Chapter 6: Multilevel Modeling

Tyson S. Barrett

Summer 2017

Utah State University

Introduction

Good Quote

“Simplicity does not precede complexity, but follows it.”

— Alan Perlis

Introduction

Multilevel data are more complex and don't meet the assumptions of regular linear or generalized linear models. But with the right modeling schemes, the results can be very interpretable and actionable.

Two powerful forms of multilevel modeling are:

1. Generalized Estimating Equations (GEE)
2. Mixed effects (ME; i.e., hierarchical linear modeling, multilevel modeling)

GEE and ME

Similarities:

- they both attempt to control for the lack of independence within clusters, although they do it in different ways.
- built on linear regression which makes them flexible and powerful at finding relationships in the data.

GEE and ME

Differences:

- The interpretation is somewhat different between the two.
- GEE is a population-averaged (e.g., marginal) model whereas ME is subject specific. In other words, *GEE is the average effect* while *ME is the effect found in the average person*.
- In a linear model, these coefficients are the same but when we use different forms such as logistic or poisson, these can be quite different (although in my experience they generally tell a similar story). - ME models are much more complex than the GEE models and can struggle with convergence compared to the GEE.
- This also means that GEE's are generally fitted much more quickly.

GEE and ME

Still the choice of the modeling technique should be driven by your hypotheses and not totally dependent on speed of the computation.

Prep the Data

For both modeling techniques we want our data in long form.

- What this implies is that each row is an observation.
- What this actually means about the data depends on the data.
- For example, if you have repeated measures, then often data is stored in wide form—a row is an individual.
- To make this long, we want each time point within a person to be a row—a single individual can have multiple rows but each row is a unique observation.

The NHANES data is in long form since we are working within community clusters within this data. So, each row is an observation and each cluster has multiple rows.

Prep the Data

Note that although these analyses will be within community clusters instead of within subjects (i.e. repeated measures), the overall steps will be the exact same.

Reminder

This is not a multilevel modeling course. For this class we are only going to demonstrate basic examples of it.

Generalized Estimating Equations

GEE

There are two packages, intimately related, that allow us to perform GEE modeling

1. `gee` and
2. `geepack`.

These have some great features and make running a fairly complex model pretty simple.

However, as great as they are, there are some annoying shortcomings.

GEE

GEE's, in general, want a few pieces of information from you.

1. The outcome and predictors
2. A correlation structure
3. A variable that is cluster ID's.
4. The family (i.e. the type of distribution).

Since this is not longitudinal, but rather clustered within communities, we'll assume for this analysis an unstructured correlation structure. It is the most flexible and we have enough power for it here.

GEE

For geepack to work, we need to filter out the missing values for the variables that will be in the model.

```
df2 <- df %>%
  filter(complete.cases(dep, famsize, sed, race, asthma))
```

GEE

We'll build the model with both packages (just for demonstration).

GEE

```r
library(gee)
fit_gee <- gee(dep ~ asthma + famsize + sed + race,
               data = df2,
               id = df2$sdmvstra,
               corstr = "unstructured")
summary(fit_gee)$coef

library(geepack)
fit_geeglm <- geeglm(dep ~ asthma + famsize + sed + race,
                     data = df2,
                     id = df2$sdmvstra,
                     corstr = "unstructured")
summary(fit_geeglm)
```

GEE

```
##       (Intercept)      asthmaAsthma            famsize                sed
##       2.500022059       1.356081567      -0.042132178       0.001362226
## raceOtherHispanic          raceWhite         raceBlack          raceOther
##       1.184995689       0.113949209       0.100536695      -0.555478773
```

```
##                      Estimate   Naive S.E.    Naive z  Robust S.E.
## (Intercept)        2.495509790 0.2867816215  8.7017773 0.2690426648
## asthmaAsthma       1.353039007 0.1867101195  7.2467363 0.2137975620
## famsize           -0.039489294 0.0461945052 -0.8548483 0.0457474654
## sed                0.001358042 0.0003362291  4.0390382 0.0003551901
## raceOtherHispanic  1.192481318 0.3075562837  3.8772783 0.3309608614
## raceWhite          0.116185743 0.2531554533  0.4589502 0.2279687738
## raceBlack          0.096800821 0.2625826864  0.3686489 0.2360498473
## raceOther         -0.555053605 0.2809301544 -1.9757708 0.2406566044
##                     Robust z
## (Intercept)        9.2755169
## asthmaAsthma       6.3285989
## famsize           -0.8632018
## sed                3.8234244
## raceOtherHispanic  3.6030886
## raceWhite          0.5096564
## raceBlack          0.4100864
## raceOther         -2.3064133
```

GEE

```
##
## Call:
## geeglm(formula = dep ~ asthma + famsize + sed + race, data = df2,
##     id = df2$sdmvstra, corstr = "unstructured")
##
##  Coefficients:
##                     Estimate    Std.err   Wald Pr(>|W|)
## (Intercept)        2.5579361  0.2700717 89.706  < 2e-16 ***
## asthmaAsthma       1.3492892  0.2156202 39.159 3.91e-10 ***
## famsize           -0.0446716  0.0457087  0.955 0.328415
## sed                0.0013015  0.0003548 13.454 0.000244 ***
## raceOtherHispanic  1.1750373  0.3318983 12.534 0.000400 ***
## raceWhite          0.0806377  0.2295661  0.123 0.725392
## raceBlack          0.0642028  0.2363255  0.074 0.785875
## raceOther         -0.5902049  0.2413379  5.981 0.014463 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)    19.49  0.7843
##
## Correlation: Structure = unstructured  Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2  0.12480 0.01654
## alpha.1:3  0.42070 0.10339
## alpha.1:4  2.89640 1.06678
```

GEE

The `gee` package doesn't directly provide p-values but provides the z-scores, which can be used to find the p-values.

The `geepack` provides the p-values in the way you'll see in the `lm()` and `glm()` functions.

These models are interpreted just as the regular GLM. It has adjusted for the correlations within the clusters and provides valid standard errors and p-values.

Mixed Effects

Mixed Effects

It is called "mixed effects" because we include both fixed and random effects into the model simultaneously.

- Random effects are those that we don't necessarily care about the specific values but want to control for it and/or estimate the variance.
- Fixed effects are those we are used to estimating in linear models and GLM's.

Mixed Effects

These are a bit more clear with an example.

- We will do the same overall model as we did with the GEE but we'll use ME.
- Use the `lme4` package

```
library(lme4)
fit_me <- lmer(dep ~ asthma + famsize + sed + race + (1 | cluster),
               data = df2,
               REML = FALSE)
summary(fit_me)
```

Mixed Effects

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: dep ~ asthma + famsize + sed + race + (1 | cluster)
##    Data: df2
##
##      AIC      BIC   logLik deviance df.resid
##    25780    25844   -12880    25760     4427
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.327 -0.635 -0.355  0.272  5.435
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  cluster  (Intercept)  0.105   0.324
##  Residual             19.389   4.403
## Number of obs: 4437, groups:  cluster, 14
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)        2.491678   0.302768    8.23
## asthmaAsthma       1.335445   0.186618    7.16
## famsize           -0.042857   0.046341   -0.92
## sed                0.001425   0.000337    4.23
## raceOtherHispanic  1.289890   0.320595    4.02
## raceWhite          0.008348   0.259449    0.03
## raceBlack          0.171658   0.273382    0.63
## raceOther         -0.552746   0.285512   -1.94
##
## Correlation of Fixed Effects:
```

Mixed Effects

There are no p-values provided here. This is because degrees of freedom are not well-defined in the ME framework.

A good way to test it can be through the `anova()` function, comparing models. Let's compare a model with and without `asthma` to see if the model is significantly better with it in.

Mixed Effects

```
fit_me1 <- lmer(dep ~ famsize + sed + race + (1 | cluster),
                data = df2,
                REML = FALSE)

anova(fit_me, fit_me1)
```

```
## Data: df2
## Models:
## fit_me1: dep ~ famsize + sed + race + (1 | cluster)
## fit_me: dep ~ asthma + famsize + sed + race + (1 | cluster)
##         Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## fit_me1  9 25829 25886 -12905    25811
## fit_me  10 25780 25844 -12880    25760  50.9      1    9.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mixed Effects

This comparison strongly suggests that `asthma` is a significant predictor ($\chi^2 = 50.5$, $p < .001$). We can do this with both fixed and random effects, as below:

```
fit_me2 <- lmer(dep ~ famsize + sed + race + (1 | cluster),
                data = df2,
                REML = TRUE)
fit_me3 <- lmer(dep ~ famsize + sed + race + (1 + asthma | cluster),
                data = df2,
                REML = TRUE)
```

Mixed Effects

```r
anova(fit_me2, fit_me3, refit = FALSE)
```

```
## Data: df2
## Models:
## fit_me2: dep ~ famsize + sed + race + (1 | cluster)
## fit_me3: dep ~ famsize + sed + race + (1 + asthma | cluster)
##         Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## fit_me2  9 25855 25912 -12918    25837
## fit_me3 11 25821 25892 -12900    25799  37.3      2      8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, including random slopes for asthma appears to be significant ($\chi^2 = 36.9$, p $< .001$).

Mixed Effects

Linear mixed effects models converge pretty well. You'll see that the conclusions and estimates are very similar to that of the GEE.

For generalized versions of ME, the convergence can be harder and more picky. As we'll see below, it complains about large eigenvalues and tells us to rescale some of the variables.

Generalized Mixed Effects

```
library(lme4)
fit_gme <- glmer(dep2 ~ asthma + famsize + sed + race + (1 | cluster),
                 data = df2,
                 family = "binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00854237 (
## 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$ch
##  - Rescale variables?;Model is nearly unidentifiable: large eigenval
##  - Rescale variables?
```

Warnings!

- `sed` is huge compared to the other variables.
- If we simply rescale it, using the `I()` function within the model formula, we can rescale it by 1,000. - Here, that is all it needed to converge.

Generalized Mixed Effects

```
library(lme4)
fit_gme <- glmer(dep2 ~ asthma + famsize + I(sed/1000) + race + (1 | cl
                 data = df2,
                 family = "binomial")
summary(fit_gme)
```

Generalized Mixed Effects

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: dep2 ~ asthma + famsize + I(sed/1000) + race + (1 | cluster)
##    Data: df2
##
##      AIC      BIC   logLik deviance df.resid
##     2665     2722    -1323     2647     4428
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -0.635 -0.329 -0.295 -0.258  5.032
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  cluster (Intercept) 0.0232   0.152
## Number of obs: 4437, groups:  cluster, 14
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.6316     0.2435  -10.81  < 2e-16 ***
## asthmaAsthma        0.5619     0.1281    4.39  1.2e-05 ***
## famsize            -0.0336     0.0374   -0.90   0.3696
## I(sed/1000)         0.5835     0.2618    2.23   0.0258 *
## raceOtherHispanic   0.7564     0.2421    3.12   0.0018 **
## raceWhite           0.0955     0.2159    0.44   0.6581
## raceBlack           0.0531     0.2277    0.23   0.8155
## raceOther          -0.4950     0.2591   -1.91   0.0560 .
## ---
```

Conclusions

Conclusions

This has been a really brief introduction into a thriving, large field of statistical analyses. These are the general methods for using R to analyze multilevel data. Our next chapter will discuss more modeling techniques in R, including mediation, mixture, and structural equation modeling.