Lesson 25

Clark

Recall, one way for us to model count data or binary data when we had over or under dispersion was through using quasi likelihood methods where instead of stipulating a likelihood, we stipulated a variance to mean ratio. That is:

We can do the same thing with clustered data. Recall that in a GLMM we have to stipulate the covariance structure. That is, we have to specify whether our error terms are independent, AR(1), spatially structured, whatever. Maybe that is wrong? Maybe they aren't? As we are perhaps seeing in our projects if we change the error term our covariate values inevitably change.

Let's look at an example. Recal lfrom Faraway there was a study form a clinical trial of 59 epileptics. The patience were randomized to treatment by the drug Progabide (31 patients) or to the placebo group (28 patience). The patience were observed for four 2-week periods and the number of seizures were recorded. The primary research question was does Progabide reduce the rate of seizures.

One model:

```r
library(tidyverse)
library(faraway)
library(INLA)
library(geepack)
data(epilepsy)
epilepsy$period <- rep(0:4,59)
epilepsy$drug <- factor(c("placebo","treatment")
                        [epilepsy$treat+1])
epilepsy$phase <- factor(c("baseline","experiment")
                         [epilepsy$expind+1])

formula <- seizures ~ offset(log(timeadj)) + expind +
  treat + I(expind*treat) + f(period, model="ar1",
                              group=id)

result <- inla(formula, family="poisson",data=epilepsy)
```

```
formula2 <- seizures ~ offset(log(timeadj)) + expind +
  treat + I(expind*treat) + f(period, model="iid",
                              group=id)

result2 <- inla(formula2, family="poisson",data=epilepsy)


epi <- epilepsy
epi$obs <- seq(1,nrow(epi))


formula3 <- seizures ~ offset(log(timeadj)) + expind +
  treat + I(expind*treat) + f(obs,model="iid")

result3 <- inla(formula3, family="poisson",data=epi)
```

Another option:

Clearly the results are impacted by the choice of our correlation structure. Similar to quasi methods, we can make our answers robust to misspecification of the assumptions of the GLMM. Similar to a quasi-model we start with stipulating a variance function and also specify a *working correlation matrix* that describes the relationship between observations within the same structure. These two pieces together form a working covariance matrix. That is:

The working covariance matrix then allows us to create a multivariate version of the score function:

We then alternatively estimate α and the β terms in the model until they do not change.

We can do this for our seizures data

```
gee_mod <- geeglm(seizures ~ offset(log(timeadj)) + expind +
  treat + I(expind*treat), data=epilepsy,id=id,family=poisson,
  corstr="ar1")
```

Some other correlation structures we might consider are:

```
gee_mod2 <- geeglm(seizures ~ offset(log(timeadj)) + expind +
  treat + I(expind*treat), data=epilepsy,id=id,family=poisson,
  corstr="exchangeable")
```

Now, an important point to note here is that we no longer are building out a model for a Y_{ij} term, rather we are building out a model for the vector Y_i. That is, our covariates represent the effect at the population level. The β terms represent the effect of the predictors averaged across all individuals with the same predictor value. What this means is that parameter estimates (and standard errors) from random effects models will be greater than those for marginal models. The GEE model ignores changes **within** subject. GEE models are used when the target of the marginal model is the population while the random effect model the target is the individual.

GEEs are NOT model based. That is, I cannot predict the outcome for an individual like I can with a GLMM, however this comes at a cost that if you misspecify the model for the GLMM you cannot trust your results anyhow. . . .

GEE models can also be useful for binary data. Consider the fall data we introduced earlier

```
data(ctsib)

ctsib$stable <- ifelse(ctsib$CTSIB==1,1,0)

bingee <- geeglm(stable ~ Sex + Age + Height+ Weight + Surface +
                   Vision, id=Subject, corstr="exchangeable",
                 data=ctsib,family=binomial)

summary(bingee)
```

```
##
## Call:
## geeglm(formula = stable ~ Sex + Age + Height + Weight + Surface +
##     Vision, family = binomial, data = ctsib, id = Subject, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)  8.61677  5.91687  2.121   0.1453
## Sexmale      1.64358  0.90325  3.311   0.0688 .
## Age         -0.01198  0.04801  0.062   0.8030
## Height      -0.10208  0.04237  5.803   0.0160 *
## Weight       0.04365  0.03398  1.650   0.1989
## Surfacenorm  3.91664  0.56665 47.775 4.78e-12 ***
## Visiondome   0.35891  0.40408  0.789   0.3744
## Visionopen   3.17998  0.46054 47.678 5.02e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.7315  0.6742
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.2169  0.2032
## Number of clusters:   40  Maximum cluster size: 12
```

To see the effect of vision we can run:

```
bingee2<- geeglm(stable ~ Sex + Age + Height+ Weight + Surface,
                 id=Subject, corstr="exchangeable",
                 data=ctsib,family=binomial)

anova(bingee2,bingee)
```

```
## 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 58.5     2e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I have no idea how to assess a GEE for goodness of fit but remember it is robust to misspecification for covariance structure so it may not matter as much as a GLM or a GLMM. If we want to analyze the impact of various covariates by looking at the Wald intervals or conducing an ANOVA type test.

The biggest factor in determining whether to use a GEE model or a GLMM boils down to this example (at least in my mind)

If you are a doctor and you want an estimate of how much a statin drug will lower your patient's odds of getting a heart attack, the subject-specific coefficient is the clear choice. On the other hand, if you are a state health official and you want to know how the number of people who die of heart attacks would change if everyone in the at-risk population took the stain drug, you would probably want to use the population–averaged coefficients. (Allison, 2009)

To see this, consider a marginal model for y with link function g is of the form $g(\mu_{ij}) = x_{ij}\beta$. For a linear model $g()$ is typically the identity so, if we have, say, y_{ij} that denotes the score on exam j for student i a marginal model would have $\mu_{ij} = \beta_{0j} + \beta_{1j}x_i$. That is, β_{1j} describes the average effect of exam j on the response (which is true for all students). This is a marginal effect.

If we consider a GLMM we have the form $g[E(y_{ij}|u_i)]$ where u_i are our random effects. If we consider a simple random effect for person in the above example our model becomes

$$E(y_{ij}|u_i) = \beta_{0j} + u_i + \beta_1 x_i$$

That is, β_1 is now the **conditional** effect of x_i given u_i.

We can take this one step further to see what the relationship between GLMMs and GEEs are by marginalizing out u_i. In general we have, for GLMMs, $E(y_{ij}|u_i) = g^{-1}(x_{ij}\beta + u_i)$. We can marginalize this:

If g is the identity link function we have:

Therefore our models are the exact same. However, if our link function is NOT the identity, than this does not hold, hence GEE models are different, in general, from GLMMs. Phew.

So. Could you use a GEE for your final project?

The `ohio` data concern 536 children from Steubenville, Ohio and were taken as part of a study on the effects of air pollution. Children were in the study for 4 years from ages 7 to 10. The response was whether they wheezed or not.

1. Construct a table that shows proportion of children who wheeze for 0,1,2,3, or 4 years broken down by maternal smoking status.
2. Fit an appropriate GLMM to check for the impacts of age and maternal smoking effect.
3. Fit a GEE with an AR(1) structure. What does an AR(1) structure mean in context of this problem?
4. Fit a GEE with an exchangeable error structure.
5. Compare the parameter estimates from the GLMM to the two GEE models. Can you conclude that age and maternal smoking are significant?