Topics for this lecture:

- *Nesting and crossing*

Associated reading: Course notes, '7 LMM nesting and crossing' chapter, Sections 1, 2, 3 (except 3.2.3); Hedeker, Ch. 13 (for hierarchical models)

1 Nested versus crossed factors

1.1 Definitions

- The term *nesting* can be applied to many things, including design structures, treatment structures, factors, data or models.

- Factors *A* and *B* are *crossed* if every level of *A* appears with every level of *B*. Such is the case for a 2-factor factorial treatment structure in a completely randomized design. Example: Myostatin experiment.

- Factors are *nested* when the levels of one factor occur with only one level of another factor. For example, consider a experiment designed to determine the effect of school and instructor on standardized test scores of kids in elementary schools, where each school has a unique set of teachers.

Examples:

Crossed (e.g., Myostatin data)

		Time		
		1	2	3
Group	A	x	x	x
	B	x	x	x

Nested (e.g., standardized test data)

		Teacher					
		1	2	3	4	5	6
School	A	x	x				
	B			x	x		
	C					x	x

- Notation for nesting:
  - On factors: for *B* nested within *A*, you may see *B* (*A*) (e.g., SAS uses this).
  - On indices: if *i* and *j* are indices for factors *A* and *B*, respectively, and *B* is nested within *A*, then an effect in the model may be denoted as $\beta_{j(i)}$.

1.2 Nesting and interaction

- In SAS, when you specify *B*(*A*), you are telling SAS that *B* is nested within *A*. Due to this, levels of *B* are unique within *A*; you have to consider the levels of *B* separate for each level of *A*, just like an interaction. The code below shows equivalency of *B*(*A*) and *B***A*.

- The data set below involves measurements taken within days (e.g., morning, noon, evening), for 3 different days. We would consider factors *A* and *B* crossed if the levels 'morning', 'noon' and 'evening' mean the same thing across days, and we would consider *B* nested within *A* if the levels of *B* could not be considered the same – e.g., if times of measurement varied across days. The data and partial output follows.

```
data time_and_day; input id day time y @@; datalines;
1 1 1 8 1 1 2 11 1 1 3 13 1 2 1 10 1 2 2 15 1 2 3 18 1 3 1 11
1 3 2 14 1 3 3 17 2 1 1 21 2 1 2 15 2 1 3 28 2 2 1 15 2 2 2 22
2 2 3 26 2 3 1 28 2 3 2 32 2 3 3 49 3 1 1 17 3 1 2 25 3 1 3 18
3 2 1 7 3 2 2 12 3 2 3 28 3 3 1 30 3 3 2 31 3 3 3 39
;
*crossed factors;
proc mixed data=time_and_day; class day time; model y=day time /
solution; random intercept / subject=id; run;
```

Solution for Fixed Effects

Effect	day	time	Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept			33.3704	4.5319	2	7.36	0.0179
day	1		-10.5556	2.6534	20	-3.98	0.0007
day	2		-10.8889	2.6534	20	-4.10	0.0006
day	3		0	.	.	.	.
time		1	-9.8889	2.6534	20	-3.73	0.0013
time		2	-6.5556	2.6534	20	-2.47	0.0226
time		3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
day	2	20	10.89	0.0006
time	2	20	7.19	0.0044

- Estimates for specific time*day combinations can be determined from the factor 'marginal' estimates (plus the intercept). 'Marginal'=main effect in this case.

```
*nested factors;
proc mixed data=time_and_day; class day time; model y=day time(day) /
solution; random intercept / subject=id; run;
```

Solution for Fixed Effects

Effect	day	time	Estimate	Standard Error	DF	t Value	Pr>|t|
Intercept			35.0000	5.1087	2	6.85	0.0206
day	1		-15.3333	4.8032	16	-3.19	0.0057
day	2		-11.0000	4.8032	16	-2.29	0.0359
day	3		0	.	.	.	.
time(day)	1	1	-4.3333	4.8032	16	-0.90	0.3803
time(day)	1	2	-2.6667	4.8032	16	-0.56	0.5864
time(day)	1	3	0	.	.	.	.
time(day)	2	1	-13.3333	4.8032	16	-2.78	0.0135
time(day)	2	2	-7.6667	4.8032	16	-1.60	0.1300
time(day)	2	3	0	.	.	.	.
time(day)	3	1	-12.0000	4.8032	16	-2.50	0.0238
time(day)	3	2	-9.3333	4.8032	16	-1.94	0.0698
time(day)	3	3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
day	2	16	9.97	0.0015
time(day)	6	16	2.58	0.0608

- Using time*day in place of time(day) will yield the same SAS output. In this case, each time*day combination has a unique estimate such that ‘marginal’ factor estimates cannot be used to obtain them.

- Determining whether factors are crossed or nested should not be based on model fit; rather, it should be based on the design of the experiment or study. But in some cases, there may be a fine line between whether factors are nested or crossed.

- In the example above, we said that factors are crossed if the levels of time meant the same thing across days. The question is, how close in actual time do the measurements need to be to be considered ‘the same’?

- For the crossed design, you could also include *time*day* in the model. The test for *time*day* will not be the same as for *time*(*day*) (or *time*day*) in the nested models above. But, the LSMEANS estimates for *time*day* combinations will be the same between the ‘full’ and nested models.

1.3 Nested subjects

- Before, we saw the use of the nested notation for subjects within groups. This relates to a *parallel experiment* in which subjects are randomly assigned to groups, and remain in those groups throughout the experiment. This differs from a crossover experiment. (These will be discussed more in the next section.)

- For a parallel study/experiment, the use of ID(GROUP) is important in SAS PROC MIXED if the same IDs are used in different groups. When using PROC GLM for repeated measures ANOVA, the use of this nested variable is important whether or not the IDs are repeated.

2 Nested versus crossed random effects

- Nesting and crossing apply to random effects as well as fixed effects. Consider the standardized test data. If the given teachers and schools form a (random) sample from this population, then we can model school as one random effect term, and teacher within school [or teacher(school)] as another. The random statement in the PROC MIXED code would be:

```
RANDOM school teacher(school);
```

- For the psychological scoring data, the first approach was to model subject and rater as random effects. Since each rater scored each subject, then the random effects were crossed:

```
RANDOM subject rater;
```

3 Hierarchical Linear Models

3.1 Two-level models

- A mixed model with subjects that are observed over time can be considered a hierarchical model, where the level 1 model consists of the responses over time for subjects (Y_{ij}), and the level 2 model consists of subject-specific random effect terms. Specifically, considering the random intercept model, we have

$$\text{Level 1: } Y_{ij} = b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij} \qquad \text{Level 2: } \begin{aligned} b_{0i} &= \beta_0 + u_{0i} \\ b_{1i} &= \beta_1 \end{aligned}$$

- If we desire to have random slopes (for time) in the 2-level model as well as random intercepts, then we have

$$\text{Level 1: } Y_{ij} = b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij} \qquad \text{Level 2: } \begin{aligned} b_{0i} &= \beta_0 + u_{0i} \\ b_{1i} &= \beta_1 + u_{1i} \end{aligned}$$

- Other types of 2-level models may not involve repeated measures.

- E.g. health care costs for individuals with different health insurance providers at one time.

  - The level 1 model would involve the subjects (Y_{ij}, i denotes provider, j denotes subject).

  - The level 2 model would involve the providers.

  - Here, subjects are nested within health-care providers, and this could be modeled using a random intercept term for provider; subject variability will be accounted for with the residual error term. The model could be expressed as:

    Level 1: $Y_{ij} = b_{0i} + \varepsilon_{ij}$ (costs for individuals)
    Level 2: $b_{0i} = \beta_0 + u_{0i}$ (random insurance provider effects)

- For the standardized test data, if there is one response per teacher, the level 1 model would involve subjects within schools, and the level 2 model would involve the schools.

- *Applying level terminology to the units.* For the standardized test data, the level 1 units involve teachers and level 2 units involve schools. For the 3-level HLM with insurance data, the level 1 units involve the repeated measures over time, the level 2 units would be subjects, and level 3 units would be the providers.

- *Distinguishing 'nesting' concepts for factors and data*. Although we may say that repeated measures are 'nested' within subjects, it is possible that the factor associated with the repeated measures (e.g., time) is crossed with subjects, if there are specific measurement times that every subject is observed at. Thus, we need to distinguish nesting/crossing of data versus nesting/crossing of factors.

3.2 Three-level models

3.2.1 Examples

- For the insurance provider data, suppose that we have repeated cost measures for subjects over time. A 3-level hierarchical model could be developed for these data, where health care costs are denoted as Y_{ijk}, where i=provider, j=subject, k=time.

    - The level 1 model involves repeated measures within subjects, the level 2 model involves the subjects themselves, and the level 3 model involves the health care providers.
    - This model may include two random intercept terms, one for provider, and one for subject within provider.
    - The lowest level (i.e., level 1) of a hierarchical model involves the smallest unit of measurement while the highest level involves the largest unit of measurement.

- Example 1: Subjects that are obtained from different sites are monitored over time (perhaps after being assigned to a treatment group).
    - This is often called a multi-site experiment or study, which is done because it is too difficult for one site alone to get all of the subjects for the experiment.
    - For example, the sites may be medical and research centers across the U.S.
    - Here, leve1 1 involves the measures on subjects over time. Level 2 involves the subjects that are nested within sites, and level 3 involves the sites. (Of course, measures on subjects are nested within subjects!)

- Example 2: Children are recruited from schools for a health study.
    - Children are obtained by selecting them from classrooms within schools within the study area.
    - The level 1 units involve the children's measurements; the level 2 units are classrooms and the level 3 units are the schools.
    - Note that children are nested within classrooms and classrooms are nested within schools.
    - We could extend this to 4-level data if repeated measures were taken on the children, where the repeated measures within subjects would become the level 1 data.

3.2.2 Case study: Kunsberg study

- An EPA-funded study at NJH has involved kids from the Kunsberg School (at NJH) that have moderate to severe asthma. One of the primary goals has been to determine how the health of children is associated with air pollution on a day-to-day basis. A number of variables have been collected on subjects (demographic, behavioral, and biological) as well as the environment (air pollution, meteorology) for this ongoing study.

- One of the difficulties with the data is that siblings are often involved in the study. For the 2001-02 school year, there were 48 subjects from 40 families; in 2002-03, there were 57 children from 52 families. In both years, there were never more than 2 kids involved from the same family.

- One of the key assumptions in our modeling is the 'independent subjects' assumption. In medical research, this assumption is often ignored. People often fret more about the normality assumption and ignore the independence assumption, the latter of which can be much more problematic.

- One of the reasons why the independence assumption is violated in medical research is because random sampling is usually unfeasible. Participants are often self selected. We can account for the possible dependency between siblings by including appropriate random terms. We can fit the model with and without the random terms to determine the impact of the dependency.

- This can be considered 3-level data, where
  - families are the level 3 unit
  - children within families are the level 2 unit
  - repeated measures within kids are the level 1 unit.

- For this analysis, LTE_4 (a specific biomarker in the body that has been shown to be associated with inflammation) was fit on the natural log scale as a function of date (linear time trend), cold (1=yes, 0=no – a time varying variable), temperature, pressure and humidity. The goal is to better understand how LTE_4 relates to different variables over time. [Side note: it has already been shown that LTE_4 is related to air pollution. See Rabinovitch, Strand, Gelfand, 2006 AJRCCM article.]

The model:

$$Y_{hij} = \beta_0 + \beta_1 date_j + \beta_2 cold_{ij} + \beta_3 temp_j + \beta_4 pressure_j + \beta_5 humidity_j + b_h + b_{i(h)} + \varepsilon_{ij}$$

$b_h \sim N(0, \sigma_F^2)$

$b_{i(h)} \sim N(0, \sigma_S^2)$

$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

We could write Y_{ij} on the left since an observation is uniquely identified by subject and date. Here I include it for sake of completeness.

h for family, *i* for subject, *j* for time.

$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, where $\mathbf{R}_i$ has the AR(1) form.

SAS code:

```
PROC MIXED DATA = newpoll METHOD = ML covtest;
  CLASS id family;
  MODEL loglte4 = date cold temp pressure humidity / s;
  RANDOM family id(family) / solution;
  REPEATED / SUBJECT = id(family) TYPE = ar(1); RUN;
```

If you wanted to include random slopes for subjects, you would need to add another random statement. For example, for random time slopes, you could include: **RANDOM date / subject=id;** . The **G** matrix will be defined for the complete data, due to the other random statement already in the model. For this particular data set the likelihood could not be solved when the 2nd random statement was included, but I did try a similar approach with another data set and it appeared to work. In general, you may need ample data to fit such a model.

Since IDs are unique study wide for these data, we could actually simplify ID(FAMILY) to ID in the code (2 places) and get the same results. However, one advantage of keeping the notation is to have a reminder of how the data is nested.

Abbreviated output:

The Mixed Procedure

Class Level Information

Class	Levels
id	53
family	48

Dimensions

Covariance Parameters	4
Columns in X	6
Columns in Z	101
Subjects	1
Max Obs Per Subject	12540

Number of Observations Used 449

The 1 'subject' is induced by the way we wrote the RANDOM statement above. This will also occur for a statement like: RANDOM ID;

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
family		0.05119	0.04229	1.21	0.1131
id(family)		0.06316	0.04249	1.49	0.0686
AR(1)	id(family)	0.4889	0.05857	8.35	<.0001
Residual		0.1522	0.01520	10.01	<.0001

Fit Statistics

-2 Log Likelihood	425.2	AIC (smaller is better)	445.2
AICC (smaller is better)	445.7	BIC (smaller is better)	463.9

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-15.6671	10.9718	5	-1.43	0.2127
date	0.001192	0.000609	393	1.96	0.0510
cold	0.1184	0.05331	393	2.22	0.0270
temp	0.006051	0.003065	393	1.97	0.0491
pressure	0.001532	0.005504	393	0.28	0.7809
humidity	0.002772	0.001609	393	1.72	0.0857

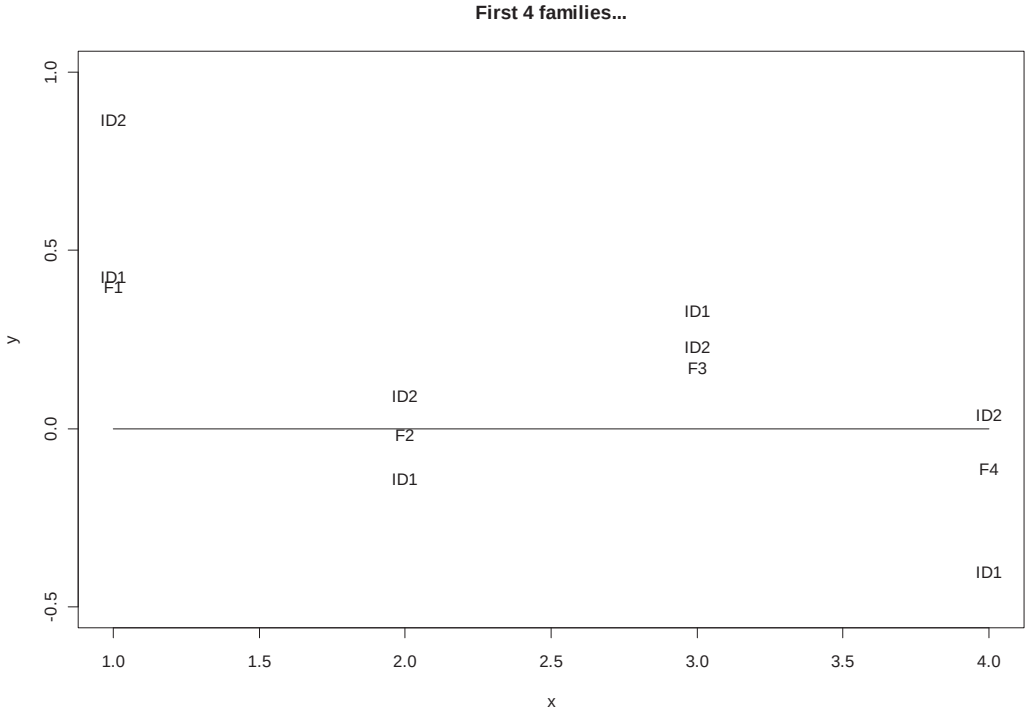
Solution for Random Effects

Effect	NewID	family	Estimate	Std Err Pred	DF	t Value	Pr > |t|
family		1	0.4030	0.1595	393	2.53	0.0119
family		2	-0.01296	0.1584	393	-0.08	0.9348
family		3	0.1757	0.1595	393	1.10	0.2715
family		4	-0.1097	0.1600	393	-0.69	0.4933
family		5	-0.1699	0.1581	393	-1.08	0.2830
family		6	-0.09309	0.1826	393	-0.51	0.6105
family		7	-0.05686	0.1833	393	-0.31	0.7566
family		8	0.1213	0.1966	393	0.62	0.5375
family		9	-0.04490	0.1821	393	-0.25	0.8053
. . .							
family		51	-0.00858	0.1821	393	-0.05	0.9625
family		52	0.1121	0.1878	393	0.60	0.5508
id(family)	224	1	0.02866	0.1795	393	0.16	0.8732
id(family)	345	1	0.4686	0.1797	393	2.61	0.0094
id(family)	139	2	-0.1233	0.1778	393	-0.69	0.4883
id(family)	402	2	0.1073	0.1785	393	0.60	0.5481
id(family)	221	3	0.1578	0.1783	393	0.89	0.3765
id(family)	222	3	0.05892	0.1812	393	0.33	0.7453
id(family)	208	4	-0.2870	0.1780	393	-1.61	0.1077
id(family)	408	4	0.1516	0.1824	393	0.83	0.4065
id(family)	346	5	0.07689	0.1768	393	0.43	0.6639
id(family)	412	5	-0.2866	0.1774	393	-1.62	0.1069
id(family)	206	6	-0.1149	0.1897	393	-0.61	0.5453
. . .							
id(family)	233	51	-0.01058	0.1889	393	-0.06	0.9554
id(family)	415	52	0.1383	0.1973	393	0.70	0.4835

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
date	1	393	3.83	0.0510
cold	1	393	4.93	0.0270
temp	1	393	3.90	0.0491
pressure	1	393	0.08	0.7809
humidity	1	393	2.97	0.0857

Consider the graph of intercepts…



- Note that family estimates are not required to be in the center of the id(family) estimates for that given family.

- In particular, family estimates are shifted downward towards 0 relative to the midpoint of the id(family) estimates when both id(family) estimates are positive, and family estimates are shifted upward towards 0 when both id(family) estimates are negative. I.e., greater shrinkage appears to exist for the family estimates compared with the id(family) estimates.

- Note: you can get specific estimates that combine both fixed and random effect estimates using the ESTIMATE statement. To learn more, see Littell, et al.: SAS System for Mixed Models, Chapter 6.

Model without family:

```
PROC MIXED DATA = newpoll METHOD = ML covtest;
  CLASS id family;
  MODEL loglte4 = date cold temp pressure humidity / s;
  random id(family) / solution;
  REPEATED / SUBJecT = id(family) TYPE = ar(1); RUN;
```

Abbreviated output:

Covariance Parameters 3

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
id(family)		0.1162	0.03080	3.77	<.0001
AR(1)	id(family)	0.4894	0.05857	8.36	<.0001
Residual		0.1523	0.01523	10.00	<.0001

Fit Statistics

-2 Log Likelihood	426.6
AIC (smaller is better)	444.6
AICC (smaller is better)	445.0
BIC (smaller is better)	462.3

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-15.6278	10.9896	50	-1.42	0.1612
date	0.001177	0.000610	393	1.93	0.0545
cold	0.1176	0.05331	393	2.21	0.0280
temp	0.006010	0.003067	393	1.96	0.0507
pressure	0.001839	0.005506	393	0.33	0.7386
humidity	0.002804	0.001610	393	1.74	0.0825

Solution for Random Effects

Effect	NewID	family	Estimate	Std Err Pred	DF	t Value	Pr > |t|
id(family)	224	1	0.3464	0.1628	393	2.13	0.0340
id(family)	345	1	0.8637	0.1634	393	5.28	<.0001
id(family)	139	2	-0.1593	0.1578	393	-1.01	0.3133
id(family)	402	2	0.1122	0.1604	393	0.70	0.4846
id(family)	221	3	0.3191	0.1586	393	2.01	0.0450
id(family)	222	3	0.2011	0.1681	393	1.20	0.2322
id(family)	208	4	-0.4276	0.1579	393	-2.71	0.0071
id(family)	408	4	0.09355	0.1717	393	0.54	0.5863
id(family)	346	5	-0.05254	0.1560	393	-0.34	0.7365
id(family)	412	5	-0.4758	0.1577	393	-3.02	0.0027
id(family)	206	6	-0.2135	0.1592	393	-1.34	0.1808
. . .							
id(family)	233	51	-0.02426	0.1560	393	-0.16	0.8765
id(family)	415	52	0.2468	0.1871	393	1.32	0.1879

Synopsis:

- The estimated variance for families and subjects within families have roughly the same order of magnitude. By looking at the 2nd output, we can see that the ID(FAMILY) variance then absorbs all of the variance that was formerly attributed to the random family effect.

- There is not a dramatic difference in fixed effect estimates for the 2 approaches. The AICs are comparable, but a bit better without the family effect. It is likely that the level of dependency due to siblings is not that great since there are few siblings involved.

- Even though removal of the random term for family (and hence not accounting for the sibling effect) did not change estimates drastically, I would generally warn against ignoring dependent data! It would be prudent to at least check for it…

- If relevant, it is also possible to examine nested effects for random slopes.
  - For example, we previously discussed the relationship between personal exposure to ambient $PM_{2.5}$ and actual ambient $PM_{2.5}$.
  - Subjects have different slopes, based somewhat on the type of housing they live in. Thus, it may be expected that subjects that live in the same house (i.e., those within the same family) will tend to have the same slopes.
  - In this case, we may try fitting two random slopes for ambient $PM_{2.5}$ to account for this dependency, one for SUBJECT=FAMILY and the other for SUBJECT=ID(FAMILY).

- For further details on hierarchical models fit to multi-level data (with an emphasis on the use of SAS PROC MIXED), see Littell et al., *SAS System for Mixed Models*.