R\_HW11\_Generalized Estimating Equations

Eom SangJun

2020 12 5

Quasi-likelihood Approach가 GLMs에 비해 갖는 장점은 response에 대해 특정 분포를 지정할 필요가 없다는 점이었다. 우리는 이러한 점을 반복측정 자료 또는 longitudinal study에도 접목시킬 수 있다. Yi가 특정 subject에 대한 반복측정 자료를 모아 놓은 vector라고 했을 때 yij 를 그 원소라고 하자. 이 때 GLMs은 Yi에 대해 분포를 가정했었다면, 우리가 하고자 하는 방식은 yij에 분포를 지정하되, 그 joint distribution인 Yi에 대해서는 특정 분포를 지정하지 않는 방식이다. 이렇게 했을 때 기존 방법과 마찬가지로 E(Yi) = , 를 지정해서 베타를 구할 수 있는데, 이 때 보통 Analytically 구해지지는 않는다. 따라서 Iteration 방법을 통해 converge할 때까지 solution을 구하게 되는데 이러한 방정식을 푸는 과정, 또는 방식을 Generalized Estimating Equations이라고 부른다. GEE의 장점은 Joint distribution을 특정하지 않아도 될 뿐더러, variance를 잘못 특정 짓더라도, 베타의 추정치가 consistent하다는 것이 있다.

R에서 GEE를 사용할 수 있는 package는 ‘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 GLMM을 사용해서 분석했던 때와 동일한 fixed effect들을 지정해줄 것이다. Grouping을 지어주는 argument는 ‘id’를 사용하며, 아쉬운 점은 nested grouping variable을 지정해주는 것은 힘들고 simple group만 허용이 된다는 점이다. ‘corstr’는 각 그룹 내에서 correlation structure를 지정해주는 argument이다. 만약 correlation이 없다고 하면, GLM과 동일해진다. ‘exchangeable’은 우리가 배웠던 compound symmetry와 동일하다. 우리는 GLMM fit과 호환성을 최대로 하기 위해 scale parameter의 값을 default value인 1로 fix하였다. 굳이 이 목적이 아니라면 scale을 fix할 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 = "exchangeable",   
## 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 Parameter인 alpha값을 보면 동일한 subject 내의 observation 값들 사이의 correlation값이 약 0.22라는 것을 알 수 있으며, std.err 값을 보았을 때 correlation이 있다고 꽤 확신할 수 있는 정도이다.

Coefficient에서 std.err 값들은 sandwich estimator를 이용해 추정된 값이며 일반적으로 likelihood 방식보다 크다는 특징이 있다(항상 그런 것은 아님). 이 standard error 값들은 옆에 있는 Wald statistic값을 도출하는 데에 사용된다. 그리고 Wald 값을 통해 p-value를 계산해보았을 때 surface와 vision이 significant하며, Height와 Gender 정도가 marginally significant하다는 것을 알 수 있다.

이는 GLMM에서의 결과와 비슷하다.

다만 GLMM과는 확연한 차이점이 있는데 그것은 바로 GEE의 Coefficients estimates값들이 GLMM의 beta값들보다 반 정도 수준이라는 것이다. GLMMs은 subject또는 individual level에서 data를 모델링한다. 또한 개인간 측정에 따른 correlation은 random effect를 통해 발생한다. 따라서 GLMM에서 betas 값은 개인에 대한 effect를 나타낸다. 반면에 GEE Model은 Population level에서 data를 모델링한다. GEE에서 베타는 동일한 predictor value를 가진 모든 개인들에 대한 predictor effect의 평균값을 나타낸다. GEE는 random effect를 사용하지 않으며, 다만 marginal 또는 correlation 수준에서의 correlation을 모델링한다.

앞서 Vision을 살펴보았을 때 특정 level은 significant한 반면, 특정 level은 significant하지 않다고 나왔다는 것을 알 수 있다. 이는 Vision의 level이 세 개인데, 따라서 두 번의 testing을 따로 진행했기 때문이다. 따라서 이러한 문제를 해결하기 위해서는 anova test를 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 Vision은 상당히 significant하다.

참고로 ordgee() function을 이용하면 ordinal response도 modeling할 수 있다.

이번에는 앞서 살펴보았던 발작 data를 geeglm을 통해 분석해보자.

data(epilepsy, package = 'faraway')

49번째 데이터는 마찬가지로 제외하고, AR(1) model을 correlation structure로 지정해주었다.   
  
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 term으로 측정된 drug effect는 significant하다는 것을 알 수 있다.

Dispersion parameter는 10.6으로 측정되었다. 이는 우리가 만약 overdispersion을 고려하지 않았다면, standard error는 훨씬 더 컸을 것이라는 의미이다. AR(1) correlation structure는 working correlation의 adjacent measurement가 0.78 correlation을 가진다는 것을 확인할 수 있다.