

Module I: Statistical Background on Multi-level Models

Francesca Dominici
Michael Griswold
The Johns Hopkins University
Bloomberg School of Public Health

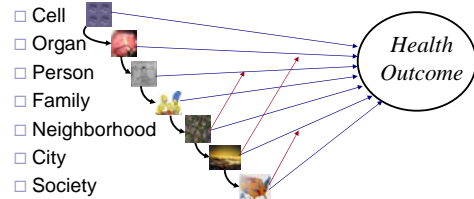
Statistical Background on MLMs

- Module 1:
 - Main Ideas on Multilevel Models
 - Review of GLMs (Generalized Linear Models)
 - Accounting for Correlated Data
 - Bayes Theorem
 - Bayesian Inference and Computation

The Main Idea...

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

Example: Alcohol Abuse

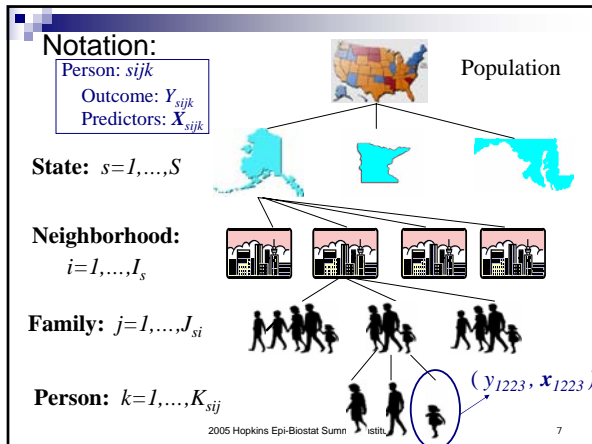
Level:

- | | |
|------------------|--|
| 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 |

Example: Alcohol Abuse; Interactions between Levels

Level:

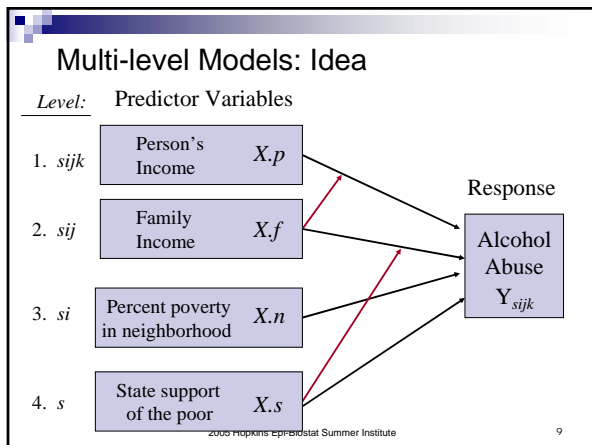
- | | | |
|---|---|---|
| 5 | } | Availability of bars <i>and</i> |
| 6 | | State laws about drunk driving |
| 4 | } | Alcohol abuse in the family <i>and</i> |
| 2 | | Person's ability to metabolize ethanol |
| 3 | } | Genetic predisposition to addiction <i>and</i> |
| 4 | | Household environment |
|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



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 $s = 1, \dots, S$
- $\mu_{sijk} = E(y_{sijk} | x_{sijk})$

2005 Hopkins Epi-Biostat Summer Institute 8



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.

2005 Hopkins Epi-Biostat Summer Institute 10

Generalized Linear Models (Review)

2005 Hopkins Epi-Biostat Summer Institute 11

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.

2005 Hopkins Epi-Biostat Summer Institute 12

Generalized Linear Models (GLMs)

$$g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (\mu = E(Y|X) = \text{mean})$$

Model	Response	$g(\mu)$	Distribution	Coef Interp
Linear	Continuous (ounces)	μ	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

2005 Hopkins Epi-Biostat Summer Institute 13

Generalized Linear Models (GLMs)

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

Example: Age & Gender

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

β_1 = Change in Average Response per 1 unit increase in Age, Comparing people of the SAME GENDER.

WHY?

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

$$\frac{\Delta E(y)}{\Delta \text{Age}} = \beta_1$$

2005 Hopkins Epi-Biostat Summer Institute 14

Generalized Linear Models (GLMs)

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

Example: Age & Gender

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

β_1 = log-OR of “+ Response” for a 1 unit increase in Age, Comparing people of the SAME GENDER.

WHY?

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

$$\frac{\Delta \log\text{-Odds}}{\Delta \text{Age}} = \beta_1$$

2005 Hopkins Epi-Biostat Summer Institute 15

Generalized Linear Models (GLMs)

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

Example: Age & Gender

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

β_1 = log-RR for a 1 unit increase in Age, Comparing people of the SAME GENDER.

WHY?

Self-Check: Verify Tonight

2005 Hopkins Epi-Biostat Summer Institute 16

Correlated Data...

2005 Hopkins Epi-Biostat Summer Institute 17

“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**

2005 Hopkins Epi-Biostat Summer Institute 18

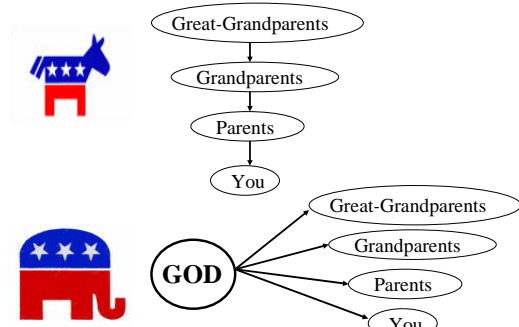
Non-independent responses (Within-Cluster Correlation)

- 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?

2005 Hopkins Epi-Biostat Summer Institute

19

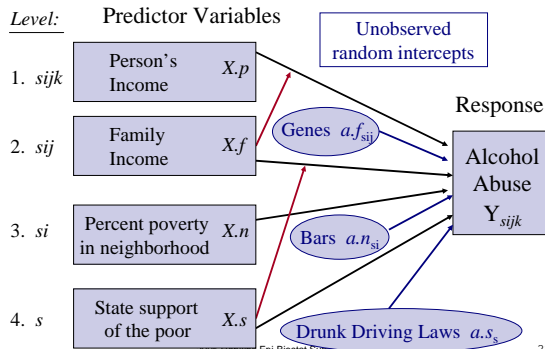
Why? (Family Wealth Example)



2005 Hopkins Epi-Biostat Summer Institute

20

Multi-level Models: Idea



2005 Hopkins Epi-Biostat Summer Institute

21

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.

2005 Hopkins Epi-Biostat Summer Institute

22

Multi-level Shmulti-level

- 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.

2005 Hopkins Epi-Biostat Summer Institute

23

Regression with Correlated Data

Must take account of correlation to:

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

2005 Hopkins Epi-Biostat Summer Institute

24

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

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

2005 Hopkins Epi-Biostat Summer Institute

25

Results: estimate, (standard error)

Variable	Model	
	Ordinary Logistic Regression	Account for correlation
Intercept (β_0)	0.66 (0.32)	0.67 (0.29)
Period (β_1)	-0.27 (0.38)	-0.30 (0.23)
Treatment (β_2)	0.56 (0.38)	0.57 (0.23)

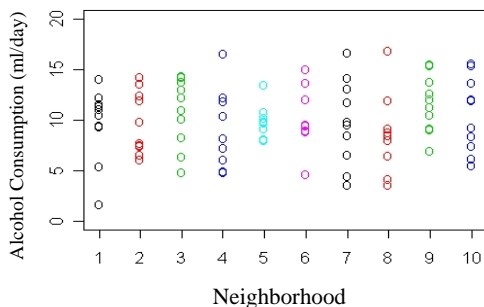
Similar Estimates,

WRONG Standard Errors (& Inferences) for OLR

2005 Hopkins Epi-Biostat Summer Institute

26

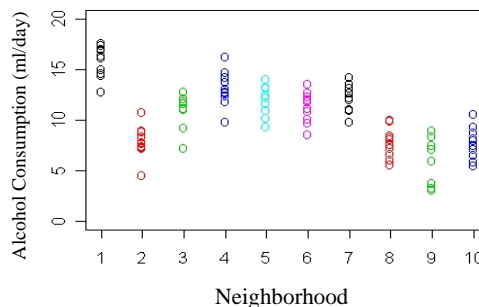
Simulated Data: Non-Clustered



2005 Hopkins Epi-Biostat Summer Institute

27

Simulated Data: Clustered



2005 Hopkins Epi-Biostat Summer Institute

28

Within-Cluster Correlation

- Correlation of two observations from same cluster =

$$\frac{\text{Total Var} - \text{Within Var}}{\text{Total Var}}$$

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

2005 Hopkins Epi-Biostat Summer Institute

29

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
 - Operating under a Bayesian paradigm

2005 Hopkins Epi-Biostat Summer Institute

30

Basic Bayes...

Diagnostic Testing

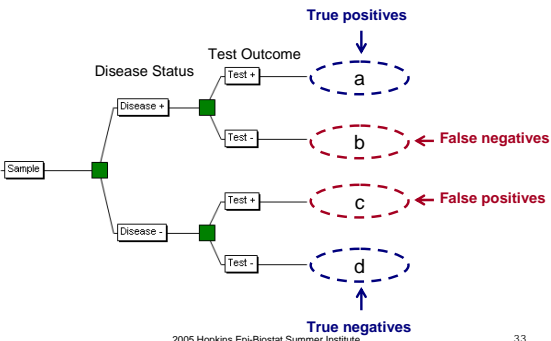
Ask
Marilyn®



BY MARILYN VOS SAVANT

A particularly interesting and important question today is that of testing for drugs. Suppose it is assumed that about 5% of the general population uses drugs. You employ a test that is 95% accurate, which we'll say means that if the individual is a user, the test will be positive 95% of the time, and if the individual is a nonuser, the test will be negative 95% of the time. A person is selected at random and is given the test. It's positive. What does such a result suggest? Would you conclude that the individual is a drug user? What is the probability that the person is a drug user?

Diagnostic Testing



Diagnostic Testing

■ “The workhorse of Epi”: The 2 × 2 table

	Disease +	Disease -	Total
Test +	a	b	a + b
Test -	c	d	c + d
Total	a + c	b + d	a + b + c + d

Diagnostic Testing

■ “The workhorse of Epi”: The 2 × 2 table

	Disease +	Disease -	Total
Test +	a	b	a + b
Test -	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$Sens = P(+|D) = \frac{a}{a+c} \quad Spec = P(-|\bar{D}) = \frac{d}{b+d}$$

Diagnostic Testing

■ “The workhorse of Epi”: The 2 × 2 table

	Disease +	Disease -	Total
Test +	a	b	a + b
Test -	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$Sens = P(+|D) = \frac{a}{a+c} \quad Spec = P(-|\bar{D}) = \frac{d}{b+d}$$

Diagnostic Testing

- Marilyn's Example $\begin{cases} \text{Sens} = 0.95 \\ \text{Spec} = 0.95 \end{cases}$

	Disease +	Disease -	Total	
Test +	48	47	95	PPV = 51%
Test -	2	903	905	NPV = 99%
Total	50	950	1000	

$P(D) = 0.05$

2005 Hopkins Epi-Biostat Summer Institute 37

Diagnostic Testing

- Marilyn's Example $\begin{cases} \text{Sens} = 0.95 \\ \text{Spec} = 0.95 \end{cases}$

	Disease +	Disease -	Total	
Test +	190	40	230	PPV = 83%
Test -	10	760	770	NPV = 99%
Total	200	800	1000	

$P(D) = 0.20$

Point: PPV depends on prior probability of disease in the population

2005 Hopkins Epi-Biostat Summer Institute 38

Diagnostic Testing & Bayes Theorem

- Bayesian Formulation:
 - Parameter of interest: "D":
 - D=0 if disease free, D=1 if diseased
 - Prior distribution of the parameter D:
 - $\Pr(D=1)$ (Prevalence of disease in general pop'n)
 - Data: to provide evidence about the parameter
 - Y=0 if test negative, Y=1 if test positive
 - Likelihood: $\Pr(\text{data} \mid \text{specific parameter value})$
 - $\Pr(Y|D) = \text{sens, spec; i.e. } \Pr(Y=1|D=1), \Pr(Y=0|D=0)$
 - Posterior distribution of the parameter D:
 - $\Pr(D|Y) = \Pr(\text{diseased} \mid \text{test outcome})$

$\propto \Pr(Y|D)P(D) = \text{Likelihood} * \text{Prior}$

2005 Hopkins Epi-Biostat Summer Institute 39

Diagnostic Testing & Bayes Theorem

- Marilyn's Example:
 - Parameter of interest: "D":
 - D=0 if disease free, D=1 if diseased
 - Prior distribution of the parameter D:
 - $\Pr(D=1) = 0.05$
 - Data: to provide evidence about the parameter
 - Suppose positive test observed: Y=1
 - Likelihood: $\Pr(\text{data} \mid \text{specific parameter value})$
 - $\Pr(Y=1|D=1) = 0.95$, (sens);
 - $\Pr(Y=1|D=0) = 1 - \Pr(Y=0|D=0) = 1 - 0.95 = 0.05$
 - Posterior distribution of the parameter D:
 - $\Pr(D=1|Y=1) = \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.05 \times 0.95} = 0.50$

2005 Hop. 40

Diagnostic Testing & Bayes Theorem

Bayes Theorem lets us combine our **Prior** beliefs with the **Likelihood** of having observed our **Data** to obtain **Posterior** inferences about the parameters we're interested in, given the data we saw.

2005 Hopkins Epi-Biostat Summer Institute 41

Key Points

- "Multi-level" Models:
 - Have covariates from many levels and their interactions
 - Acknowledge correlation among observations from within a level (cluster)
- Bayesian Inference:
 - Assumptions about the latent variables determine the nature of the within cluster correlations
 - Information can be borrowed across clusters (levels) to improve individual estimates

2005 Hopkins Epi-Biostat Summer Institute 42