

Marginal Model for Categorical Data: Case Studies

Crossover Trial

The data is available at the textbook's website:

<http://faculty.washington.edu/heagerty/Books/AnalysisLongitudinal/datasets.html>

```
> library (gee)
> xover3 <- read.table ("../data/xover3.data",
+                       col.names = c("id", "class", "relief",
+                                     "intercept", "tx2", "tx3", "p2", "p3",
+                                     "ptx1", "ptx2", "ptx3"))
> xover3$period <- ifelse (xover3$p2 == 1, 2,
+                          ifelse (xover3$p3 == 1, 3, 1))
> xover3$treatment <- ifelse (xover3$tx2 == 1, 2,
+                             ifelse (xover3$tx3 == 1, 3, 1))
> with (xover3, ftable (period, relief, treatment))
      treatment  1  2  3
period relief
1      0          24  4 10
      1           7 19 22
2      0          20 11  5
      1           9 18 23
3      0          20  7  5
      1           6 23 25
```

- Just from the table, it appears that the two different doses of the analgesic treatment (2 and 3) both did better than the placebo (1). The high dose may be slightly better than the low dose.
- Is there any carry-over effect?

```
> xover3$ptx <- ifelse (xover3$ptx1 == 1, 1,
+                      ifelse (xover3$ptx2 == 1, 2, 3))
> xover3$ptx[xover3$period == 1] <- 0
> with (xover3, ftable (ptx, relief, treatment))
      treatment  1  2  3
ptx relief
0      0          24  4 10
      1           7 19 22
1      0           0  6  5
      1           0 23 26
2      0          20  0  5
      1           9  0 22
3      0          20 12  0
      1           6 18  0
```

We fit two models, assuming working independence or unstructured correlation.

Model 1: Working independence correlation.

```
> xover.gee <- gee (relief ~ p2 + p3 + tx2 + tx3 + ptx2 + ptx3,
+                 data = xover3, scale.fix = TRUE, id = id,
+                 family = binomial)
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
[1] -1.0864970 0.4123752 0.5866249 2.1055755 2.0683345 -0.1284182
    -0.9285768
> summary (xover.gee)
```

```
GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)
```

Model:

```
Link:                               Logit
Variance to Mean Relation: Binomial
Correlation Structure:               Independent
```

```
Call: gee(formula = relief ~ p2 + p3 + tx2 + tx3 + ptx2 + ptx3, id = id,
          data = xover3, family = binomial, scale.fix = TRUE)
```

Summary of Residuals:

	Min	1Q	Median	3Q	Max
	-0.8328140	-0.2522785	0.1819871	0.2652069	0.8323952

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-1.0864970	0.3282391	-3.3100778	0.3180752	-3.4158489
p2	0.4123752	0.4608082	0.8948955	0.4118732	1.0012187
p3	0.5866249	0.4750698	1.2348183	0.4560502	1.2863166
tx2	2.1055755	0.4017464	5.2410563	0.4191703	5.0231983
tx3	2.0683345	0.3830991	5.3989548	0.4185131	4.9421017
ptx2	-0.1284182	0.5016688	-0.2559821	0.5022616	-0.2556799
ptx3	-0.9285768	0.4872822	-1.9056242	0.4457959	-2.0829638

Estimated Scale Parameter: 1

Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]
[1,]	1	0	0
[2,]	0	1	0
[3,]	0	0	1

Model 2: Working unstructured correlation.

```

> xover.gee <- gee (relief ~ p2 + p3 + tx2 + tx3 + ptx2 + ptx3,
+                 data = xover3, scale.fix = TRUE,
+                 family = binomial, corstr = "unstructured")

[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
[1] -1.0864970 0.4123752 0.5866249 2.1055755 2.0683345 -0.1284182
    -0.9285768
> summary (xover.gee)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                               Logit
Variance to Mean Relation: Binomial
Correlation Structure:              Unstructured

Call:  gee(formula = relief ~ p2 + p3 + tx2 + tx3 + ptx2 + ptx3, data = xover3,
          family = binomial, corstr = "unstructured", scale.fix = TRUE)

Summary of Residuals:
      Min       1Q   Median       3Q      Max
-0.8289259 -0.2496275  0.1798135  0.2631610  0.8308251

Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
(Intercept) -1.10060006  0.3260360 -3.3757006  0.3217271 -3.4209122
p2           0.37925136  0.4755893  0.7974346  0.4099182  0.9251879
p3           0.54843106  0.4614023  1.1886181  0.4512896  1.2152530
tx2          2.10098624  0.3905658  5.3793404  0.4219768  4.9789141
tx3          2.13020321  0.3792224  5.6172930  0.4199083  5.0730204
ptx2         -0.09101728  0.4777062 -0.1905298  0.5041583 -0.1805331
ptx3         -0.87013747  0.4635191 -1.8772417  0.4336075 -2.0067400

Estimated Scale Parameter: 1
Number of Iterations: 3

Working Correlation
      [,1]      [,2]      [,3]
[1,] 1.00000000 -0.17244343 0.02875624
[2,] -0.17244343 1.00000000 0.03688866
[3,] 0.02875624 0.03688866 1.00000000

```

- The results are similar. The two doses of treatments seemed to be highly effective and the difference between them is small (odds ratios ~ 8).
- There is a slight evidence for the carry-over effect after the high dose analgesic (OR is 0.4).
- DHLZ Example 8.2 used different models (c.f., models 1 and 2 in Table 8.4, p152) with similar conclusions.

Seizure Data

A little data manipulation such that each person will have an additional row of baseline (or pre-treatment) data.

```
> seize <- read.table("../data/seize.data",
+                     col.names = c("id", "seizure", "week",
+                                   "progabide", "baseline8", "age"))
> seize.base <- data.frame(id = seize$id, seizure = seize$baseline8,
+                          progabide = seize$prog, age = seize$age,
+                          week = seize$week)
> seize.base <- seize.base[seize.base$week == 1,]
> seize.base$week <- 0
> seize.full <- rbind(seize[,nam], seize.base[,nam])
> seize.full <- seize.full[order(seize.full$id, seize.full$week),]
> seize.full$time <- ifelse(seize.full$week == 0, 8, 2)
> seize.full$post <- seize.full$week != 0
> seize.full[1:10,]
```

	id	age	progabide	week	seizure	time	post
1131	101	18	1	0	76	8	FALSE
113	101	18	1	1	11	2	TRUE
114	101	18	1	2	14	2	TRUE
115	101	18	1	3	9	2	TRUE
116	101	18	1	4	8	2	TRUE
1171	102	32	1	0	38	8	FALSE
117	102	32	1	1	8	2	TRUE
118	102	32	1	2	7	2	TRUE
119	102	32	1	3	9	2	TRUE
120	102	32	1	4	4	2	TRUE

Fit a Poisson family GEE model with exchangeable correlation.

```
> sg2 <- gee (seizure ~ progabide + post + post:progabide +
+           offset (log(time)),
+           data = seize.full, id = id, family = "poisson",
+           cor = "exchangeable")
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
[1] 1.347609 0.026515 0.110798 -0.103681
> summary (sg2)
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link: Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure: Exchangeable

Call: gee(formula = seizure ~ progabide + post + post:progabide
 + offset(log(time)), id = id, data = seize.full,
 family = "poisson", corstr = "exchangeable")

Summary of Residuals:

Min	1Q	Median	3Q	Max
-4.2991	-1.2991	2.0202	10.3746	147.0484

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	1.347609	0.15119	8.91364	0.15736	8.56402
progabide	0.026515	0.20727	0.12792	0.22185	0.11951
postTRUE	0.110798	0.15470	0.71620	0.11610	0.95434
progabide:postTRUE	-0.103681	0.21995	-0.47138	0.21361	-0.48537

Estimated Scale Parameter: 19.703
 Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.00000	0.77159	0.77159	0.77159	0.77159
[2,]	0.77159	1.00000	0.77159	0.77159	0.77159
[3,]	0.77159	0.77159	1.00000	0.77159	0.77159
[4,]	0.77159	0.77159	0.77159	1.00000	0.77159
[5,]	0.77159	0.77159	0.77159	0.77159	1.00000

- Offset (a term with constant coefficient) is used to take into account that the time periods are different (8 weeks vs 2 weeks).

$$\log(\mu_i) = \log(t_i \lambda_i) = \log(t_i) + \log \lambda_i = \log(t_i) + \mathbf{X}_i^T \boldsymbol{\beta}.$$

We can also allow the dispersion parameter to differ between the treatment and placebo group.

```
> library (geepack)
> sg2 <- geese (seizure ~ progabide + post + post:progabide +
+             offset (log(time)), sformula = ~ progabide,
+             data = seize.full, id = id, family = "poisson",
+             corstr = "exchangeable")
> summary (sg2)
```

```
Call: geese(formula = seizure ~ progabide + post + post:progabide +
  offset(log(time)), sformula = ~progabide, id = id,
  data = seize.full,
  family = "poisson", corstr = "exchangeable")
```

```
Mean Model:
Mean Link:          log
Variance to Mean Relation: poisson
```

```
Coefficients:
```

	estimate	san.se	wald	p
(Intercept)	1.347609	0.16202	69.185798	1.1102e-16
progabide	0.026515	0.22518	0.013865	9.0627e-01
postTRUE	0.110798	0.12040	0.846881	3.5744e-01
progabide:postTRUE	-0.103681	0.21598	0.230453	6.3119e-01

```
Scale Model:
Scale Link:          identity
```

```
Estimated Scale Parameters:
```

	estimate	san.se	wald	p
(Intercept)	14.2285	4.5929	9.59699	0.0019490
progabide	9.9102	16.6576	0.35395	0.5518873

```
Correlation Model:
Correlation Structure:  exchangeable
Correlation Link:      identity
```

```
Estimated Correlation Parameters:
```

	estimate	san.se	wald	p
alpha	0.74512	0.082562	81.45	0

```
Returned Error Value: 0
Number of clusters: 59
Maximum cluster size: 5
```

ALR Example

The `alr` package of R implements the ALR algorithm (http://www.biostat.harvard.edu/~carey/vcwww_4.html).

```
> x <- as.matrix (xover3[,c("p2", "p3", "tx2",
+                          "tx3", "ptx2", "ptx3")])
> y <- xover3$relief
> xover.alr <- alr (y ~ x, id = xover3$id,
+                 ainit = 0.01, depmodel = "exchangeable")
```

```
ALR: ALTERNATING LOGISTIC REGRESSION
alr S-function, version 4.4 98/02/24
```

```
Call: alr(formula = y ~ x, id = xover3$id, ainit = 0.01,
depmodel = "exchangeable")
```

Summary of Residuals:

	Min	1Q	Median	3Q	Max
	-0.8323602	-0.2528351	0.1818345	0.2674213	0.8300472

Coefficients:

	Estimate	Robust S.E.	Robust z
(Intercept)	-1.0835486	0.3175031	-3.4127182
xp2	0.4211660	0.4143912	1.0163488
xp3	0.5947027	0.4589374	1.2958254
xtx2	2.0912939	0.4180508	5.0024877
xtx3	2.0709832	0.4169999	4.9663874
xptx2	-0.1526329	0.5066495	-0.3012593
xptx3	-0.9235791	0.4440375	-2.0799573

Alpha:

	Estimate	Robust S.E.	Robust z
a1	-0.2243342	0.3753329	-0.5976938

```
Number of observations : 258
Number of Iterations  : 3
```

For an unstructured correlation matrix, there are $n_i(n_i - 1)$ parameters. For the balanced design here we have $n_i = 3$ and 3 odds ratios to estimate. We need first make a design matrix with $n_i(n_i - 1) \times q$ (6×3) elements where q is the number of parameters to estimate.

```
> z <- matrix (c(1, 0, 0,
+               0, 1, 0,
+               0, 0, 1,
+               1, 0, 0,
+               0, 1, 0,
+               0, 0, 1),
+             ncol = 3, byrow = TRUE)
```

- The rows correspond to pairs 12, 13, 23, 21, 31, 32.
- 12 and 21 are the same so they have the same rows.

```
> xover.alr <- alr (y ~ x, id = xover3$id,
+                 z = z,
+                 zmast = 1,
+                 zlocs = rep (1:3, 86),
+                 ainit = rep (0.01, 3),
+                 depmodel = "general")
```

```
[1] "alternating logistic regression - Splus, @(#) alr.q 4.4 98/02/24"
[1] "Running glm to get initial estimates"
[1] -1.0864970 0.4123752 0.5866249 2.1055755 2.0683345 -0.1284182 0.9
[1] "nobs"
[1] 258
```

```
> summary (xover.alr)
```

```
ALR: ALTERNATING LOGISTIC REGRESSION
alr S-function, version 4.4 98/02/24
```

```
Call: alr(formula = y ~ x, id = xover3$id, z = z, zmast = 1,
          zlocs = rep(1:3, 86), ainit = rep(0.01, 3), depmodel = "general")
```

```
Summary of Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.8339973	-0.2478472	0.1788862	0.2659418	0.8342610

Coefficients:

	Estimate	Robust S.E.	Robust z
(Intercept)	-1.1101269	0.3220709	-3.4468402
xp2	0.4314657	0.4209058	1.0250883
xp3	0.5989153	0.4620797	1.2961299
xtx2	2.1162817	0.4199172	5.0397601
xtx3	2.1254376	0.4205242	5.0542578
xptx2	-0.1681888	0.5134973	-0.3275359
xptx3	-0.9374704	0.4438234	-2.1122601

Alpha:

	Estimate	Robust S.E.	Robust z
a1	-0.3837662	0.4111212	-0.9334625
a2	0.1217843	0.7090241	0.1717632
a3	-0.3837668	0.4140482	-0.9268650

Number of observations : 258

Number of Iterations : 4

ALR in SAS

The implementation of ALR is perhaps more mature in SAS:

```
proc genmod data = xover3 descending;
  class id;
  model relief = p2 p3 tx2 tx3 ptx2 ptx3 / dist = bin;
  repeated subject = id / logor = fullclust modelse;
run;
```

GEE Model Information

Log Odds Ratio Structure	Fully Parameterized Clusters	Subject
Effect	id (86 levels)	Number of
Clusters	86	Correlation
Matrix Dimension	3	Maximum Cluster
Size	3	Minimum Cluster Size 3

Log Odds Ratio Parameter Information

Parameter	Group
Alpha1	(1, 2)
Alpha2	(1, 3)
Alpha3	(2, 3)

Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept	-1.0967	0.3204	-1.7247	-0.4686	-3.42	0.0006
p2	0.3757	0.4068	-0.4217	1.1731	0.92	0.3557
p3	0.5543	0.4488	-0.3253	1.4338	1.24	0.2168
tx2	2.0868	0.4201	1.2635	2.9101	4.97	<.0001
tx3	2.1214	0.4184	1.3014	2.9413	5.07	<.0001
ptx2	-0.0909	0.4999	-1.0708	0.8889	-0.18	0.8557
ptx3	-0.8627	0.4315	-1.7085	-0.0170	-2.00	0.0456
Alpha1	-0.9635	0.5923	-2.1244	0.1974	-1.63	0.1038
Alpha2	0.1072	0.7007	-1.2662	1.4807	0.15	0.8784
Alpha3	0.2877	0.6947	-1.0740	1.6493	0.41	0.6788

Other possible choices for `logor` are: `exch` (exchangeable), `logorvar` (allows the log OR to depend on another variable, e.g., center).

Seizure Data

For illustration, we assume an exchangeable correlation structure where the correlation may depend on age.

$$\begin{aligned}\text{Cor}(Y_{ij}, Y_{ik}) &= \rho_i \\ \log\left(\frac{1 + \rho_i}{1 - \rho_i}\right) &= \alpha_1 + \text{Age}_i \alpha_2\end{aligned}$$

This model can be fitted using **geese**. The design matrix for the correlation model has to be constructed by hand. The matrix \mathbf{Z} has the same number of rows as the number of clusters (the covariate should be invariant within a cluster).

```
> z <- cbind(1, seize.full$age[seize.full$week == 0])
> sg2 <- geese(seizure ~ progabide + post + post:progabide +
+             offset(log(time)), sformula = ~ progabide,
+             data = seize.full, id = id, family = "poisson",
+             cor.link = "fisherz", zcor = z,
+             corstr = "exchangeable")
> summary(sg2)
```

```
Call: geese(formula = seizure ~ progabide + post + post:progabide +
  offset(log(time)), sformula = ~progabide, id = id,
  data = seize.full,
  zcor = z, family = "poisson", cor.link = "fisherz",
  corstr = "exchangeable")
```

Mean Model:

```
Mean Link:                log
Variance to Mean Relation: poisson
```

Coefficients:

	estimate	san.se	wald	p
(Intercept)	1.413723297	0.1697429	6.936582e+01	1.110223e-16
progabide	0.003521332	0.2227135	2.499890e-04	9.873851e-01
postTRUE	0.119463923	0.1448016	6.806551e-01	4.093612e-01
progabide:postTRUE	-0.321152168	0.2880464	1.243073e+00	2.648795e-01

Scale Model:

Scale Link: identity

Estimated Scale Parameters:

	estimate	san.se	wald	p
(Intercept)	13.30317	4.188812	10.0862304	0.001493814
progabide	14.30612	19.040456	0.5645324	0.452439804

Correlation Model:

Correlation Structure: exchangeable

Correlation Link: fisherz

Estimated Correlation Parameters:

	estimate	san.se	wald	p
alpha:1	5.8652470	2.80875050	4.360600	0.03677934
alpha:2	-0.1370331	0.08139154	2.834606	0.09225342

Returned Error Value: 0

Number of clusters: 59

Maximum cluster size: 5

Further Reading

- Chapter 8 of DHLZ.