Lecture 1
Introduction to Multi-level Models

Course Website:
http://www.biostat.jhsph.edu/~ejohnson/multilevel.htm

All lecture materials extracted and further developed from the Multilevel Model course taught by Francesca Dominici:
http://www.biostat.jhsph.edu/~fdominic/teaching/bio656/ml.html

Statistical Background on MLMs

- Main Ideas
- Accounting for Within-Cluster Associations
- Marginal & Conditional Models
- A Simple Example
- Key MLM components
The Main Idea...

Multi-level Models – Main Idea

• Biological, psychological and social processes that influence health occur at many **levels**:  
  – Cell  
  – Organ  
  – Person  
  – Family  
  – Neighborhood  
  – City  
  – Society  

• An analysis of risk factors should consider:  
  – Each of these levels  
  – Their interactions
**Example: Alcohol Abuse**

| **Level:** | Neurochemistry |
| 1. Cell: | Ability to metabolize ethanol |
| 2. Organ: | Genetic susceptibility to addiction |
| 3. Person: | Alcohol abuse in the home |
| 4. Family: | Availability of bars |
| 5. Neighborhood: | Regulations; organizations; social norms |

**Example: Alcohol Abuse; Interactions between Levels**

<table>
<thead>
<tr>
<th>Level</th>
<th>Interactions between Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Availability of bars <em>and</em></td>
</tr>
<tr>
<td>6</td>
<td>State laws about drunk driving <em>and</em></td>
</tr>
<tr>
<td>4</td>
<td>Alcohol abuse in the family <em>and</em></td>
</tr>
<tr>
<td>2</td>
<td>Person’s ability to metabolize ethanol</td>
</tr>
<tr>
<td>3</td>
<td>Genetic predisposition to addiction <em>and</em></td>
</tr>
<tr>
<td>4</td>
<td>Household environment</td>
</tr>
<tr>
<td>6</td>
<td>State regulations about intoxication <em>and</em></td>
</tr>
<tr>
<td>3</td>
<td>Job requirements</td>
</tr>
</tbody>
</table>
Notation:
Population

Person: $sijk$
Outcome: $Y_{sijk}$
Predictors: $X_{sijk}$

State: $s=1,\ldots,S$

Neighborhood:
$i=1,\ldots,I_s$

Family: $j=1,\ldots,J_{si}$

Person: $k=1,\ldots,K_{sij}$

Notation (cont.)

- $(y_{sijk}, x_{sijk})$ are (response, predictors) for
  - person $k = 1, \ldots, K_{sij}$ in
  - family $j = 1, \ldots, J_{si}$ in
  - neighborhood $i = 1, \ldots, I_s$ in
  - state $s = 1, \ldots, S$
- $\mu_{sijk} = E(y_{sijk} | x_{sijk})$
Multi-level Models: Idea

**Level:** Predictor Variables

1. Person’s Income
2. Family Income
3. Percent poverty in neighborhood
4. State support of the poor

Response: Alcohol Abuse

**A Rose is a Rose is a…**

- Multi-level model
- Random effects model
- Mixed model
- Random coefficient model
- Hierarchical model
- Meta-analysis (in some cases)

Many names for similar models, analyses, and goals.
Digression on Statistical Models

• A statistical model is an approximation to reality
• There is not a “correct” model;
  – ( forget the holy grail )
• A model is a tool for asking a scientific question;
  – ( screw-driver vs. sludge-hammer )
• A useful model combines the data with prior information to address the question of interest.
• Many models are better than one.

Generalized Linear Models (GLMs)

g(μ) = β₀ + β₁*X₁ + … + βₚ*Xₚ

\( \mu = E(Y|X) = \text{mean} \)

<table>
<thead>
<tr>
<th>Model</th>
<th>Response</th>
<th>g(μ)</th>
<th>Distribution</th>
<th>Coef Interp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>Continuous (ounces)</td>
<td>μ</td>
<td>Gaussian</td>
<td>Change in avg(Y) per unit change in X</td>
</tr>
<tr>
<td>Logistic</td>
<td>Binary (disease)</td>
<td>( \log \left( \frac{\mu}{1-\mu} \right) )</td>
<td>Binomial</td>
<td>Log Odds Ratio</td>
</tr>
<tr>
<td>Log-linear</td>
<td>Count/Times to events</td>
<td>log(μ)</td>
<td>Poisson</td>
<td>Log Relative Risk</td>
</tr>
</tbody>
</table>
Generalized Linear Models (GLMs)

\[ g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \]

**Example: Age & Gender**

**Gaussian – Linear:**

\[ \text{E}(y) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} \]

\( \beta_1 \) = Change in Average Response per 1 unit increase in Age, Comparing people of the SAME GENDER.

** WHY?**

Since: \( \text{E}(y|\text{Age+1,Gender}) = \beta_0 + \beta_1 (\text{Age+1}) + \beta_2 \text{Gender} \)

And: \( \text{E}(y|\text{Age ,Gender}) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} \)

\[ \Delta \text{E}(y) = \beta_1 \]

**Generalized Linear Models (GLMs)**

\[ g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \]

**Example: Age & Gender**

**Binary – Logistic:**

\[ \log\{\text{odds}(Y)\} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} \]

\( \beta_1 \) = log-OR of “+ Response” for a 1 unit increase in Age, Comparing people of the SAME GENDER.

** WHY?**

Since: \( \log\{\text{odds}(y|\text{Age+1,Gender})\} = \beta_0 + \beta_1 (\text{Age+1}) + \beta_2 \text{Gender} \)

And: \( \log\{\text{odds}(y|\text{Age ,Gender})\} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} \)

\[ \Delta \log\{-\text{Odds}\} = \beta_1 \]

\[ \log\{-\text{OR}\} = \beta_1 \]
Generalized Linear Models (GLMs)

\[ g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \]

**Example: Age & Gender**

Counts – Log-linear: \[ \log\{E(Y)\} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} \]

\[ \beta_1 = \log\text{-RR for a 1 unit increase in Age,} \]
\[ \text{Comparing people of the SAME GENDER.} \]

**WHY?**

*Self-Check: Verify Tonight*

---

“Quiz”: Most Important Assumptions of Regression Analysis?

A. Data follow normal distribution

B. **All the key covariates are included in the model**

C. Xs are fixed and known

D. Responses are independent
Non-independent responses
(Within-Cluster Correlation)

• Fact: two responses from the same family tend to be more like one another than two observations from different families

• Fact: two observations from the same neighborhood tend to be more like one another than two observations from different neighborhoods

• Why?

Why? (Family Wealth Example)

Great-Grandparents

Grandparents

Parents

You

Great-Grandparents

Grandparents

Parents

You
Key Components of Multi-level Models

• Specification of predictor variables from multiple levels (Fixed Effects)
  – Variables to include
  – Key interactions

• Specification of correlation among responses from same clusters (Random Effects)

• Choices must be driven by scientific understanding, the research question and empirical evidence.

Correlated Data…
(within-cluster associations)
Multi-level analyses

• Multi-level analyses of social/behavioral phenomena: an important idea

• Multi-level models involve predictors from multi-levels and their interactions

• They must account for associations among observations within clusters (levels) to make efficient and valid inferences.

Regression with Correlated Data

Must take account of correlation to:

• Obtain valid inferences
  – standard errors
  – confidence intervals

• Make efficient inferences
Logistic Regression Example: Cross-over trial

- Response: 1-normal; 0- alcohol dependence
- Predictors: period (x₁); treatment group (x₂)
- Two observations per person (cluster)
- Parameter of interest: log odds ratio of alcohol dependence: placebo vs. treatment

Mean Model:  \[ \log(\text{odds(AD)}) = \beta_0 + \beta_1 \text{Period} + \beta_2 \text{Placebo} \]

Results: estimate (standard error)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model</th>
<th></th>
<th>Account for correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ( (\beta_0) )</td>
<td><a href="0.32">0.66</a></td>
<td><a href="0.29">0.67</a></td>
<td></td>
</tr>
<tr>
<td>Period ( (\beta_1) )</td>
<td>-0.27 (0.38)</td>
<td>-0.30 (0.23)</td>
<td></td>
</tr>
<tr>
<td>Placebo ( (\beta_2) )</td>
<td>0.56 (0.38)</td>
<td>0.57 (0.23)</td>
<td></td>
</tr>
</tbody>
</table>

Similar Estimates, **WRONG** Standard Errors (& Inferences) for OLR
Simulated Data: Non-Clustered

Simulated Data: Clustered
Within-Cluster Correlation

• Correlation of two observations from same cluster =

\[
\frac{\text{Tot Var} - \text{Var Within}}{\text{Tot Var}}
\]

• Non-Clustered = (9.8-9.8) / 9.8 = 0
• Clustered = (9.8-3.2) / 9.8 = 0.67

Models for Clustered Data

• Models are tools for inference
• Choice of model determined by scientific question
• Scientific Target for inference?
  – Marginal mean:
    • Average response across the population
  – Conditional mean:
    • Given other responses in the cluster(s)
    • Given unobserved random effects
• We will deal mainly with conditional models (but we'll mention some important differences)
Marginal vs Conditional Models…

Marginal Models

• Focus is on the “mean model”: E(Y|X)
• Group comparisons are of main interest, i.e. neighborhoods with high alcohol use vs. neighborhoods with low alcohol use
• Within-cluster associations are accounted for to correct standard errors, but are not of main interest.

\[
\log\{ \text{odds(AD)} \} = \beta_0 + \beta_1 \text{Period} + \beta_2 \text{Placebo}
\]
Marginal Model Interpretations

• log{ odds(AD) } = \beta_0 + \beta_1\text{Period} + \beta_2\text{Placebo}
  = 0.67 + (-0.30)\text{Period} + (0.57)\text{Placebo}

TRT Effect: (placebo vs. trt)
OR = \exp(0.57) = 1.77, 95% CI (1.12, 2.80)

\text{Risk of Alcohol Dependence is almost twice as high on placebo, regardless of, (adjusting for), time period}

WHY?
Since: log{odds(AD|\text{Period, placebo})} = \beta_0 + \beta_1\text{Period} + \beta_2
And: log{odds(AD|\text{Period, trt})} = \beta_0 + \beta_1\text{Period}

\Delta \text{log-Odds} = \beta_2
\text{OR} = \exp(\beta_2)

Random Effects Models

• Conditional on unobserved latent variables or “random effects”
  – Alcohol use within a family is related because family members share an unobserved “family effect”: common genes, diets, family culture and other unmeasured factors
  – Repeated observations within a neighborhood are correlated because neighbors share: common traditions, access to services, stress levels,…
  – log{ odds(AD) } = b_1 + \beta_0 + \beta_1\text{Period} + \beta_2\text{Placebo}
Random Effects Model Interpretations

WHY?
Since: \[ \log(\text{odds}(\text{AD}_i|\text{Period, Placebo, } b_i)) = \beta_0 + \beta_1\text{Period} + \beta_2 + b_i \]
And: \[ \log(\text{odds}(\text{AD}_i|\text{Period, TRT, } b_i)) = \beta_0 + \beta_1\text{Period} + b_i \]
\[ \Delta \log\text{-Odds} = \beta_2 \]
\[ \text{OR} = \exp(\beta_2) \]

• In order to make comparisons we must keep the subject-specific latent effect \(b_i\) the same.
• In a Cross-Over trial we have outcome data for each subject on both placebo & treatment
• In other study designs we may not.

Marginal vs. Random Effects Models

• For **linear models**, regression coefficients in random effects models and marginal models are identical:
  \[ \text{average of linear function} = \text{linear function of average} \]
• For **non-linear models**, (logistic, log-linear,…) coefficients have different meanings/values, and address different questions
  - Marginal models -> **population-average** parameters
  - Random effects models -> **cluster-specific** parameters
### Comparison of Marginal and Random Effect Logistic Regressions

- Regression coefficients in the random effects model are roughly 3.3 times as large

  - **Marginal:** population odds (prevalence with/prevalence without) of AD is \( \exp(0.57) = 1.8 \) greater for placebo than on active drug;
    *population-average parameter*

  - **Random Effects:** a person’s odds of AD is \( \exp(1.8) = 6.0 \) times greater on placebo than on active drug;
    *cluster-specific, here person-specific, parameter*

Which model is better? They ask different questions.
Refresher: Forests & Trees

Multi-Level Models:
– Explanatory variables from multiple levels
  • i.e. person, family, n'bhd, state, …
  • Interactions
– Take account of correlation among responses from same clusters:
  • i.e. observations on the same person, family,…
  • Marginal: GEE, MMM
  • Conditional: RE, GLMM

Remainder of the course will focus on these.

Key Points

• “Multi-level” Models:
  – Have covariates from many levels and their interactions
  – Acknowledge correlation among observations from within a level (cluster)
• Random effect MLMs condition on unobserved “latent variables” to account for the correlation
• Assumptions about the latent variables determine the nature of the within cluster correlations
• Information can be borrowed across clusters (levels) to improve individual estimates
Examples of two-level data

• Studies of health services: assessment of quality of care are often obtained from patients that are clustered within hospitals. Patients are level 1 data and hospitals are level 2 data.

• In developmental toxicity studies: pregnant mice (dams) are assigned to increased doses of a chemical and examined for evidence of malformations (a binary response). Data collected in developmental toxicity studies are clustered. Observations on the fetuses (level 1 units) nested within dams/litters (level 2 data)

• The “level” signifies the position of a unit of observation within the hierarchy

Examples of three-level data

• Observations might be obtained in patients nested within clinics, that in turn, are nested within different regions of the country.

• Observations are obtained on children (level 1) nested within classrooms (level 2), nested within schools (level 3).
Why use marginal model when I can use a multi-level model?

- Public health problems: what is the impact of intervention/exposure on the population?
  - Most translation into policy makes sense at the population level

- Clinicians may be more interested in subject specific or hospital unit level analyses
  - What impact does a policy shift within the hospital have on patient outcomes or unit level outcomes?

Why use marginal model when I can use a multi-level model?

- Your study design may induce a correlation structure that you are not interested in
  - Sampling individuals within neighborhoods or households
  - Outcome: population mortality
  - Marginal model allows you to adjust inferences for the correlation while focusing attention on the model for mortality

- Dose-response or growth-curve
  - Here we are specifically interested in an individual trajectory
  - And also having an estimate of how the individual trajectories vary across individuals is informative.
### Additional Points: Marginal Model

- We focus attention on the population level associations in the data and we try to model these best we can (mean model)
- We acknowledge that there is correlation and adjust for this in our statistical inferences.
- These methods (GEE) are robust to misspecification of the correlation
- We are obtaining estimates of the target of interest and valid inferences even when we get the form of the correlation structure wrong.

### Multi-level Models

- Suppose you have hospital level summaries of patient outcomes
  - The fixed effect portion of your model suggests that these outcomes may differ by whether the hospital is teaching/non-teaching or urban/rural
  - The hospital level random effect represents variability across hospitals in the summary measures of patient outcomes; this measure of variability may be of interest
  - Additional interest lies in how large the hospital level variability is relative to a measure of total variability; what fraction of variability is attributable to hospital differences?
Additional considerations:

- Interpretations in the multi-level models can be tricky!

- Think about interpretation of gender in a random effects model:
  - \( E(Y|\text{gender,bi}) = b_0 + b_1\text{gender} + bi \)
  - Interpretation of \( b_1 \):
    Among persons with similar unobserved latent effect \( bi \), the difference in average \( Y \) if those same people had been males instead of females

  - Imagine the counter-factual world….does it make sense?

Comparison of Estimates:
Linear Model and Non-linear model

- A hypothetical cross-over trial
  - \( N = 15 \) participants
  - 2 periods
  - treatment vs placebo

- Two outcomes of interest
  - Continuous response: say alcohol consumption \((Y)\)
  - Binary response: say alcohol dependence \((AD)\)
### Linear model

\[ E(Y|\text{Period, Treatment}) = b_0 + b_1\text{Period} + b_2\text{Treatment} \]

<table>
<thead>
<tr>
<th></th>
<th>Ordinary Least Squares</th>
<th>GEE (Indep)</th>
<th>GEE (Exchange)</th>
<th>Random subject effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (b0)</td>
<td>15.2 (1.22)</td>
<td>15.2 (1.16)</td>
<td>15.2 (1.07)</td>
<td>15.2 (1.13)</td>
</tr>
<tr>
<td>Period (b1)</td>
<td>2.57 (1.38)</td>
<td>2.57 (1.31)</td>
<td>2.57 (1.01)</td>
<td>2.57 (1.08)</td>
</tr>
<tr>
<td>Treatment (b2)</td>
<td>-0.43 (1.38)</td>
<td>-0.43 (1.31)</td>
<td>-0.43 (1.01)</td>
<td>-0.43 (1.08)</td>
</tr>
</tbody>
</table>

SAME estimates . . . DIFFERENT standard errors . . .

### Non-Linear model

\[ \log(\text{Odds(AD}|\text{Period, Treatment})) = b_0 + b_1\text{Period} + b_2\text{Treatment} \]

<table>
<thead>
<tr>
<th></th>
<th>Ordinary Logistic Regression</th>
<th>GEE (Indep)</th>
<th>GEE (Exchange)</th>
<th>Random subject effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (b0)</td>
<td>-1.14 (0.75)</td>
<td>-1.14 (0.75)</td>
<td>-1.11 (0.83)</td>
<td>-1.14 (0.75)</td>
</tr>
<tr>
<td>Period (b1)</td>
<td>0.79 (0.83)</td>
<td>0.79 (0.83)</td>
<td>0.76 (1.02)</td>
<td>0.79 (0.83)</td>
</tr>
<tr>
<td>Treatment (b2)</td>
<td>1.82 (0.83)</td>
<td>1.82 (0.83)</td>
<td>1.80 (1.03)</td>
<td>1.82 (0.83)</td>
</tr>
</tbody>
</table>

SAME estimates and standard errors

Estimates and standard errors change (a little)
What happened in the GEE models?

- In non-linear models (binary, count, etc), the mean of the outcome is linked to the variance of outcome:
  - $X \sim \text{Binomial}, \text{mean } p, \text{variance } p(1-p)$
  - $X \sim \text{Poisson}, \text{mean } \lambda, \text{variance } \lambda$

- When we change the structure of the correlation/variance, we change the estimation of the mean too!

- The target of estimation is the same and our estimates are unbiased.

Why similarity between GEE and random effects here?

- No association in AD within person
- Little variability across persons
- Odds ratio of exposure across persons $\sim 1$

```
tab AD0 AD1
<table>
<thead>
<tr>
<th></th>
<th>1 AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 AD</td>
<td>0</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
</tr>
</tbody>
</table>
```