomLORgee(formula = form_gee, data = house, id = ID, repeated = Time)

## Change all 0's to 3's (so that it will be the last category)
library(tidyverse)
house <- house %>% mutate( Housing_2 = ifelse( Housing == 0, 3, Housing))
table( house$Housing, house$Housing_2)
```

/	1	2	3
0	0	0	294
1	484	0	0
2	0	511	0

```
form_gee <- Housing_2 ~ factor(Sec) + factor(Time) + factor(Sec)*factor(Time)
nom_log_0 <- nomLORgee(formula = form_gee, data = house, id = ID, repeated = Time)

### Using 2 as the referent category
cbind( coefficients(nom_log_2), coefficients(nom_log_0) )
```

```
## Warning in kable_pipe(x = structure(c("beta10", "factor(Sec)1:1",
## "factor(Time)6:1", : The table should have a header (column names)
```

beta10	1.6607312	-0.4942963
factor(Sec)1:1	-0.5368011	0.4297578
factor(Time)6:1	-1.8701034	1.6155588
factor(Time)12:1	-2.9250570	2.3545358
factor(Time)24:1	-2.8135894	1.7725770

factor(Sec)1:factor(Time)6:1	-1.1821813	-0.4412346
factor(Sec)1:factor(Time)12:1	0.0791567	-2.1276618
factor(Sec)1:factor(Time)24:1	0.0327299	-1.0823813
beta20	1.1664349	-1.6607312
factor(Sec)1:2	-0.1070433	0.5368011
factor(Time)6:2	-0.2545448	1.8701033
factor(Time)12:2	-0.5705218	2.9250576
factor(Time)24:2	-1.0410113	2.8135884
factor(Sec)1:factor(Time)6:2	-1.6234164	1.1821822
factor(Sec)1:factor(Time)12:2	-2.0485043	-0.0791573
factor(Sec)1:factor(Time)24:2	-1.0496530	-0.0327280

```
### Using 0 as the referent category

summary(nom_log_0)
```

```
## GEE FOR NOMINAL MULTINOMIAL RESPONSES
## version 1.6.0 modified 2017-07-10
##
## Link : Baseline Category Logit
##
## Local Odds Ratios:
## Structure:         time.exch
## Model:             3way
## Homogenous scores: TRUE
##
## call:
## nomLORgee(formula = form_gee, data = house, id = ID, repeated = Time)
##
## Summary of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.719981 -0.338889 -0.143646  0.001088  0.414072  0.894444
##
## Number of Iterations: 3
##
## Coefficients:
##                               Estimate   san.se   san.z Pr(>|san.z|)
## beta10                       -0.49430  0.16246 -3.0426      0.00235 **
## factor(Sec)1:1                0.42976  0.22853  1.8805      0.06004 .
## factor(Time)6:1               1.61556  0.23756  6.8006      < 2e-16 ***
## factor(Time)12:1              2.35454  0.30856  7.6307      < 2e-16 ***
## factor(Time)24:1              1.77258  0.29071  6.0974      < 2e-16 ***
## factor(Sec)1:factor(Time)6:1  -0.44123  0.40993 -1.0764      0.28176
## factor(Sec)1:factor(Time)12:1 -2.12766  0.44362 -4.7961      < 2e-16 ***
## factor(Sec)1:factor(Time)24:1 -1.08238  0.41763 -2.5917      0.00955 **
## beta20                       -1.66073  0.25026 -6.6359      < 2e-16 ***
## factor(Sec)1:2                0.53680  0.33704  1.5927      0.11122
## factor(Time)6:2               1.87010  0.31876  5.8668      < 2e-16 ***
## factor(Time)12:2              2.92506  0.36829  7.9424      < 2e-16 ***
## factor(Time)24:2              2.81359  0.34258  8.2130      < 2e-16 ***
## factor(Sec)1:factor(Time)6:2   1.18218  0.46036  2.5680      0.01023 *
## factor(Sec)1:factor(Time)12:2 -0.07916  0.48306 -0.1639      0.86984
## factor(Sec)1:factor(Time)24:2 -0.03273  0.46558 -0.0703      0.94396
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Local Odds Ratios Estimates:
##       [,1]   [,2]  [,3]   [,4]  [,5]   [,6]  [,7]   [,8]
## [1,] 0.000  0.000 3.687  0.144 3.687  0.144 3.687  0.144
## [2,] 0.000  0.000 0.144 17.748 0.144 17.748 0.144 17.748
## [3,] 3.687  0.144 0.000  0.000 3.687  0.144 3.687  0.144
## [4,] 0.144 17.748 0.000  0.000 0.144 17.748 0.144 17.748
## [5,] 3.687  0.144 3.687  0.144 0.000  0.000 3.687  0.144
## [6,] 0.144 17.748 0.144 17.748 0.000  0.000 0.144 17.748
## [7,] 3.687  0.144 3.687  0.144 3.687  0.144 0.000  0.000
## [8,] 0.144 17.748 0.144 17.748 0.144 17.748 0.000  0.000
##
## pvalue of Null model: <0.0001
```

Options for the marginal model in the function `ordLORgee` include cumulative link models or an adjacent categories logit model.

2.0.2 The Proportional-Odds Cumulative Logit Model

- Proportional-odds cumulative logit model is possibly the most popular model for ordinal data (and is used quite frequently for survival data).
- Here we build the model using cumulative probabilities, i.e.,

$$\Pr(Y \leq j|X) = \pi_1(X) + \pi_2(X) + \ldots + \pi_j(X)$$

  the probability that the category is less than or equal to category j.

- Then a cumulative logit is defined as

$$\log\left\{\frac{\Pr(Y \leq j|X)}{\Pr(Y > j|X)}\right\} = \log\left\{\frac{\Pr(Y \leq j|X)}{1 - \Pr(Y \leq j|X)}\right\} = X\beta_j$$

  which we can write as $\text{logit}\{\Pr(Y \leq j|X)\} = X\beta_j$.

\end{frame}

```r
form_gee <- Housing ~ factor(Sec) + factor(Time)
cum_log <- ordLORgee(formula = form_gee, data = house, id = ID, repeated = Time)
summary(cum_log)
```

```
## GEE FOR ORDINAL MULTINOMIAL RESPONSES
## version 1.6.0 modified 2017-07-10
##
## Link : Cumulative logit
##
## Local Odds Ratios:
## Structure:         category.exch
## Model:             3way
##
## call:
## ordLORgee(formula = form_gee, data = house, id = ID, repeated = Time)
##
## Summary of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.612912 -0.311452 -0.153290  0.001443  0.540097  0.933739
```

```
##
## Number of Iterations: 3
##
## Coefficients:
##                Estimate   san.se    san.z Pr(>|san.z|)
## beta10          0.45957  0.12824   3.5837      0.00034 ***
## beta20          2.50317  0.15846  15.7965      < 2e-16 ***
## factor(Sec)1   -0.93657  0.14429  -6.4909      < 2e-16 ***
## factor(Time)6  -1.82612  0.14193 -12.8662      < 2e-16 ***
## factor(Time)12 -2.16587  0.15571 -13.9094      < 2e-16 ***
## factor(Time)24 -2.16860  0.15957 -13.5905      < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Local Odds Ratios Estimates:
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
## [1,] 0.000 0.000 1.692 1.692 1.384 1.384 1.332 1.332
## [2,] 0.000 0.000 1.692 1.692 1.384 1.384 1.332 1.332
## [3,] 1.692 1.692 0.000 0.000 4.408 4.408 2.466 2.466
## [4,] 1.692 1.692 0.000 0.000 4.408 4.408 2.466 2.466
## [5,] 1.384 1.384 4.408 4.408 0.000 0.000 6.834 6.834
## [6,] 1.384 1.384 4.408 4.408 0.000 0.000 6.834 6.834
## [7,] 1.332 1.332 2.466 2.466 6.834 6.834 0.000 0.000
## [8,] 1.332 1.332 2.466 2.466 6.834 6.834 0.000 0.000
##
## pvalue of Null model: <0.0001
```