

## Analysis of Multilevel Data and Applications to Social Contagion

Brian G. Leroux

leroux@u.washington.edu

Health Disparities Symposium

May 19, 2003

### Outline

1. Overview of multilevel models
2. 3 formulations (direct, random-effects, conditional)
  - parameter interpretation
3. Applications to smoking prevention research:
  - a. community-randomized trials
  - b. modeling smoking epidemics

**Multi-level model** (most general meaning):

any statistical model for outcomes that characterize *elementary units* that are grouped into *primary sampling units* (“clusters”)

References:

Hierarchical Linear Models, by Bryk and Raudenbush, 1992

Variance Components, by Searle, Casella, and McCulloch, 1992

Random-Coefficient Models, by Longford, 1993

Models for Repeated Measurements, by Lindsey, 1993

Multi-Level Statistical Models, by Goldstein, 1995

Generalized, Linear and Mixed Models, by McCulloch and Searle, 2001

Longitudinal Data Analysis, by Diggle, Heagerty, Liang, and Zeger, 2002

Examples:

1. Health Surveys:

Community  $\longrightarrow$  Household  $\longrightarrow$  Person (outcome)

2. Community-Randomized Trial:

Community (randomized to treatment condition)  $\longrightarrow$  Person (outcome)

3. Matched Community-Randomized Trial:

Community-Pair  $\longrightarrow$  Community (randomized to treatment condition)  $\longrightarrow$  Person (outcome)

**Applications.**

<b>Field</b>	<b>Outcome Variable</b>	<b>e.u.</b>	<b>p.s.u.</b>	<b>Terminology</b>
1. Agriculture	Crop yield	A plot of land	A field of plots	Agricultural field trial
2. Cancer	Biomarker measurement	One measure	Multiple measures	Cancer screening study
3. Child development	Height	One measurement	Series of measures	Growth study
4. Dentistry	Presence of caries	A tooth	Several teeth in a patient	Caries study
5. Education	Test score Test score	A student A student	Students in a classroom	Educational assessment
6. Epidemiology	Presence of disease	Person	Geographical region	Environmental risk factor study
7. Genetics	Biological trait	A twin	A pair of twins	Twin study
8. Health promotion	Smoking acquisition	A child	Children in a school	School-based smoking prevention
9. Health Services	Frequency of medical procedure	A physician	Physicians at a hospital	Small-area analysis
10. Medicine	Disease diagnosis	One test	Multiple tests on same patient	Diagnostic Test Evaluation
11. Medicine	Treatment effect	One trial	Multiple trials	Meta-analysis
12. Nutrition	Vitamin intake	Person	Household	Household nutrition survey
13. Ophthalmology	Visual acuity	An eye	Pair of eyes in one person	Paired study
14. Pharmacology	Serum concentration	One measurement	Multiple measures	Pharmacokinetic study
15. Psychology	Response to stimulus	One stimulus	Multiple stimuli	Repeated Measures Study
16. Radiology	Radiographic diagnosis	One rater	Multiple raters rating one film	Reliability Study
17. Toxicology	Toxic response	One animal	Litter of animals	Littermate study
18. Toxicology	Potency	One bioassay	Multiple bioassays	Inter-laboratory Comparison

**Illustration:** Hutchinson Smoking Prevention Project (HSPP)

*Design:* long-term matched community-randomized trial

20 SD pairs  $\rightarrow$  2 SDs (trt/control)  $\rightarrow \approx 200$  3rd grade students  
(total  $N \approx 8000$ )

School districts randomized to intervention or control conditions within each of 20 matched pairs of school districts

*Primary Goal:* determine if a school-based smoking prevention program reduces prevalence of smoking at 12th grade

*Primary outcome variable:* indicator of daily smoking at 12th grade.

*Secondary Goal:* determine predictors of 12th grade smoking

*Ref:* Peterson AV Jr, Kealey KA, Mann SL, Marek PM, Sarason IG. Hutchinson Smoking Prevention Project: long-term randomized trial in school-based tobacco use prevention—results on smoking. J Natl Cancer Inst. 2000 Dec 20;92(24):1979-91.

HSPP/Parent Protective Factors/Social Environments Investigators

Art Peterson  
Kathleen Kealey  
Sue Mann  
Pat Marek  
Irwin Sarason  
Brian Leroux  
Robyn Anderson  
Jonathan Bricker  
Bharat Rajan

**Advantages of Multi-Level Designs**

1. Richer class of scientific questions can be addressed compared with single-level designs:

1a. can distinguish

B=between-cluster effect = comparison of mean outcome for two units with predictor values differing by 1, in different clusters

versus

W=within-cluster effect = comparison of mean outcome for two units with predictor values differing by 1, in the same cluster

(In longitudinal data setting, W=longitudinal, B=cross-sectional.)

Q: Do i want 1) B only, 2) W only, 3) both, 4) estimate a common value?

1b. mean outcome for one unit may depend on predictors on

other units in the same cluster (eg, treatment of one tooth effects other teeth in same mouth, intervention for some members of community may impact others in the community, mean outcome may depend on past values of the predictor as well as the current value)

1c. can estimate variability of mean outcome across clusters and estimate (predict) mean outcome for individual clusters

(eg, small area analysis). This is what random-effects models are particularly good at.

2. Increased precision about any scientific question:

more observations = more information (to a point)

For estimation of within-cluster effects, can gain very much precision.

For estimation of between-cluster effects, law of diminishing returns.

For example, if you have only a small number (eg, 10) communities, it may be impossible to achieve adequate power with a community-randomized design even with infinitely many persons per community!

### Challenges of Multi-Level Designs

1. choice of model
2. parameter interpretation
3. model fitting
4. accounting for correlation between outcomes in same cluster

(will focus on 1 and 2)

### Three Formulations of Models for Multi-level Data

1. Direct Formulation ("Marginal Model"):

specify means, variances, and correlations outcomes directly, ie, as functions of predictors but not involving other outcomes or unobserved random variables

2. Random-Effects Formulation ("RE Model", "Mixed Model", "Hierarchical Model", ...)

specify means, variances, and correlations of outcomes indirectly in terms of conditional means, variances and correlations given unobserved random effects, and specify distribution of random effects.

3. Conditional Formulation (eg, "Transition Model")

specify means, variances, and correlations of outcomes indirectly in terms of conditional means, variances and correlations given other outcomes

Example: Community-randomized trials

$Y_{ij}$  = disease indicator for person j, community i

**Direct Formulation:**

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_i,$$

(Note :  $\text{logit}(p) = \log[p/(1-p)]$ )

$$\text{var}(Y_{ij}) = p_{ij}(1 - p_{ij})$$

$$\text{corr}(Y_{ij}) = \rho$$

**Random-Effects Formulation:**

$$Y_{ij}|a_i \sim \text{Bernoulli}(p_{ij}^a)$$

$$\text{logit}(p_{ij}^a) = \beta_0 + \beta_1 x_i + a_i,$$

$$\begin{aligned}\text{var}(Y_{ij}|a_i) &= p_{ij}^a(1 - p_{ij}^a) \\ \text{corr}(Y_{ij}, Y_{ij'}|a_i) &= 0 \\ a_i &\sim \text{independent } N(0, \sigma_a^2)\end{aligned}$$

**Conditional Formulation:**

$$Y_{ij}|Y_{(-i)} \sim \text{Bernoulli}(p_{ij}^c)$$

(Note:  $Y_{(-i)} = \{Y_{ij'}, j' \neq j\}$ )

$$\begin{aligned}\text{logit}(p_{ij}^c) &= \beta_0 + \beta_1 x_i + f(Y_{(-i)}), \\ \text{var}(Y_{ij}|Y_{(-i)}) &= p_{ij}^c(1 - p_{ij}^c)\end{aligned}$$

**Which formulation should i use?**

Let your scientific question be your guide.

Warning: choice of model for variance and correlation can change the interpretation of the regression parameters

Notes:

It is not necessary to model var and corr correctly to get valid inference for a marginal model (GEE). Lesson: don't use random effects or conditional models simply to account for correlation.

You usually don't want to condition on other outcomes when estimating a treatment effect (unless studying mediation), so conditional model often not appropriate.

All three formulations allow flexible modeling of mean structures and variance structures, eg,

Family	Link Function	Formula	Variance	
			Function	Formula
Gaussian	Identity	$\mu$	Constant	$\sigma^2$
Binomial	Logit	$\text{logit}(\mu)$	Binomial	$\mu(1 - \mu)$
Poisson	Log	$\log(\mu)$	Poisson	$\mu$

May want to mix and match, eg, log link and Binomial variance to estimate a relative risk.

**Advantages of Modeling the Working Correlation Structure**

Possible to gain precision in coefficient estimates... how much gained depend on many things:

- 1) magnitude of within-cluster correlation,
- 2) correlation structure,
- 3) cluster size and variation in cluster sizes,
- 4) the distribution of the predictor variables,
- 5) link function,
- 6) coefficient values.

Rules of thumb:

- 1) precision of coefficient estimates increases as the working correlation structure becomes closer to the true correlation structure
- 2) large gains in precision are possible by using a non-independence working correlation structure compared with independence

3) law of diminishing returns sets in quickly so further refinements may not help much

Ref: Mancl and Leroux (1996) Biometrics.

Example 1: community-randomized trial analysis

(use HSPP data, ignore matching)

$Y_{ij}$  = indicator of daily smoking for student  $j$ , school district  $i$

**Direct Formulation (linear model):**

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$p_{ij} = \beta_0 + \beta_1 x_i,$$

$x_i$  is treatment indicator (cluster level).

Independence working correlation

			Robust		
y		Coef.	Std. Err.	t	P> t
trt		-.0030038	.0169861	-0.18	0.861
_cons		.2567079	.0139214	18.44	0.000

95% C.I. for trt: -.0373614, .0313539

Notes: observed prevalences are .257 control vs. .254 intervention.

HSPP investigators used a permutation test approach.

**Direct Formulation (logit model):**

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_i,$$

$$\text{var}(Y_{ij}) = p_{ij}(1 - p_{ij})$$

Independent working correlation

			Robust		
y		Coef.	Std. Err.	z	P> z
trt		-.0158031	.0892427	-0.18	0.859
_cons		-1.06315	.0729547	-14.57	0.000

95% Conf. Interval for log-odds ratio: -.1907155, .1591093

p-values are very similar for linear and logit models

Random effects linear model

$$Y_{ij}|a_i \sim \text{Bernoulli}(p_{ij}^a)$$

$$p_{ij}^a = \beta_0 + \beta_1 x_i + a_i,$$

$$a_i \sim \text{independent } N(0, \sigma_a^2)$$

y	Coef.	Std. Err.	z	P> z
trt	-.0023205	.0215277	-0.11	0.914
_cons	.2554698	.0151877	16.82	0.000

95\% Conf. Interval for trt: -.0445141, .0398731

$$\hat{\sigma}_a = .06$$

SE for trt is larger than for marginal model

Random effects logit model:

$$\text{logit}(p_{ij}^a) = \beta_0 + \beta_1 x_i + a_i,$$

$$\text{var}(Y_{ij}|a_i) = p_{ij}^a(1 - p_{ij}^a)$$

$$a_i \sim \text{independent } N(0, \sigma_a^2)$$

y	Coef.	Std. Err.	z	P> z
trt	.0060894	.1005414	0.06	0.952
_cons	-1.100588	.0767685	-14.34	0.000

95\% Conf. Interval for log-odds: -.1909681, .203147

$$\hat{\sigma}_a = .28 \text{ (variation in prevalence on logit scale)}$$

### Example 2: Prediction of 12th Grade Smoking

Large variation in prevalence between districts, most in the random effect ( $\hat{\sigma}_a = .06$ ), not binomial sampling variability.  
Possible explanation: *Social Contagion*.

**Direct Formulation (linear model):**

$Y_{ij}$  = indicator of daily smoking for student  $j$ , school district  $i$

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$p_{ij} = \beta_0 + \beta_1 x_{ij},$$

$x_{ij}$  = number of parents who smoke (individual level) = “nparsm”

Independence working correlation

Control districts only.

-----				
y	Coef.	Robust Std. Err.	t	P> t
-----				
nparsm	.1263992	.0097295	12.99	0.000
_cons	.1944862	.0123785	15.71	0.000
-----				

95\% Conf. Interval for nparsm: .106035, .1467634

**Direct Formulation (logit model):**

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij},$$

$$\text{var}(Y_{ij}) = p_{ij}(1 - p_{ij})$$

-----				
y	Coef.	Robust Std. Err.	z	P> z
-----				
nparsm	.6207752	.0413289	15.02	0.000
_cons	-1.411143	.0778805	-18.12	0.000
-----				

95\% Conf. Interval for log-odds ratio: .539772, .7017784

Note: naive SE for nparsm=.056 > robust!

Random effects linear model

$$Y_{ij}|a_i \sim \text{Bernoulli}(p_{ij}^a)$$

$$p_{ij}^a = \beta_0 + \beta_1 x_{ij} + a_i,$$

$$a_i \sim \text{independent } N(0, \sigma_a^2)$$

-----				
y	Coef.	Std. Err.	z	P> z
-----				



```

nparsm | .1249306 .0109742 11.38 0.000
_cons | .1904093 .0172459 11.04 0.000
-----+-----

```

95\% Conf. Interval for nparsm: .1034217, .1464396

How much between-district variability is explained?

Estimate of random effects std. dev.:

Without nparsm:  $\hat{\sigma}_a = .067$

With nparsm:  $\hat{\sigma}_a = .064$

The between-cluster effect of parent smoking is poorly determined:

B (Between):  $.281 \pm .177$

W (Within):  $.124 \pm .011$  (similar to overall)

The overall RE model estimate is closer to W, but the marginal model estimate weights B more.

→ Look for cluster-level predictors (as well as within-cluster) ....

Random effects logit model

$$\text{logit}(p_{ij}^a) = \beta_0 + \beta_1 x_{ij} + a_i,$$

$$\text{var}(Y_{ij}|a_i) = p_{ij}^a(1 - p_{ij}^a)$$

$$a_i \sim \text{independent } N(0, \sigma_a^2)$$

```

-----+-----
      y |      Coef.   Std. Err.      z    P>|z|
-----+-----
nparsm | .6242505   .0572635    10.90   0.000
_cons | -1.488469   .093661    -15.89   0.000
-----+-----

```

95\% Conf. Interval for log-odds ratio: .5120162, .7364849

## Modeling Smoking Epidemics

Use data that are at the level of the Individual and also Longitudinal.

Model probability of smoking acquisition and smoking cessation. For example,

$$p_{ij} = P(\text{student starts smoking between grades 7,9} | \text{not smoking at grade 7})$$

Model  $p_{ij}$  as function of # of smokers in social environments at start of interval. Predictors may be individual level (eg, parents) or cluster level (eg, smoking prevalence of 12th graders).

Which link function? Usual choices are logit, log, identity (linear model). But parameter interpretation is difficult.

New Model (use probability of not smoking):

$$1 - p_{ij} = (1 - p_0)(1 - \pi_1)^{x_{1ij}}(1 - \pi_2)^{x_{2ij}}$$

where

$$x_{1ij} = \text{number of parents who smoke at start of interval}$$

$x_{2ij}$  = number of 12th graders who smoke at start of interval

Parameter interpretation:

$\pi_1$  = probability of transmission of smoking from one parent to child given that child does not acquire smoking from other influences

$\pi_2$  = probability of transmission of smoking from one 12th grader to 7th grader given that child does not acquire smoking from other influences

## Modeling Smoking Acquisition and Cessation

(eg, young adults)

3-state process:

NEVER SMOKER  $\rightarrow$  SMOKER  $\rightarrow$  FORMER SMOKER

## References

### Books

- Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.
- Cox, D.R. and Solomon, P.J. (2002). *Components of Variance*. Chapman & Hall/CRC. [Mathematical theory of estimation of variance components in random effects models.]
- Crowder, M.J. and Hand, D.J. (1990). *Analysis of Repeated Measures*. Chapman & Hall. [Presents theory of univariate and multivariate repeated measures ANOVA for continuous outcomes.]
- Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London. [Mathematical theory of models for longitudinal continuous outcome data, and examples of applications to pharmacokinetics]
- Diggle, P.J., Heagerty, P.J., Liang, K.-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data (2nd ed.)*. Oxford: Oxford University Press. [A second edition of Diggle, Liang and Zeger's 1994 book on longitudinal data analysis, presents both mathematical theory and practical aspects and applications of methods for analysis of longitudinal data.]
- Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd ed.)*. Springer, New York. [Mathematical theory of methods for correlated data.]
- Goldstein, H. (1995). *Multilevel Statistical Models*. New York: Halsted Press.
- Hardin, J.W. and Hilbe, J.M. (2003). *Generalized Estimating Equations*. Chapman & Hall/CRC. [Contains some useful programs and data sets.]
- Hougaard P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.
- Jones, B. and Kenward, M.G. (1989). *Design and Analysis of Cross-Over Trials*. Chapman and Hall. [Practical aspects of analysis of data from clinical trials in which each patient receives two or more treatments in succession.]
- Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Hoboken. An update of a classic reference on mathematical theory of survival analysis, includes methods for multivariate survival

analysis for correlated data.]

Leyland, A.H. and Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*. John Wiley & Sons, New York.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press. [Mathematical theory of random effects models emphasizing linear models for continuous outcomes, includes several detailed worked-out examples.]

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York. [Mathematical theory of random effects models.]

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. Wiley, New York. [Mathematical theory of random effects models.]

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York. [Covers practical aspects of fitting random effects models with continuous outcomes and use of SAS PROC MIXED.]

## Articles on Results from HSPP and related studies

Andersen, M.R., Leroux, B.G., Marek, P.M., et al. (2002). Mothers' attitudes and concerns about their children smoking: Do they influence kids? *Preventive Medicine* **34**:198–206.

Bricker JB, Leroux BG, Peterson AV Jr, Kealey KA, Sarason IG, Andersen MR, Marek PM: Nine-year prospective relationship between parental smoking cessation and children's daily smoking. *Addiction* 98:585-93, 2003.

Peterson AVP Jr, Kealey KA, Mann SL, Marek PM, Sarason IG. Hutchinson Smoking Prevention Project: long-term randomized trial in school-based tobacco use prevention—results on smoking. *J Natl Cancer Inst.* 2000 Dec 20;92(24):1979-91. [Presented the results of the Hutchinson Smoking Prevention Project.]

Rajan KB et al: Nine-year prospective association between older siblings' smoking and youth smoking. *J Adolescent Health* 33:xxx-xxx, 2003 (to appear).