R_HW11_Generalized Estimating Equations

Eom SangJun

2020 12 5

Quasi-likelihood Approach 가 GLMs 에 비해 갖는 장점은 response 에 대해 특정 분포를 지정할 필요가 없다는 점이었다. 우리는 이러한 점을 반복측정 자료 또는 longitudinal study 에도 접목시킬 수 있다. Yi 가 특정 subject 에 대한 반복측정 자료를 모아 놓은 vector 라고 했을 때 yij 를 그 원소라고 하자. 이 때 GLMs 은 Yi 에 대해 분포를 가정했었다면, 우리가 하고자 하는 방식은 yij 에 분포를 지정하되, 그 joint distribution 인 Yi 에 대해서는 특정 분포를 지정하지 않는 방식이다. 이렇게 했을 때 기존 방법과 마찬가지로 E(Yi) = μ_i , $g(\mu_i) = x_i^T\beta$를 지정해서 베타를 구할 수 있는데, 이 때 보통 Analytically 구해지지는 않는다. 따라서 Iteration 방법을 통해 converge 할 때까지 solution 을 구하게 되는데 이러한 방정식을 푸는 과정, 또는 방식을 Generalized Estimating Equations 이라고 부른다. GEE 의 장점은 Joint distribution 을 특정하지 않아도 될 뿐더러, variance 를 잘못 특정 짓더라도, 베타의 추정치가 consistent 하다는 것이 있다.

R 에서 GEE 를 사용할 수 있는 package 는 'geepack'이다.

```
data(ctsib, package='faraway')
ctsib$stable <- ifelse(ctsib$CTSIB==1,1,0)
library(geepack)
```

```
## Warning: package 'geepack' was built under R version 4.0.3
```

우선 우리는 앞서 GLMM 을 사용해서 분석했던 때와 동일한 fixed effect 들을 지정해줄 것이다. Grouping 을 지어주는 argument 는 'id'를 사용하며, 아쉬운 점은 nested grouping variable 을 지정해주는 것은 힘들고 simple group 만 허용이 된다는 점이다. 'corstr'는 각 그룹 내에서 correlation structure 를 지정해주는 argument 이다. 만약 correlation 이 없다고 하면, GLM 과 동일해진다. 'exchangeable'은 우리가 배웠던 compound symmetry 와 동일하다. 우리는 GLMM fit 과 호환성을 최대로 하기 위해 scale parameter 의 값을 default value 인 1 로 fix 하였다. 굳이 이 목적이 아니라면 scale 을 fix 할 필요는 없다.

```
modgeep <- geeglm(stable ~ Sex + Age + Height + Weight + Surface + Vision,
                  id=Subject, corstr='exchangeable', scale.fix=TRUE,
```

```
                  data = ctsib, family=binomial)
summary(modgeep)

##
## Call:
## geeglm(formula = stable ~ Sex + Age + Height + Weight + Surface +
##     Vision, family = binomial, data = ctsib, id = Subject, corstr = "excha
ngeable",
##     scale.fix = TRUE)
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)  8.62332  5.91992  2.122   0.1452
## Sexmale      1.64488  0.90347  3.315   0.0687 .
## Age         -0.01205  0.04802  0.063   0.8019
## Height      -0.10211  0.04239  5.801   0.0160 *
## Weight       0.04365  0.03399  1.649   0.1991
## Surfacenorm  3.91632  0.56682 47.738 4.87e-12 ***
## Visiondome   0.35888  0.40403  0.789   0.3744
## Visionopen   3.17990  0.46063 47.657 5.08e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.2185 0.04467
## Number of clusters:   40  Maximum cluster size: 12
```

→ Estimated Correlation Parameter 인 alpha 값을 보면 동일한 subject 내의 observation 값들 사이의 correlation 값이 약 0.22 라는 것을 알 수 있으며, std.err 값을 보았을 때 correlation 이 있다고 꽤 확신할 수 있는 정도이다.

Coefficient 에서 std.err 값들은 sandwich estimator 를 이용해 추정된 값이며 일반적으로 likelihood 방식보다 크다는 특징이 있다(항상 그런 것은 아님). 이 standard error 값들은 옆에 있는 Wald statistic 값을 도출하는 데에 사용된다. 그리고 Wald 값을 통해 p-value 를 계산해보았을 때 surface 와 vision 이 significant 하며, Height 와 Gender 정도가 marginally significant 하다는 것을 알 수 있다.

이는 GLMM 에서의 결과와 비슷하다.

다만 GLMM 과는 확연한 차이점이 있는데 그것은 바로 GEE 의 Coefficients estimates 값들이 GLMM 의 beta 값들보다 반 정도 수준이라는 것이다. GLMMs 은 subject 또는 individual level 에서 data 를 모델링한다. 또한 개인간 측정에 따른 correlation 은 random effect 를 통해 발생한다. 따라서 GLMM 에서 betas 값은 개인에 대한 effect 를 나타낸다. 반면에 GEE Model 은 Population level 에서 data 를 모델링한다. GEE 에서 베타는 동일한 predictor value 를 가진 모든 개인들에 대한 predictor effect 의 평균값을 나타낸다. GEE 는 random effect 를 사용하지 않으며, 다만 marginal 또는 correlation 수준에서의 correlation 을 모델링한다.

앞서 Vision 을 살펴보았을 때 특정 level 은 significant 한 반면, 특정 level 은 significant 하지 않다고 나왔다는 것을 알 수 있다. 이는 Vision 의 level 이 세 개인데, 따라서 두 번의 testing 을 따로 진행했기 때문이다. 따라서 이러한 문제를 해결하기 위해서는 anova test 를 해볼 수 있다.

```
modgeep2 <- geeglm(stable ~ Sex + Age + Height + Weight + Surface,
                   id = Subject, corstr = 'exchangeable', scale.fix = TRUE,
                   data=ctsib, family=binomial)
anova(modgeep2, modgeep)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 stable ~ Sex + Age + Height + Weight + Surface + Vision
## Model 2 stable ~ Sex + Age + Height + Weight + Surface
##   Df   X2 P(>|Chi|)
## 1  2 58.4   2.1e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ 우리가 앞서 살펴본 것과 같이 Vision 은 상당히 significant 하다.

참고로 ordgee() function 을 이용하면 ordinal response 도 modeling 할 수 있다.

이번에는 앞서 살펴보았던 발작 data 를 geeglm 을 통해 분석해보자.

```
data(epilepsy, package = 'faraway')
```

49 번째 데이터는 마찬가지로 제외하고, AR(1) model 을 correlation structure 로 지정해주었다.

```r
modgeep <- geeglm(seizures ~ offset(log(timeadj)) + expind + treat + I(expind
*treat),
                  id=id, family=poisson, corstr = 'ar1',
                  data=epilepsy, subset=(id!=49))
summary(modgeep)
```

```
##
## Call:
## geeglm(formula = seizures ~ offset(log(timeadj)) + expind + treat +
##     I(expind * treat), family = poisson, data = epilepsy, subset = (id !=
##     49), id = id, corstr = "ar1")
##
##  Coefficients:
##                    Estimate Std.err  Wald Pr(>|W|)
## (Intercept)          1.3138  0.1616 66.10  4.4e-16 ***
## expind               0.1509  0.1108  1.86    0.173
## treat               -0.0797  0.1983  0.16    0.688
## I(expind * treat)   -0.3987  0.1745  5.22    0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     10.6    2.35
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.783  0.0519
## Number of clusters:   58  Maximum cluster size: 5
```

→ interaction term 으로 측정된 drug effect 는 significant 하다는 것을 알 수 있다.

Dispersion parameter 는 10.6 으로 측정되었다. 이는 우리가 만약 overdispersion 을 고려하지 않았다면, standard error 는 훨씬 더 컸을 것이라는 의미이다. AR(1) correlation structure 는 working correlation 의 adjacent measurement 가 0.78 correlation 을 가진다는 것을 확인할 수 있다.