Lecture 9: Marginal Models for Clustered Data

1 Data and Model

This example uses the depression data in Section 12.1.1 of the Agresti textbook. The longitudinal study is interested in comparing a new drug with a standard drug for treatment of $n = 340$ subjects suffering from mental depression. Subjects were classified into two initial groups according to the severity of their depression. In each group, subjects were randomly assigned a treatment. Each subject's suffering from mental depression was classified as normal or abnormal at follow-up times of 1, 2 and 4 weeks. We are interested in modeling the binary response variable normal ($outcome = 1$) vs. abnormal as a function of two covariates and the treatment indicator:

1. severity: 0 = mild, 1 = severe
2. treat: 0 = standard drug, 1 = new drug
3. time: 0 = 1 week, 1 = 2 weeks, 2 = 4 weeks

```
rm(list=ls())
data = read.csv("depression.csv", header = T)
names(data)
```

```
## [1] "case"     "severity" "treat"    "time"     "outcome"
```

```
dim(data)
```

```
## [1] 1020    5
```

```
data[1:9,]
```

```
##   case severity treat time outcome
## 1    1        0     0    0       1
## 2    1        0     0    1       1
## 3    1        0     0    2       1
## 4    2        0     0    0       1
## 5    2        0     0    1       1
## 6    2        0     0    2       1
## 7    3        0     0    0       1
## 8    3        0     0    1       1
## 9    3        0     0    2       1
```

The data is in subject level (long) form where each of the 340 subjects has 3 observations for each of the three observation times.

We will use GEE to fit the logistic model from Agresti:

$$logit\left[P(outcome = 1)\right] = \beta_0 + \beta_1 * severity + \beta_2 * treat + \beta_3 * time + \beta_4 * treat * time.$$

2 GEE with Exchangeable Correlation Structure

First, let's assume the exchangeable correlation structure - this is a reasonable assumption given the nature of the repeated measures. Note that the $scale.fix = T$ option fixes the scale parameter to 1 which is necessary for binary data (as opposed to binomial).

```r
library(gee)
```

```
## Warning: package 'gee' was built under R version 3.4.4
```

```r
fit.gee.exch = gee(outcome ~ severity + treat + time + treat*time, id = case,
                   family=binomial, corstr="exchangeable", scale.fix=T, data=data)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
## (Intercept)    severity       treat        time  treat:time
## -0.02798843 -1.31391092 -0.05960381  0.48241209  1.01744498
```

```r
summary(fit.gee.exch)
```

```
##
##  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:     Exchangeable
##
## Call:
## gee(formula = outcome ~ severity + treat + time + treat * time,
##     id = case, data = data, family = binomial, corstr = "exchangeable",
##     scale.fix = T)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.94843397 -0.40683122  0.05156603  0.38832332  0.80238627
##
##
## Coefficients:
##                Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept) -0.02809866  0.1637503 -0.1715945   0.1741791 -0.1613205
## severity    -1.31391033  0.1459325 -9.0035505   0.1459630 -9.0016667
## treat       -0.05926689  0.2221626 -0.2667725   0.2285569 -0.2593091
## time         0.48246420  0.1149581  4.1968686   0.1199383  4.0226037
## treat:time   1.01719312  0.1890913  5.3793750   0.1877014  5.4192084
##
## Estimated Scale Parameter:  1
## Number of Iterations:  2
##
## Working Correlation
##              [,1]         [,2]         [,3]
## [1,]  1.000000000 -0.003432732 -0.003432732
## [2,] -0.003432732  1.000000000 -0.003432732
## [3,] -0.003432732 -0.003432732  1.000000000
```

The working correlation matrix indicates that the correlation between observations within a subject is estimated to be $-.003$. Because we specified an exchangeable correlation structure, this correlation is the same for all pairs in a group. This estimated correlation is rather small, our first indication that accounting for clustering may not be necessary.

```
summary(fit.gee.exch)$coefficients
```

```
##                Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept) -0.02809866  0.1637503 -0.1715945   0.1741791 -0.1613205
## severity    -1.31391033  0.1459325 -9.0035505   0.1459630 -9.0016667
## treat       -0.05926689  0.2221626 -0.2667725   0.2285569 -0.2593091
## time         0.48246420  0.1149581  4.1968686   0.1199383  4.0226037
## treat:time   1.01719312  0.1890913  5.3793750   0.1877014  5.4192084
```

The coefficient estimates are interpreted as usual in a logistic regression model. Unlike the usual ML fitting of a logistic regression model, the GEE analysis produces two sets of columns of standard errors and z-statistics. The *Naive* columns are model-based estimates and the *Robust* columns are the sandwich estimate. The naive S.E. is the standard error under the assumption that the correlation matrix has been correctly specified and estimated. The robust S.E. modifies the naive S.E. by a term that depends on the covariance of the model residuals across groups. It is robust in the sense that using it allows one to draw correct inferences from the data even if the correlation model was incorrectly specified.

The reported z-statistics can be treated in the usual way to carry out hypothesis tests and find confidence intervals.

3 GEE with Independent Correlation Structure

Next we fit the logistic model via GEE with the independent correlation structure.

```
fit.gee.ind = gee(outcome ~ severity + treat + time + treat*time, id = case,
                  family=binomial, corstr="independence", scale.fix=T, data=data)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)    severity       treat        time  treat:time
## -0.02798843 -1.31391092 -0.05960381  0.48241209  1.01744498
```

```
summary(fit.gee.ind)$coefficients
```

```
##                Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept) -0.02798843  0.1639083 -0.1707566   0.1741865 -0.1606808
## severity    -1.31391092  0.1464151 -8.9738733   0.1459845 -9.0003423
## treat       -0.05960381  0.2222080 -0.2682343   0.2285385 -0.2608042
## time         0.48241209  0.1147626  4.2035644   0.1199350  4.0222784
## treat:time   1.01744498  0.1887954  5.3891398   0.1876938  5.4207709
```

```
fit.gee.ind$working.correlation
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

The diagonal working correlation clearly indicates the assumption that the observations within a subject are assumed independent. The coefficient and standard error estimates are very similar to the fit assuming the exchangeable correlation structure. This similarity provides further evidence that accounting for the cluster is not necessary.

It is important to note that fitting the model using GEE with an indepedent correlation structure yields the same result as fitting the usual binomial GLM assuming all 1020 observations are independent:

```
fit.glm = glm(outcome ~ severity + treat + time + treat*time, family=binomial, data=data)
summary(fit.glm)
```

```
##
## Call:
## glm(formula = outcome ~ severity + treat + time + treat * time,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4352  -1.0220   0.3254   0.9915   1.8009
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02799    0.16391  -0.171    0.864
## severity    -1.31391    0.14641  -8.974  < 2e-16 ***
## treat       -0.05960    0.22221  -0.268    0.789
## time         0.48241    0.11476   4.204 2.63e-05 ***
## treat:time   1.01744    0.18879   5.389 7.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.9  on 1019  degrees of freedom
## Residual deviance: 1161.9  on 1015  degrees of freedom
## AIC: 1171.9
##
## Number of Fisher Scoring iterations: 4
```

Note that these estimates are also reported in the GEE results because the GLM estimates are used as the initial estimates in the GEE algorithm.

4 GEE with Unstructured Correlation Structure

The choice of the correlation structure depends on the nature of the data. For this data, the unstructured correlation is likely too general as we expect similar relations between all pairs. Nonetheless, it is interesting to assess the estimates, specifically the estimated correlation.

```
fit.gee.unstr = gee(outcome ~ severity + treat + time + treat*time, id = case,
                    family=binomial, corstr="unstructured", scale.fix=T, data=data)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

## (Intercept)    severity       treat        time  treat:time
## -0.02798843 -1.31391092 -0.05960381  0.48241209  1.01744498
```

```
fit.gee.unstr$working.correlation
```

```
##             [,1]        [,2]        [,3]
## [1,]  1.00000000  0.07393977 -0.02741128
## [2,]  0.07393977  1.00000000 -0.05669559
## [3,] -0.02741128 -0.05669559  1.00000000
```

We find that all of the pairwise correlations are estimated to be small, with the smallest correlation between week 1 and week 4.