Lesson 17

Clark

Last class we talked about clustered observations and said sometimes, like in Chicago, we have clustering by geographical area. Here we have multiple observations that all occur on the same cluster. We can generalize this a bit more and consider multiple forms of clustering.

For example, let's consider example 9.2.3 from Agresti. Here he discusses a study of the efficacy of two programs for discouraging young people from starting or continuing to smoke. The study compared four groups, defined by a 2×2 full factorial design according to whether a student was exposed to a school-based curriculum, and a television-based prevention program. The subjects (and this is key here) were 1600 sevent grade students from 135 classrooms in 28 Los Angeles schools. The schools were randomly assigned to the four intervention conditions. We will assume our response is normally distributed and measures tobacco and health knwoledge of an individual student. They also provide information on the pre knowledge of the student prior to getting the training.

What are our observational units?

Remember our observational units are what the response is measured on. Our *experimental units* are what we apply the treatment on. What are the experimental units here?

When we are in a world where our experimental units are different than our observational units, we need to account for this in someways. Let's start with our mechanism of interest, that is the fixed effects we want to observe. Here we can structure this as:

However, are there unique aspects of each school that might make the mechanism of interest manifest differently? What about each classroom? What about each student? Therefore, we can write:

We can fit this as:

```
library(lme4)
library(tidyverse)
smoking <- read.table("https://users.stat.ufl.edu/~aa/glm/data/Smoking.dat",header=TRUE)

model <- lmer(y~PTHK + SC + TV + (1|school) + (1|class), data=smoking)
summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ PTHK + SC + TV + (1 | school) + (1 | class)
##    Data: smoking
##
## REML criterion at convergence: 5374.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5202 -0.6975 -0.0177  0.6875  3.1630
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  class    (Intercept) 0.06853  0.2618
##  school   (Intercept) 0.03925  0.1981
##  Residual             1.60108  1.2653
## Number of obs: 1600, groups:  class, 135; school, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.78493    0.11295  15.803
## PTHK         0.30524    0.02590  11.786
## SC           0.47147    0.11330   4.161
## TV           0.01956    0.11330   0.173
##
## Correlation of Fixed Effects:
##      (Intr) PTHK   SC
## PTHK -0.493
## SC   -0.503  0.025
## TV   -0.521  0.015 -0.002
```

If we ignore our clustering we get:

```
model2 <- lm(y~PTHK + SC + TV, data=smoking)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ PTHK + SC + TV, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5635 -0.9130 -0.0626  0.8981  4.2416
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73734    0.07866  22.088  < 2e-16 ***
## PTHK         0.32525    0.02589  12.561  < 2e-16 ***
```

```
## SC           0.47987    0.06529   7.350 3.15e-13 ***
## TV           0.04534    0.06518   0.696    0.487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.303 on 1596 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.112
## F-statistic: 68.21 on 3 and 1596 DF,  p-value: < 2.2e-16
```

What experiment does this model assume? While it might be tempting to want to use model 2 as the t values are bigger, this is disingenous as it assumes an experiment that we didn't actually perform. What experiment does `model2` assume conducted?

Going back to model 1. If we wanted to fit this model, we could use the likelihood from the joint distribution. To find this, let's consider the covariance between two students who not in the same school, nor the same class, in this case the covariance is:

Now consider the covariance between two students who are in the same school but not the same class:

Now two students who are in the same class:

Therefore, our entire covariance matrix looks like:

Once we have this we can find the MLEs maximizing:

Let's consider another example where we want to, perhaps, think of our random effects a little differently. Let's look at a dataset where 24 patients were randomly assigned to each of three treatment groups and compared on a measure of respiratory ability, FEV. The study observed FEV for a baseline measurement and then for each of 8 hours after a drug was administered.

The data look like:

```
FEV <- read.table("https://users.stat.ufl.edu/~aa/glm/data/FEV2.dat",header=TRUE)

FEV %>% head()
```

```
##   Obs patient base drug hour  fev
## 1   1       1 2.46    a    1 2.68
## 2   2       2 3.50    a    1 3.95
## 3   3       3 1.96    a    1 2.28
## 4   4       4 3.44    a    1 4.08
## 5   5       5 2.80    a    1 4.09
## 6   6       6 2.36    a    1 3.79
```

What are our observational units here? What are our experimental units?

Let's write out our random effects model how we did before with a term for patient i observed at hour j. Perhaps we have covariates for the measurement at baseline, the drug they were given, and how long the drug has been in their system.

What does this same about the covariance between patient 1 and hours 1 and 2?

What about hours 1 and 3?

For a given patient, we can write out the covariance matrix of u_i as:

What are some potential issues with this?

Another, perhaps, reasonable assumption would be that the relationship between patient 1 at hours 1 and 2 is stronger than the relationship of patient 1 at hours 1 and 3. That is, we may assume that the correlation structure for u_i to look like:

This is what is called an *autoregressive* correlation structure.

```
library(nlme)

ar_mod <- lme(fev~base+factor(drug)+hour,
              random=~1|patient,
              correlation=corAR1(form=~1|patient),
              data=FEV,
              method="ML")

summary(ar_mod)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: FEV
##       AIC      BIC    logLik
##   223.931 258.7798 -103.9655
##
## Random effects:
##  Formula: ~1 | patient
##         (Intercept)  Residual
## StdDev:   0.3938976 0.3337413
##
## Correlation Structure: AR(1)
##  Formula: ~1 | patient
##  Parameter estimate(s):
##     Phi
## 0.64471
## Fixed effects:  fev ~ base + factor(drug) + hour
##                    Value  Std.Error  DF   t-value p-value
## (Intercept)    1.0720258 0.28452329 503  3.767796  0.0002
## base           0.8918907 0.10012033  68  8.908188  0.0000
## factor(drug)c  0.2130917 0.12952249  68  1.645210  0.1045
## factor(drug)p -0.3138347 0.12953998  68 -2.422686  0.0181
## hour          -0.0690975 0.00768843 503 -8.987206  0.0000
##  Correlation:
##                (Intr) base   fctr(drg)c fctr(drg)p
## base           -0.939
## factor(drug)c -0.245  0.019
## factor(drug)p -0.251  0.025  0.500
## hour          -0.122  0.000  0.000      0.000
##
## Standardized Within-Group Residuals:
```

```
##          Min           Q1          Med           Q3          Max
## -3.448308129 -0.508151496 -0.004933778  0.499614068  2.423924086
##
## Number of Observations: 576
## Number of Groups: 72
```

We can compare this to a model with just the random intercept

```
library(nlme)

re_mod <- lme(fev~base+factor(drug)+hour,
              random=~1|patient,
              data=FEV,
              method="ML")

summary(re_mod)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: FEV
##        AIC     BIC    logLik
##   368.6692 399.162 -177.3346
##
## Random effects:
##  Formula: ~1 | patient
##         (Intercept)  Residual
## StdDev:   0.4393671 0.2714002
##
## Fixed effects:  fev ~ base + factor(drug) + hour
##                    Value  Std.Error  DF    t-value p-value
## (Intercept)    1.0492317 0.28523050 503   3.678540  0.0003
## base           0.9028516 0.10080988  68   8.955983  0.0000
## factor(drug)c  0.2258930 0.13041454  68   1.732115  0.0878
## factor(drug)p -0.2814907 0.13043215  68  -2.158139  0.0345
## hour          -0.0745734 0.00495693 503 -15.044284  0.0000
##  Correlation:
##               (Intr) base   fctr(drg)c fctr(drg)p
## base          -0.943
## factor(drug)c -0.246  0.019
## factor(drug)p -0.252  0.025  0.500
## hour          -0.078  0.000  0.000      0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.24932716 -0.53092128  0.02861887  0.55803625  2.57472933
##
## Number of Observations: 576
## Number of Groups: 72
```

Note that this model is also called a Compound Symmetry model. As we used the MLEs we can compute:

```
AIC(re_mod)
```

```
## [1] 368.6692
```

```
AIC(ar_mod)
```

```
## [1] 223.931
```

Apparently you CAN compute AIC for REML https://doi.org/10.1111/anzs.12254 , it's not entirely clear to me how as I haven't dug into it, but I would hesitate to compare AIC for a model I fit with REML to AIC for a model I fit with log-likelihood.

More here: https://stats.stackexchange.com/questions/131272/lme4-why-is-aic-no-longer-displayed-when-using-reml

I want to make one more change to the syllabus (I'm sorry!) I want to cover GLMMs first before we talk about Bayesian inference.