# Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 12a: Multilevel regression

# Multilevel regression

- Simple multilevel modeling
- Partial pooling
- Multilevel regression

# Simplest example of multilevel modeling

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | $-3$ | 16 |
| D | 7 | 11 |
| E | $-1$ | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

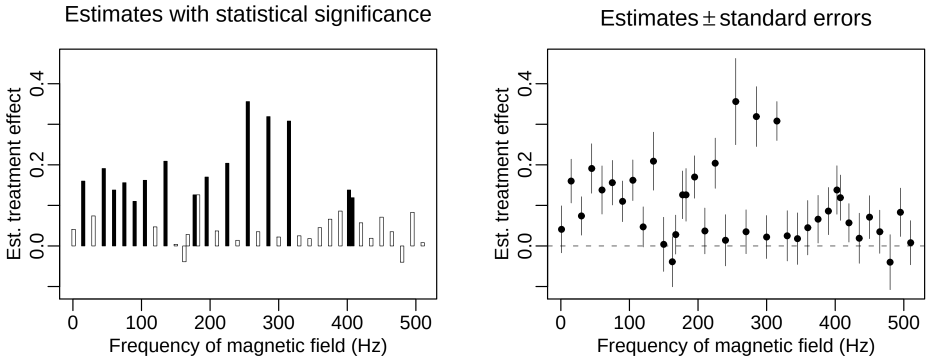
- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlapping confidence intervals for the 8 school effects
  - Statements such as $\Pr(\text{effect in A} > \text{effect in C}) = 0.7$

# The mathematics of multilevel modeling

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | $-3$ | 16 |
| D | 7 | 11 |
| E | $-1$ | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

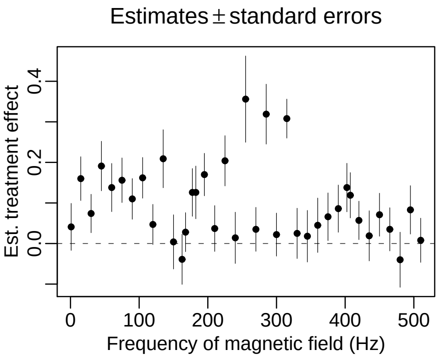
- What happens when you ...
  - Decrease the standard errors by a factor of 10?
  - Increase the standard errors by a factor of 10?
  - Move the estimates apart? Together?
  - Take the estimate for school A and move it away from all the others?

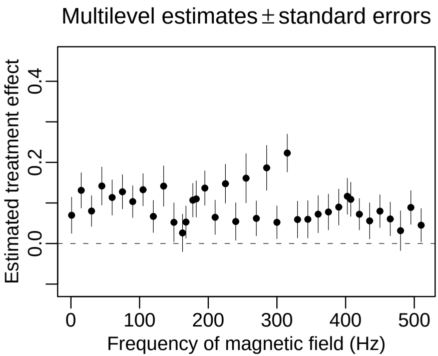
# Another example




- Effects of electromagnetic fields at 38 frequencies
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly

# Separate estimates and multilevel estimates



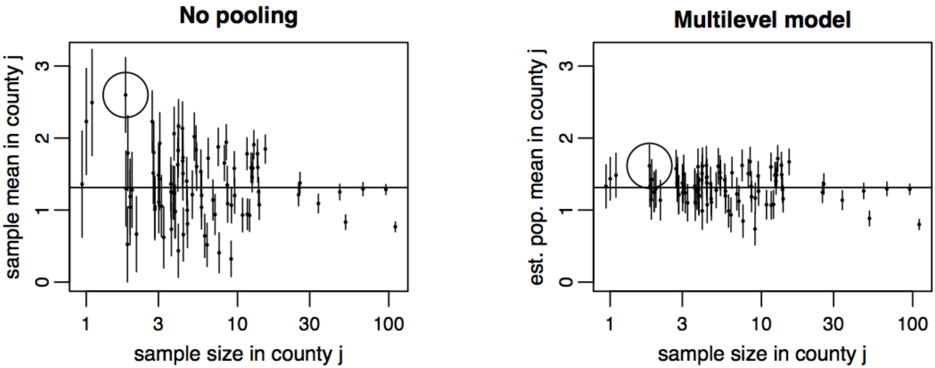




- What should we believe?

# Some other examples

- Teacher effects in schools
- Longitudinal studies
- State-level opinions
- Controlling for "country effects"
- Time series cross sections
- Common theme:
  - Sparse data
  - Small sample size in some groups
  - Partial pooling

# Sample size and partial pooling



# Multilevel regression

- Individual-level predictors
- Group-level predictors
- In either case, partially pool toward the regression line
- Simple example on blackboard: Advanced Placement scores and grades in Calculus 2
- More complicated example: public opinion among low-income white Catholics in Montana

# Sample size and variability

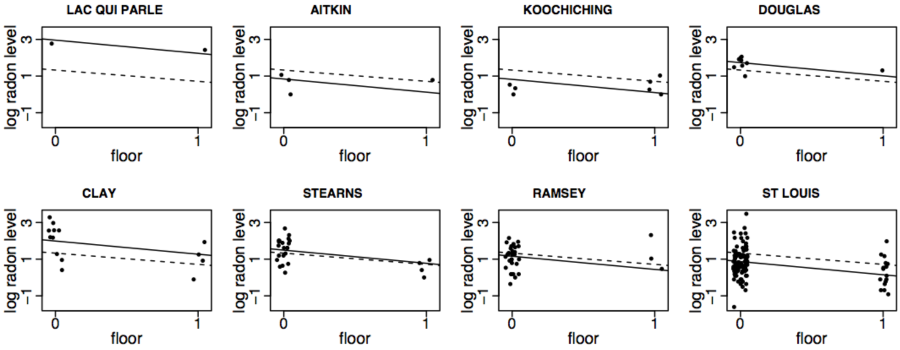



Figure 12.2 *Complete-pooling (dashed lines,* $y = \alpha + \beta x$*) and no-pooling (solid lines,*

# Sample size and partial pooling

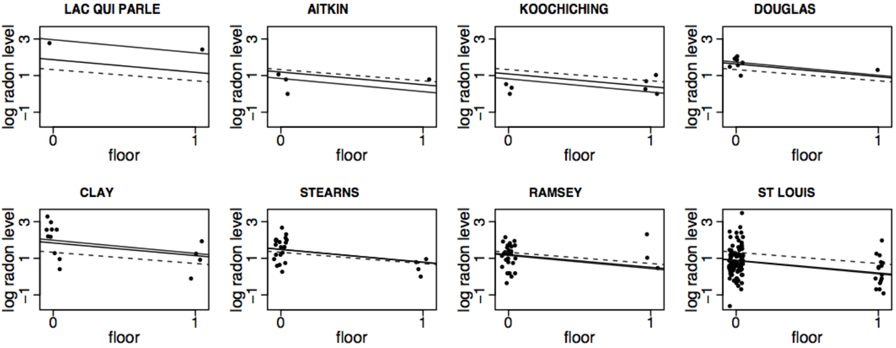
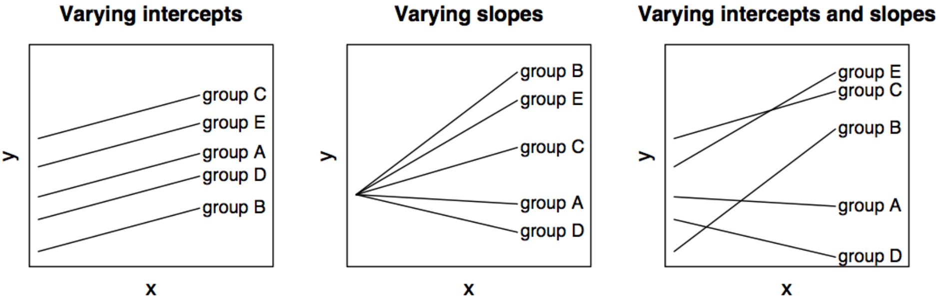



Figure 12.4 *Multilevel (partial pooling) regression lines* $y = \alpha_j + \beta x$ *fit to radon data from Minnesota, displayed for eight counties. Light-colored dashed and solid lines show the complete-pooling and no-pooling estimates, respectively, from Figure 12.3a.*

# Varying intercepts and slopes



# How *not* to code your multilevel data

| ID | dad age | mom race | informal support | city ID | city name | enforce intensity | benefit level | city indicators 1 | 2 | ⋯ | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 19 | hisp | 1 | 1 | Oakland | 0.52 | 1.01 | 1 | 0 | ⋯ | 0 |
| 2 | 27 | black | 0 | 1 | Oakland | 0.52 | 1.01 | 1 | 0 | ⋯ | 0 |
| 3 | 26 | black | 1 | 1 | Oakland | 0.52 | 1.01 | 1 | 0 | ⋯ | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ |
| 248 | 19 | white | 1 | 3 | Baltimore | 0.05 | 1.10 | 0 | 0 | ⋯ | 0 |
| 249 | 26 | black | 1 | 3 | Baltimore | 0.05 | 1.10 | 0 | 0 | ⋯ | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ |
| 1366 | 21 | black | 1 | 20 | Norfolk | −0.11 | 1.08 | 0 | 0 | ⋯ | 1 |
| 1367 | 28 | hisp | 0 | 20 | Norfolk | −0.11 | 1.08 | 0 | 0 | ⋯ | 1 |

# How to code your multilevel data

| ID | dad age | mom race | informal support | city ID |
|---|---|---|---|---|
| 1 | 19 | hisp | 1 | 1 |
| 2 | 27 | black | 0 | 1 |
| 3 | 26 | black | 1 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 248 | 19 | white | 1 | 3 |
| 249 | 26 | black | 1 | 3 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1366 | 21 | black | 1 | 20 |
| 1367 | 28 | hisp | 0 | 20 |

| city ID | city name | enforce-ment | benefit level |
|---|---|---|---|
| 1 | Oakland | 0.52 | 1.01 |
| 2 | Austin | 0.00 | 0.75 |
| 3 | Baltimore | −0.05 | 1.10 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 20 | Norfolk | −0.11 | 1.08 |

# Before doing multilevel regression

- Complete pooling
- No pooling
- Two-stage regression
- Difficulties with these approaches

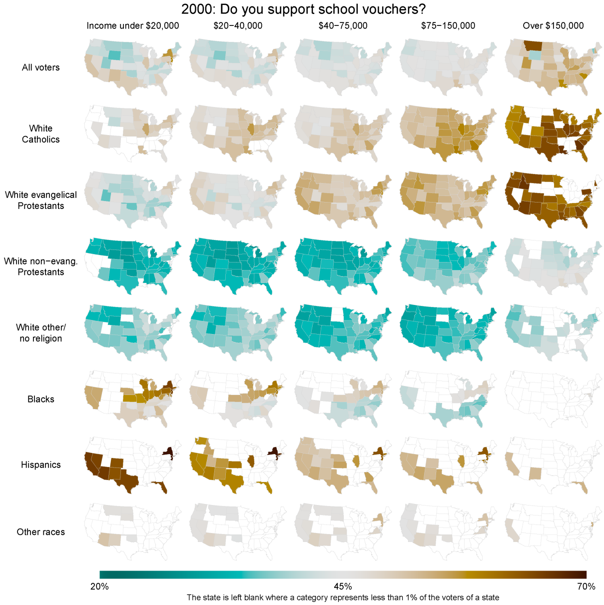
# Multilevel regression in R

- lmer
- glmer
- Individual-level predictors
- Group-level predictors
- Main effects
- Two-way interactions
- Varying intercepts and slopes
- Fake-data simulation
- Checking model fit

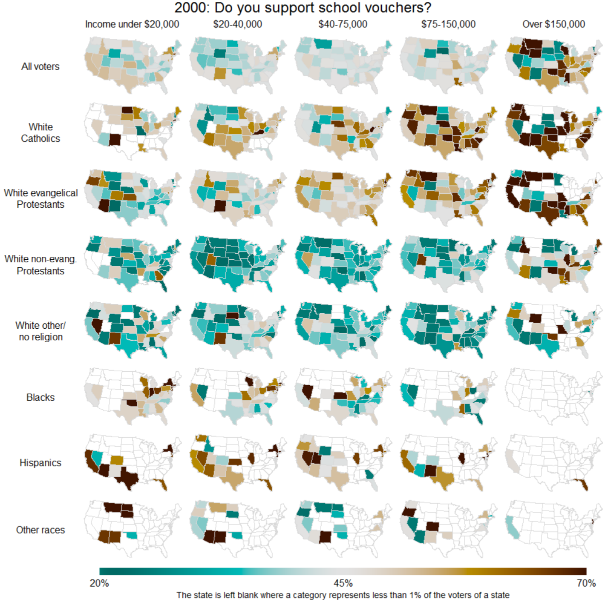
# Why multilevel regression and poststratification?

- Goal 1: State-level estimates from national surveys
  - Need to correct for known differences between sample and population
  - Compare to alternatives
- Goal 2: Inference for small subgroups of the population

# Ethnicity/religion, income, and school vouchers



# The raw data


2000: Do you support school vouchers?
Income under $20,000
$20-40,000
$40-75,000
$75-150,000
Over $150,000
All voters
White Catholics
White evangelical Protestants
White non-evang. Protestants
White other/ no religion
Blacks
Hispanics
Other races
20%
45%
70%
The state is left blank where a category represents less than 1% of the voters of a state

# Applying MRP to U.S. politics

- Census
- Post-election supplement
- Adjusting for state-level vote
- State-level predictors