

# Lecture 1

## Introduction to Multi-level Models

Course web site

<http://www.biostat.jhsph.edu/~fdominic/teaching/bio656/ml.html>

1

## Statistical Background on MLMs

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

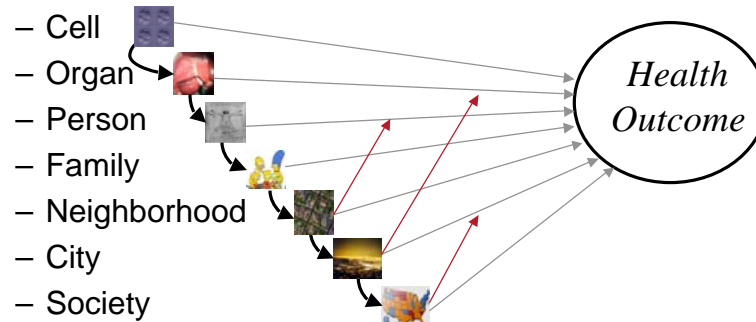
2

## The Main Idea...

3

### Multi-level Models – Main Idea

- Biological, psychological and social processes that influence health occur at many **levels**:



- An analysis of risk factors should consider:
  - Each of these levels
  - **Their interactions**

4

## Example: Alcohol Abuse

Level:

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

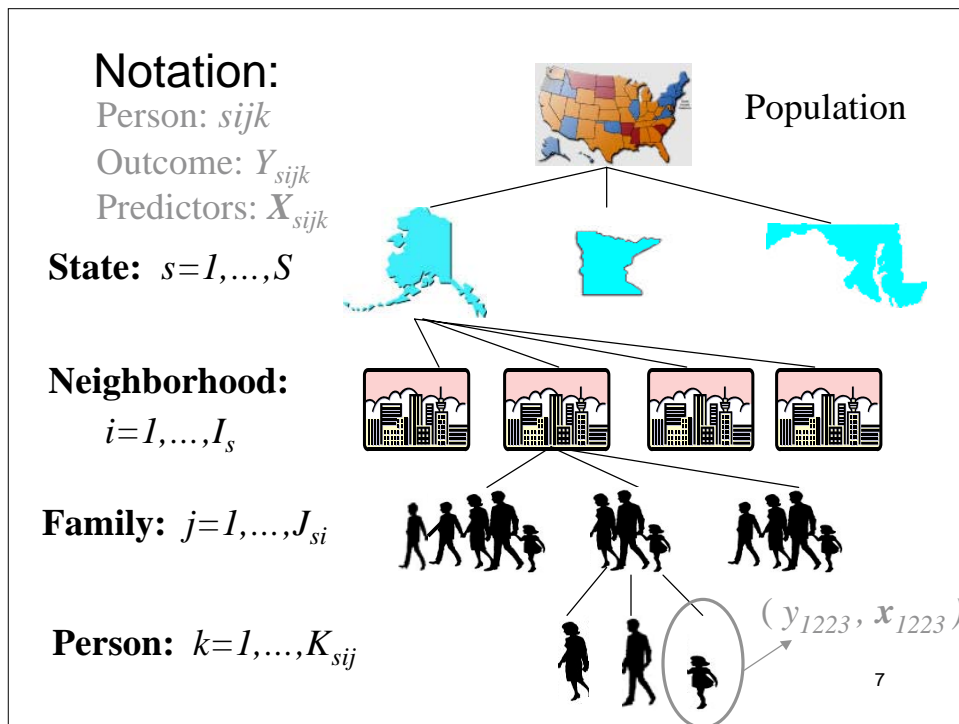
5

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

Level:

- |   |   |   |
|---|---|---|
| 5 | } | Availability of bars <i>and</i>                 |
| 6 |   | State laws about drunk driving                  |
| 4 | } | Alcohol abuse in the family <i>and</i>          |
| 2 |   | Person's ability to metabolize ethanol          |
| 3 | } | Genetic predisposition to addiction <i>and</i>  |
| 4 |   | Household environment                           |
| 6 | } | State regulations about intoxication <i>and</i> |
| 3 |   | Job requirements                                |

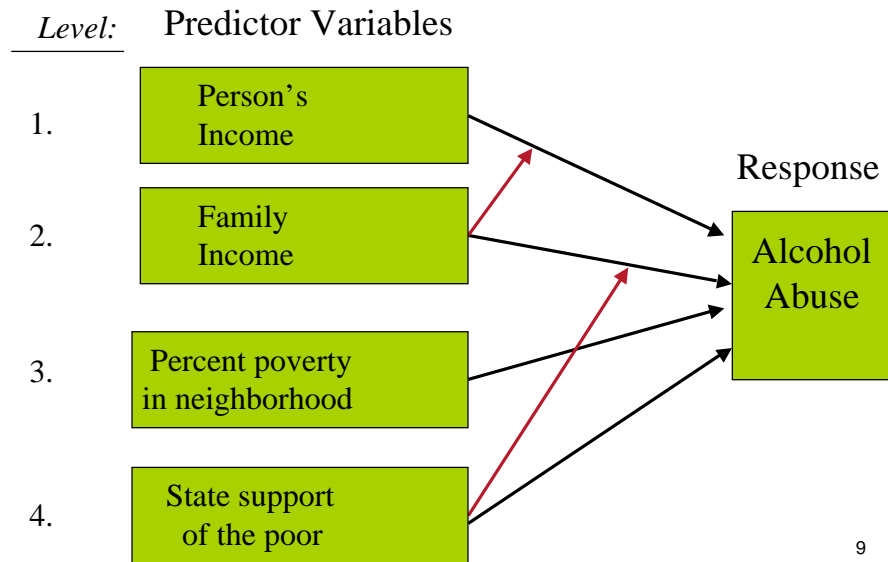
6



### Notation (cont.)

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

## Multi-level Models: Idea



## A Rose is a Rose is a...

- Multi-level model
- Random effects model
- Mixed model
- Random coefficient model
- Hierarchical model

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.

11

## Generalized Linear Models (GLMs)

$$g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

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

Model	Response	$g(\mu)$	Distribution	Coef Interp
Linear	Continuous (ounces)	$\mu$	Gaussian	Change in avg(Y) per unit change in X
Logistic	Binary (disease)	$\log\left(\frac{\mu}{1-\mu}\right)$	Binomial	Log Odds Ratio
Log-linear	Count/Times to events	$\log(\mu)$	Poisson	Log Relative Risk

12

## Generalized Linear Models (GLMs)

$$g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

### Example: Age & Gender

$$\text{Gaussian – Linear: } E(y) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender}$$

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

WHY?

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

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

---


$$\Delta E(y) = \beta_1 \quad 13$$

## Generalized Linear Models (GLMs)

$$g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

### Example: Age & Gender

$$\text{Binary – Logistic: } \log\{\text{odds}(Y)\} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender}$$

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

WHY?

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

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

---


$$\begin{array}{l} \Delta \log\text{-Odds} \\ \longrightarrow \log\text{-OR} \end{array} = \beta_1 \quad 14$$

## Generalized Linear Models (GLMs)

$$g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

### Example: Age & Gender

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

$\beta_1$  = log-RR for a 1 unit increase in Age,  
Comparing people of the SAME GENDER.

WHY?

*Self-Check: Verify Tonight*

15

## “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**

16

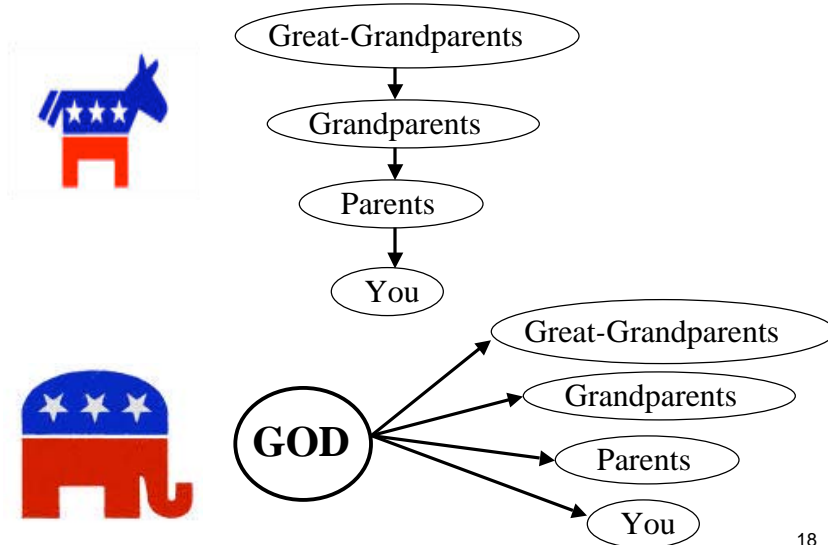


## Non-independent responses (Within-Cluster Correlation)

- 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?

17

## Why? (Family Wealth Example)



18

## 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.

19

Correlated Data...  
(within-cluster associations)

20

## 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.

21

## Regression with Correlated Data

Must take account of correlation to:

- Obtain valid inferences
  - standard errors
  - confidence intervals
- Make efficient inferences

22

## Logistic Regression Example: Cross-over trial

- Response: 1-normal; 0- alcohol dependence
- Predictors: period ( $x_1$ ); treatment group ( $x_2$ )
- Two observations per person (cluster)
- Parameter of interest: log odds ratio of dependence: treatment vs placebo

Mean Model:  $\log\{\text{odds(AD)}\} = \beta_0 + \beta_1 \text{Period} + \beta_2 \text{Trt}$

23

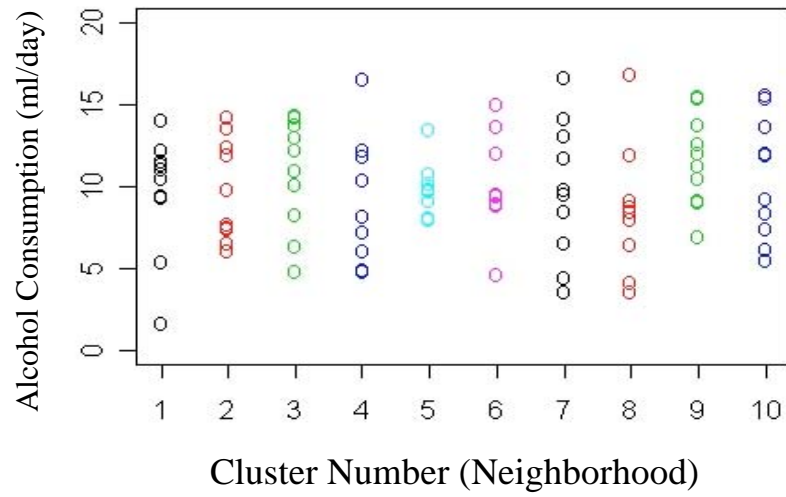
## Results: estimate, (standard error)

Variable	Model	
	Ordinary Logistic Regression	Account for correlation
Intercept ( $\beta_0$ )	0.66 <span style="color: red;">(0.32)</span>	0.67 <span style="color: red;">(0.29)</span>
Period ( $\beta_1$ )	-0.27 <span style="color: red;">(0.38)</span>	-0.30 <span style="color: red;">(0.23)</span>
Treatment ( $\beta_2$ )	0.56 <span style="color: red;">(0.38)</span>	0.57 <span style="color: red;">(0.23)</span>

Similar Estimates,

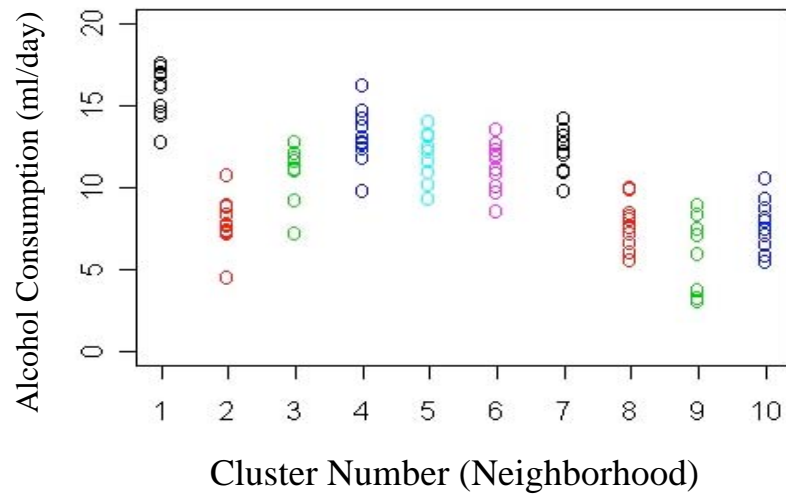
**WRONG** Standard Errors (& Inferences) for OLR <sup>24</sup>

## Simulated Data: Non-Clustered



25

## Simulated Data: Clustered



26

## 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 =  $(9.8-3.2) / 9.8 = 0.67$

27

## 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)

28

## Marginal vs Conditional Models...

29

### 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.

30

## Marginal Model Interpretations

- $\log\{\text{odds(AlcDep)}\} = \beta_0 + \beta_1\text{Period} + \beta_2\text{pl}$   
 $= 0.67 + (-0.30)\text{Period} + (0.57)\text{pl}$

TRT Effect: (placebo vs. trt)

OR =  $\exp(0.57) = 1.77$ , 95% CI (1.12, 2.80)

→ *Risk of Alcohol Dependence is almost twice as high on placebo, regardless of, (adjusting for), time period*

WHY?

Since:  $\log\{\text{odds(AlcDep|Period, pl)}\} = \beta_0 + \beta_1\text{Period} + \beta_2$

And:  $\log\{\text{odds(AlcDep|Period, trt)}\} = \beta_0 + \beta_1\text{Period}$

$$\Delta \log\text{-Odds} = \beta_2$$

$$\rightarrow \text{OR} = \exp(\beta_2)$$

31

## 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,...

32



## Random Effects Model Interpretations

WHY?

Since:  $\log\{\text{odds}(\text{AlcDep}_i | \text{Period, pl}, b_i)\} = \beta_0 + \beta_1 \text{Period} + \beta_2 + b_i$

And:  $\log\{\text{odds}(\text{AlcDep}_i | \text{Period, trt}, b_i)\} = \beta_0 + \beta_1 \text{Period} + \beta_2 + b_i$

$$\begin{array}{rcl} \Delta \log\text{-Odds} & = & \beta_2 \\ \longrightarrow \text{OR} & = & \exp(\beta_2) \end{array}$$

- 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.

33

## Marginal vs. Random Effects Models

- For **linear models**, regression coefficients in random effects models and marginal models are identical:  
average of linear function = 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

34

## Marginal -vs- Random Intercept Models; Cross-over Example

Variable	Model		
	Ordinary Logistic Regression	Marginal (GEE) Logistic Regression	Random-Effect Logistic Regression
Intercept	0.66 (0.32)	0.67 (0.29)	2.2 (1.0)
Period	-0.27 (0.38)	-0.30 (0.23)	-1.0 (0.84)
Treatment	0.56 (0.38)	0.57 (0.23)	1.8 (0.93)
Log OR (assoc.)	0.0	3.56 (0.81)	5.0 (2.3)

35

## 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 AlcDep is  $\exp(.57) = 1.8$  greater for placebo than on active drug;  
*population-average parameter*
  - Random Effects: **a person's odds** of AlcDep 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.**

36

# 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.

37

## 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

38

## 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

39

## 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).

40