Some questions (and a few answers) about multilevel models

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

3 May 2005

Themes

- Multilevel models are *necessary*
- Tools needed to *build*, *fit*, *check*, and *understand* mlms
- Analogy to linear regression
- Mlm as regression with categorical inputs

Themes

- Multilevel models are *necessary*
- Tools needed to *build*, *fit*, *check*, and *understand* mlms
- Analogy to linear regression
- Mlm as regression with categorical inputs

Themes

- Multilevel models are *necessary*
- Tools needed to *build*, *fit*, *check*, and *understand* mlms
- Analogy to linear regression
- Mlm as regression with categorical inputs

Themes

- Multilevel models are *necessary*
- Tools needed to *build*, *fit*, *check*, and *understand* mlms
- Analogy to linear regression
- Mlm as regression with categorical inputs

Themes

- Multilevel models are *necessary*
- Tools needed to *build*, *fit*, *check*, and *understand* mlms
- Analogy to linear regression
- Mlm as regression with categorical inputs

Fitting and understanding multilevel models

- Some of my experiences with multilevel models
- Some challenges and solutions
- Lots of time for questions
- Collaborators:
  - Iain Pardoe, Dept of Decision Sciences, University of Oregon
  - David Park, Joseph Bafumi, Boris Shor, Dept of Political Science, Columbia University
  - Samantha Cook, Zaiying Huang, Jouni Kerman, Shouhao Zhao, Dept of Statistics, Columbia University
  - Phillip Price, Energy and Environment Division, Lawrence Berkeley National Laboratory

Fitting and understanding multilevel models

- Some of my experiences with multilevel models
- Some challenges and solutions
- Lots of time for questions
- Collaborators:
  - Iain Pardoe, Dept of Decision Sciences, University of Oregon
  - David Park, Joseph Bafumi, Boris Shor, Dept of Political Science, Columbia University
  - Samantha Cook, Zaiying Huang, Jouni Kerman, Shouhao Zhao, Dept of Statistics, Columbia University
  - Phillip Price, Energy and Environment Division, Lawrence Berkeley National Laboratory

Fitting and understanding multilevel models

- Some of my experiences with multilevel models
- Some challenges and solutions
- Lots of time for questions
- Collaborators:
  - Iain Pardoe, Dept of Decision Sciences, University of Oregon
  - David Park, Joseph Bafumi, Boris Shor, Dept of Political Science, Columbia University
  - Samantha Cook, Zaiying Huang, Jouni Kerman, Shouhao Zhao, Dept of Statistics, Columbia University
  - Phillip Price, Energy and Environment Division, Lawrence Berkeley National Laboratory

Fitting and understanding multilevel models

- Some of my experiences with multilevel models
- Some challenges and solutions
- Lots of time for questions
- Collaborators:
  - Iain Pardoe, Dept of Decision Sciences, University of Oregon
  - David Park, Joseph Bafumi, Boris Shor, Dept of Political Science, Columbia University
  - Samantha Cook, Zaiying Huang, Jouni Kerman, Shouhao Zhao, Dept of Statistics, Columbia University
  - Phillip Price, Energy and Environment Division, Lawrence Berkeley National Laboratory

Fitting and understanding multilevel models

- Some of my experiences with multilevel models
- Some challenges and solutions
- Lots of time for questions
- Collaborators:
  - Iain Pardoe, Dept of Decision Sciences, University of Oregon
  - David Park, Joseph Bafumi, Boris Shor, Dept of Political Science, Columbia University
  - Samantha Cook, Zaiying Huang, Jouni Kerman, Shouhao Zhao, Dept of Statistics, Columbia University
  - Phillip Price, Energy and Environment Division, Lawrence Berkeley National Laboratory

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions ...

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions . . .

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions . . .

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions . . .

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions . . .

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions ...

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions . . .

Plan of talk

- Rodents in NYC: apts within buildings within neighborhoods
- State-level opinions from national polls: mlm and poststratification
- Mlm when number of groups is small
- Finite-population and superpopulation inference
- Understanding a fitted multilevel regression: Anova, average predictive effects, partial pooling, and R^2
- Why I don't use the terms "fixed" and "random" effects
- Questions . . .

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:
  - Fit to subset of the data (900 apts in 500 bldgs)
  - Fit to all the data, separate model for each neighborhood

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:
  - Fit to subset of the data (900 apts in 500 bldgs)
  - Fit to all the data, separate model for each neighborhood

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:
  - Fit to subset of the data (900 apts in 500 bldgs)
  - Fit to all the data, separate model for each neighborhood

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:
  - Fit to subset of the data (900 apts in 500 bldgs)
  - Fit to all the data, separate model for each neighborhood

NYC Dept of Health study

- Survey of 16000 apts in 9000 bldgs in 55 neighborhoods in NYC
- Do you have rodents?
- Hierarchical logistic regression:

$$\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i + \alpha_{\text{bldg}(i)} + \gamma_{\text{neighborhood}(i)})$$

- Try to fit in WinBUGS, but too slow! Solutions:
  - Fit to subset of the data (900 apts in 500 bldgs)
  - Fit to all the data, separate model for each neighborhood

National opinion trends

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex × ethnicity × age × education × state

State-level opinion trends

- Goal: estimating time series within each state
- One poll at a time: small-area estimation
- It works! Validated for pre-election polls
- Combining surveys: model for parallel time series
- Multilevel modeling + poststratification
- Poststratification cells: sex $\times$ ethnicity $\times$ age $\times$ education $\times$ state

Multilevel modeling of opinions

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Bayesian inference, summarize by posterior simulations of β:

Simulation	θ_1	$\cdots$	θ_{75}
1	**	$\cdots$	**
$\vdots$	$\vdots$	$\ddots$	$\vdots$
1000	**	$\cdots$	**

Multilevel modeling of opinions

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Bayesian inference, summarize by posterior simulations of β:

Simulation	θ_1	$\cdots$	θ_{75}
1	**	$\cdots$	**
$\vdots$	$\vdots$	$\ddots$	$\vdots$
1000	**	$\cdots$	**

Multilevel modeling of opinions

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Bayesian inference, summarize by posterior simulations of β:

Simulation	θ_1	$\cdots$	θ_{75}
1	**	$\cdots$	**
$\vdots$	$\vdots$	$\ddots$	$\vdots$
1000	**	$\cdots$	**

Multilevel modeling of opinions

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Bayesian inference, summarize by posterior simulations of β:

Simulation	θ_1	$\cdots$	θ_{75}
1	**	$\cdots$	**
$\vdots$	$\vdots$	$\ddots$	$\vdots$
1000	**	$\cdots$	**

Multilevel modeling of opinions

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Bayesian inference, summarize by posterior simulations of β:

Simulation	θ_1	$\cdots$	θ_{75}
1	**	$\cdots$	**
$\vdots$	$\vdots$	$\ddots$	$\vdots$
1000	**	$\cdots$	**

Multilevel modeling of opinions

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Bayesian inference, summarize by posterior simulations of β:

Simulation	θ_1	$\cdots$	θ_{75}
1	**	...	**
$\vdots$	$\vdots$	$\ddots$	$\vdots$
1000	**	...	**

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Interlude: why "multilevel" $\neq$ "hierarchical"

- Logistic regression: $\Pr(y_i = 1) = \text{logit}^{-1}((X\beta)_i)$
- X includes demographic and geographic predictors
- Group-level model for the 16 age $\times$ education predictors
- Group-level model for the 50 state predictors
- Crossed (nonnested) structure of age, education, state
- Several overlapping "hierarchies"

Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells j (e.g., black female, age 18–29, college graduate, Georgia)
- Poststratification
  - Within each state s, average over 64 cells:
    $\sum_{j \in s} N_j \theta_j / \sum_{j \in s} N_j$
    N_j = population in cell j (from Census)
    1000 simulation draws propagate to uncertainty for each θ_s

Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells j (e.g., black female, age 18–29, college graduate, Georgia)
- Poststratification
  - Within each state s, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - N_j = population in cell j (from Census)
  - 1000 simulation draws propagate to uncertainty for each θ_s

Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells j (e.g., black female, age 18–29, college graduate, Georgia)
- Poststratification
  - Within each state s, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - N_j = population in cell j (from Census)
  - 1000 simulation draws propagate to uncertainty for each θ_j

Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells j (e.g., black female, age 18–29, college graduate, Georgia)
- Poststratification
  - Within each state s, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - N_j = population in cell j (from Census)
  - 1000 simulation draws propagate to uncertainty for each θ_j

Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells j (e.g., black female, age 18–29, college graduate, Georgia)
- Poststratification
  - Within each state s, average over 64 cells: $\sum_{j\in s} N_j\theta_j \Big/ \sum_{j\in s} N_j$
  - N_j = population in cell j (from Census)
  - 1000 simulation draws propagate to uncertainty for each θ_j

Poststratification to estimate state opinions

- Implied inference for $\theta_j = \text{logit}^{-1}(X\beta)$ in each of 3264 cells j (e.g., black female, age 18–29, college graduate, Georgia)
- Poststratification
  - Within each state s, average over 64 cells: $\sum_{j \in s} N_j \theta_j \Big/ \sum_{j \in s} N_j$
  - N_j = population in cell j (from Census)
  - 1000 simulation draws propagate to uncertainty for each θ_j

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
    Complete pooling: no state predictors
    Hierarchical model and poststratify
- Mean absolute state errors:

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

CBS/New York Times pre-election polls from 1988

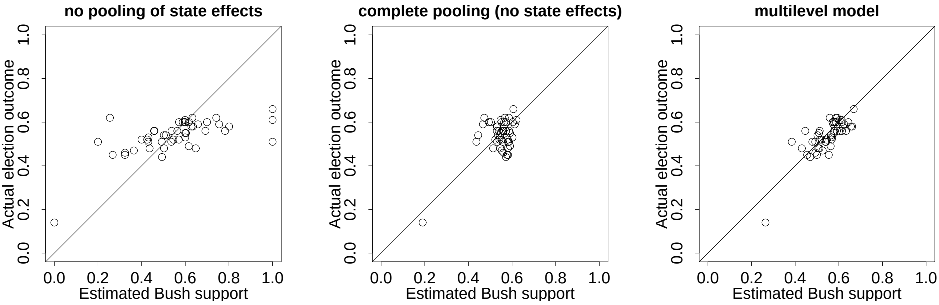
- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

CBS/New York Times pre-election polls from 1988

- Validation study: fit model on poll data and compare to election results
- Competing estimates:
  - No pooling: separate estimate within each state
  - Complete pooling: no state predictors
  - Hierarchical model and poststratify
- Mean absolute state errors:
  - No pooling: 10.4%
  - Complete pooling: 5.4%
  - Hierarchical model with poststratification: 4.5%

Validation study: comparison of state errors

1988 election outcome vs. poll estimate



How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
    Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category
- Extends to interactions (16 age × education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category
- Extends to interactions (16 age × education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category"
- Extends to interactions (16 age × education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category"
- Extends to interactions (16 age × education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category"
- Extends to interactions (16 age $\times$ education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category"
- Extends to interactions (16 age $\times$ education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category"
- Extends to interactions (16 age × education categories)

How many groups do you need to fit a mlm?

- 9000 bldgs, 55 neighborhoods, 50 states: that's ok
- But why do mlm with only 4 categories?
  - Age 18–29, 30–44, 45–64, 65+
  - Education less than HS, HS, some college, college grad
- Simple to set up as mlm
- No need to choose a "baseline" category"
- Extends to interactions (16 age $\times$ education categories)

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\text{age}}, \ldots, \beta_4^{\text{age}}$
- Finite-population centering:

$$\tilde{\beta}_j^{\text{age}} = \beta_j^{\text{age}} - \bar{\beta}^{\text{age}}, \text{ for } j = 1, \ldots, 4$$
$$\tilde{\beta}_0 = \beta_0 + \bar{\beta}^{\text{age}}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\text{age}}, \ldots, \beta_4^{\text{age}}$
- Finite-population centering:

$$\begin{aligned}\tilde{\beta}_j^{\text{age}} &= \beta_j^{\text{age}} - \bar{\beta}^{\text{age}}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\text{age}}\end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\text{age}}, \ldots, \beta_4^{\text{age}}$
- Finite-population centering:

$$\begin{aligned} \tilde{\beta}_j^{\text{age}} &= \beta_j^{\text{age}} - \bar{\beta}^{\text{age}}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\text{age}} \end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\text{age}}, \ldots, \beta_4^{\text{age}}$
- Finite-population centering:

$$\begin{aligned} \tilde{\beta}_j^{\text{age}} &= \beta_j^{\text{age}} - \bar{\beta}^{\text{age}}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\text{age}} \end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\text{age}}, \ldots, \beta_4^{\text{age}}$
- Finite-population centering:

$$\begin{aligned} \tilde{\beta}_j^{\text{age}} &= \beta_j^{\text{age}} - \bar{\beta}^{\text{age}}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\text{age}} \end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects
  - $\beta_j^{\text{age}} \sim \text{N}(0, \sigma_{\text{age}}^2)$, for j=1, . . . ,4
  - Superpopulation sd: σ_{age}
  - Finite-population sd: $\sqrt{\frac{1}{3}\sum_{j=1}^{4}(\beta_j^{\text{age}} - \bar{\beta}^{\text{age}})^2}$

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\text{age}}, \ldots, \beta_4^{\text{age}}$
- Finite-population centering:

$$
\begin{aligned}
\tilde{\beta}_j^{\text{age}} &= \beta_j^{\text{age}} - \bar{\beta}^{\text{age}}, \text{ for } j = 1, \ldots, 4 \\
\tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\text{age}}
\end{aligned}
$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects
  - $\beta_j^{\text{age}} \sim \text{N}(0, \sigma_{\text{age}}^2)$, for j=1,...,4
  - Superpopulation sd: σ_{age}
  - Finite-population sd: $\sqrt{\frac{1}{3}\sum_{j=1}^4(\beta_j^{\text{age}} - \bar{\beta}^{\text{age}})^2}$

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\mathrm{age}}, \ldots, \beta_4^{\mathrm{age}}$
- Finite-population centering:

$$\begin{aligned}\tilde{\beta}_j^{\mathrm{age}} &= \beta_j^{\mathrm{age}} - \bar{\beta}^{\mathrm{age}}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\mathrm{age}}\end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects
  - $\beta_j^{\mathrm{age}} \sim \mathrm{N}(0, \sigma_{\mathrm{age}}^2)$, for j=1,...,4
  - Superpopulation sd: σ_{age}
  - Finite-population sd: $\sqrt{\frac{1}{3}\sum_{j=1}^4(\beta_j^{\mathrm{age}} - \bar{\beta}^{\mathrm{age}})^2}$

Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\rm age}, \ldots, \beta_4^{\rm age}$
- Finite-population centering:

$$\begin{aligned}\tilde{\beta}_j^{\rm age} &= \beta_j^{\rm age} - \bar{\beta}^{\rm age}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\rm age}\end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects
  - $\beta_j^{\rm age} \sim \mathrm{N}(0, \sigma^2_{\rm age})$, for j=1,...,4
  - Superpopulation sd: $\sigma_{\rm age}$
  - Finite-population sd: $\sqrt{\frac{1}{3}\sum_{j=1}^4 (\beta_j^{\rm age} - \bar{\beta}^{\rm age})^2}$

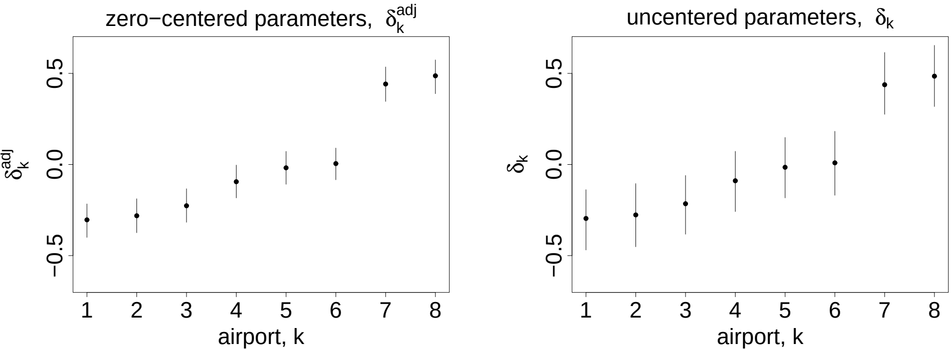
Finite-population and superpopulation estimands

- Consider the 4 coefficients, $\beta_1^{\rm age}, \ldots, \beta_4^{\rm age}$
- Finite-population centering:

$$\begin{aligned} \tilde{\beta}_j^{\rm age} &= \beta_j^{\rm age} - \bar{\beta}^{\rm age}, \text{ for } j = 1, \ldots, 4 \\ \tilde{\beta}_0 &= \beta_0 + \bar{\beta}^{\rm age} \end{aligned}$$

- Adjusted parameters are more precisely estimated
- Especially when # of groups is small
- Sd of group effects
  - $\beta_j^{\rm age} \sim {\rm N}(0, \sigma^2_{\rm age})$, for j=1,...,4
  - Superpopulation sd: $\sigma_{\rm age}$
  - Finite-population sd: $\sqrt{\frac{1}{3}\sum_{j=1}^4(\beta_j^{\rm age} - \bar{\beta}^{\rm age})^2}$

Example of finite-pop and superpop ests



Redundant parameterization

- Data model: $\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)$
- Usual model for the coefficients:

$$\begin{aligned}\beta_j^{\text{age}} &\sim \text{N}(0, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4\\ \beta_j^{\text{state}} &\sim \text{N}(0, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50\end{aligned}$$

- Additively redundant model:

$$\begin{aligned}\beta_j^{\text{age}} &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4\\ \beta_j^{\text{state}} &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50\end{aligned}$$

- Why add the redundant $\mu_{\text{age}}, \mu_{\text{state}}$?

Redundant parameterization

- Data model: $\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)$
- Usual model for the coefficients:

$$\begin{aligned} \beta^{\text{age}}_j &\sim \text{N}(0, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(0, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Additively redundant model:

$$\begin{aligned} \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Why add the redundant $\mu_{\text{age}}, \mu_{\text{state}}$?

Redundant parameterization

- Data model: $\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)$
- Usual model for the coefficients:

$$\begin{aligned} \beta^{\text{age}}_j &\sim \text{N}(0, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \dots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(0, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \dots, 50 \end{aligned}$$

- Additively redundant model:

$$\begin{aligned} \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \dots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \dots, 50 \end{aligned}$$

- Why add the redundant $\mu_{\text{age}}, \mu_{\text{state}}$?
  - Iterative algorithm moves more smoothly

Redundant parameterization

- Data model: $\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)$
- Usual model for the coefficients:

$$\begin{aligned} \beta_j^{\text{age}} &\sim \text{N}(0, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta_j^{\text{state}} &\sim \text{N}(0, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Additively redundant model:

$$\begin{aligned} \beta_j^{\text{age}} &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta_j^{\text{state}} &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Why add the redundant $\mu_{\text{age}}, \mu_{\text{state}}$?
  - Iterative algorithm moves more smoothly

Redundant parameterization

- Data model: $\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)$
- Usual model for the coefficients:

$$\begin{aligned} \beta^{\text{age}}_j &\sim \text{N}(0, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(0, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Additively redundant model:

$$\begin{aligned} \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Why add the redundant $\mu_{\text{age}}, \mu_{\text{state}}$?
  - Iterative algorithm moves more smoothly

Redundant parameterization

- Data model: $\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)$
- Usual model for the coefficients:

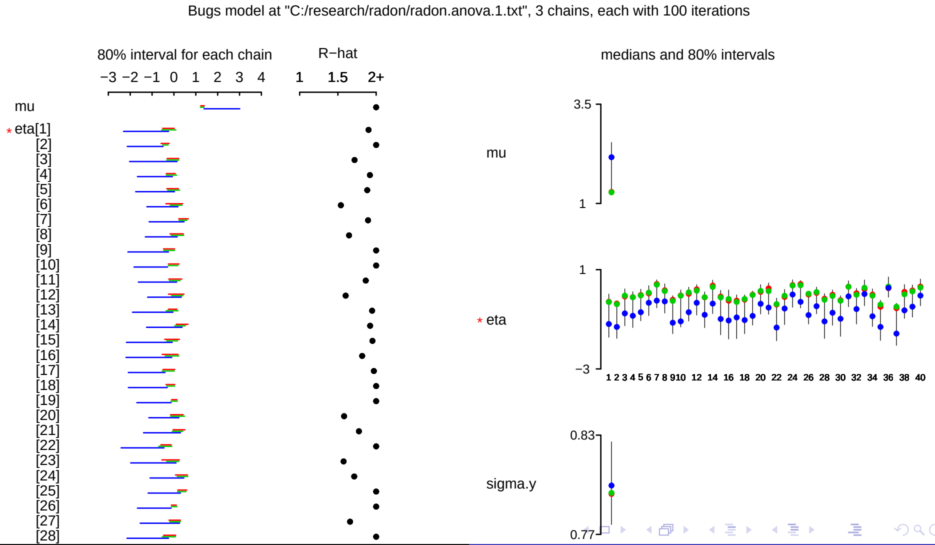
$$\begin{aligned}\beta^{\text{age}}_j &\sim \text{N}(0, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \dots, 4\\ \beta^{\text{state}}_j &\sim \text{N}(0, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \dots, 50\end{aligned}$$

- Additively redundant model:

$$\begin{aligned}\beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \dots, 4\\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \dots, 50\end{aligned}$$

- Why add the redundant $\mu_{\text{age}}, \mu_{\text{state}}$?
  - Iterative algorithm moves more smoothly

Motivation for redundant parameterization



Redundant additive parameterization

- Model

$$\begin{aligned}\Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right)\\ \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \dots, 4\\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \dots, 50\end{aligned}$$

- Identify using centered parameters:

$$\begin{aligned}\tilde{\beta}^{\text{age}}_j &= \beta^{\text{age}}_j - \bar{\beta}^{\text{age}}, \quad \text{for } j = 1, \dots, 4\\ \tilde{\beta}^{\text{state}}_j &= \beta^{\text{state}}_j - \bar{\beta}^{\text{state}}, \quad \text{for } j = 1, \dots, 50\end{aligned}$$

- Redefine the constant term:

$$\tilde{\beta}^0 = \beta^0 + \bar{\beta}^{\text{age}} + \bar{\beta}^{\text{age}}$$

Redundant additive parameterization

- Model

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right) \\ \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Identify using centered parameters:

$$\begin{aligned} \tilde{\beta}^{\text{age}}_j &= \beta^{\text{age}}_j - \bar{\beta}^{\text{age}}, \quad \text{for } j = 1, \ldots, 4 \\ \tilde{\beta}^{\text{state}}_j &= \beta^{\text{state}}_j - \bar{\beta}^{\text{state}}, \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Redefine the constant term:

$$\tilde{\beta}^0 = \beta^0 + \bar{\beta}^{\text{age}} + \bar{\beta}^{\text{age}}$$

Redundant additive parameterization

- Model

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right) \\ \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \dots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \dots, 50 \end{aligned}$$

- Identify using centered parameters:

$$\begin{aligned} \tilde{\beta}^{\text{age}}_j &= \beta^{\text{age}}_j - \bar{\beta}^{\text{age}}, \quad \text{for } j = 1, \dots, 4 \\ \tilde{\beta}^{\text{state}}_j &= \beta^{\text{state}}_j - \bar{\beta}^{\text{state}}, \quad \text{for } j = 1, \dots, 50 \end{aligned}$$

- Redefine the constant term:

$$\tilde{\beta}^0 = \beta^0 + \bar{\beta}^{\text{age}} + \bar{\beta}^{\text{age}}$$

Redundant additive parameterization

- Model

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \beta^{\text{age}}_{\text{age(i)}} + \beta^{\text{state}}_{\text{state(i)}}\right) \\ \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Identify using centered parameters:

$$\begin{aligned} \tilde\beta^{\text{age}}_j &= \beta^{\text{age}}_j - \bar\beta^{\text{age}}, \quad \text{for } j = 1, \ldots, 4 \\ \tilde\beta^{\text{state}}_j &= \beta^{\text{state}}_j - \bar\beta^{\text{state}}, \quad \text{for } j = 1, \ldots, 50 \end{aligned}$$

- Redefine the constant term:

$$\tilde\beta^0 = \beta^0 + \bar\beta^{\text{age}} + \bar\beta^{\text{age}}$$

Redundant multiplicative parameterization

- New model

$$\begin{aligned}\Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \xi^{\text{age}}\beta^{\text{age}}_{\text{age(i)}} + \xi^{\text{state}}\beta^{\text{state}}_{\text{state(i)}}\right)\\ \beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}),\quad \text{for } j = 1,\ldots,4\\ \beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}),\quad \text{for } j = 1,\ldots,50\end{aligned}$$

- Identify using centered and scaled parameters:

$$\begin{aligned}\tilde{\beta}^{\text{age}}_j &= \xi^{\text{age}}(\beta^{\text{age}}_j - \bar{\beta}^{\text{age}}),\quad \text{for } j = 1,\ldots,4\\ \tilde{\beta}^{\text{state}}_j &= \xi^{\text{state}}\left(\beta^{\text{state}}_j - \bar{\beta}^{\text{state}}\right),\quad \text{for } j = 1,\ldots,50\end{aligned}$$

- Faster convergence
- More general model, connections to factor analysis

Redundant multiplicative parameterization

- New model

$$\begin{aligned}
\Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \xi^{\text{age}}\beta^{\text{age}}_{\text{age(i)}} + \xi^{\text{state}}\beta^{\text{state}}_{\text{state(i)}}\right) \\
\beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Identify using centered and scaled parameters:

$$\begin{aligned}
\tilde{\beta}^{\text{age}}_j &= \xi^{\text{age}}(\beta^{\text{age}}_j - \bar{\beta}^{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\tilde{\beta}^{\text{state}}_j &= \xi^{\text{state}}\left(\beta^{\text{state}}_j - \bar{\beta}^{\text{state}}\right), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Faster convergence
- More general model, connections to factor analysis

Redundant multiplicative parameterization

- New model

$$\begin{aligned}
\Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \xi^{\text{age}} \beta^{\text{age}}_{\text{age(i)}} + \xi^{\text{state}} \beta^{\text{state}}_{\text{state(i)}}\right) \\
\beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Identify using centered and scaled parameters:

$$\begin{aligned}
\tilde{\beta}^{\text{age}}_j &= \xi^{\text{age}}(\beta^{\text{age}}_j - \bar{\beta}^{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\tilde{\beta}^{\text{state}}_j &= \xi^{\text{state}}\left(\beta^{\text{state}}_j - \bar{\beta}^{\text{state}}\right), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Faster convergence
- More general model, connections to factor analysis

Redundant multiplicative parameterization

- New model

$$\begin{aligned}
\Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \xi^{\text{age}}\beta^{\text{age}}_{\text{age(i)}} + \xi^{\text{state}}\beta^{\text{state}}_{\text{state(i)}}\right) \\
\beta^{\text{age}}_j &\sim \mathrm{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\beta^{\text{state}}_j &\sim \mathrm{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Identify using centered and scaled parameters:

$$\begin{aligned}
\tilde{\beta}^{\text{age}}_j &= \xi^{\text{age}}(\beta^{\text{age}}_j - \bar{\beta}^{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\tilde{\beta}^{\text{state}}_j &= \xi^{\text{state}}\left(\beta^{\text{state}}_j - \bar{\beta}^{\text{state}}\right), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Faster convergence
- More general model, connections to factor analysis

Redundant multiplicative parameterization

- New model

$$\begin{aligned}
\Pr(y_i = 1) &= \text{logit}^{-1}\left(\beta^0 + \xi^{\text{age}} \beta^{\text{age}}_{\text{age(i)}} + \xi^{\text{state}} \beta^{\text{state}}_{\text{state(i)}}\right) \\
\beta^{\text{age}}_j &\sim \text{N}(\mu_{\text{age}}, \sigma^2_{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\beta^{\text{state}}_j &\sim \text{N}(\mu_{\text{state}}, \sigma^2_{\text{state}}), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Identify using centered and scaled parameters:

$$\begin{aligned}
\tilde{\beta}^{\text{age}}_j &= \xi^{\text{age}}(\beta^{\text{age}}_j - \bar{\beta}^{\text{age}}), \quad \text{for } j = 1, \ldots, 4 \\
\tilde{\beta}^{\text{state}}_j &= \xi^{\text{state}}\left(\beta^{\text{state}}_j - \bar{\beta}^{\text{state}}\right), \quad \text{for } j = 1, \ldots, 50
\end{aligned}$$

- Faster convergence
- More general model, connections to factor analysis

MLM and partial pooling

- Goal is to more accurately estimate coefficients that are grouped
- A reparameterization can change a model
(even if it leaves the likelihood unchanged)
- Redundant additive parameterization
- Redundant multiplicative parameterization
- Weakly-informative prior distribution for group-level variance parameters

MLM and partial pooling

- Goal is to more accurately estimate coefficients that are grouped
- A reparameterization can change a model (even if it leaves the likelihood unchanged)
- Redundant additive parameterization
- Redundant multiplicative parameterization
- Weakly-informative prior distribution for group-level variance parameters

MLM and partial pooling

- Goal is to more accurately estimate coefficients that are grouped
- A reparameterization can change a model (even if it leaves the likelihood unchanged)
- Redundant additive parameterization
- Redundant multiplicative parameterization
- Weakly-informative prior distribution for group-level variance parameters

MLM and partial pooling

- Goal is to more accurately estimate coefficients that are grouped
- A reparameterization can change a model (even if it leaves the likelihood unchanged)
- Redundant additive parameterization
- Redundant multiplicative parameterization
- Weakly-informative prior distribution for group-level variance parameters

MLM and partial pooling

- Goal is to more accurately estimate coefficients that are grouped
- A reparameterization can change a model
  (even if it leaves the likelihood unchanged)
- Redundant additive parameterization
- Redundant multiplicative parameterization
- Weakly-informative prior distribution for group-level variance parameters

Displaying and summarizing inferences

- Displaying parameters in groups rather than as a long list
- Analysis of variance
- Average predictive effects
- R^2 and partial pooling factors

Displaying and summarizing inferences

- Displaying parameters in groups rather than as a long list
- Analysis of variance
- Average predictive effects
- R^2 and partial pooling factors

Displaying and summarizing inferences

- Displaying parameters in groups rather than as a long list
- Analysis of variance
- Average predictive effects
- R^2 and partial pooling factors

Displaying and summarizing inferences

- Displaying parameters in groups rather than as a long list
- Analysis of variance
- Average predictive effects
- R^2 and partial pooling factors

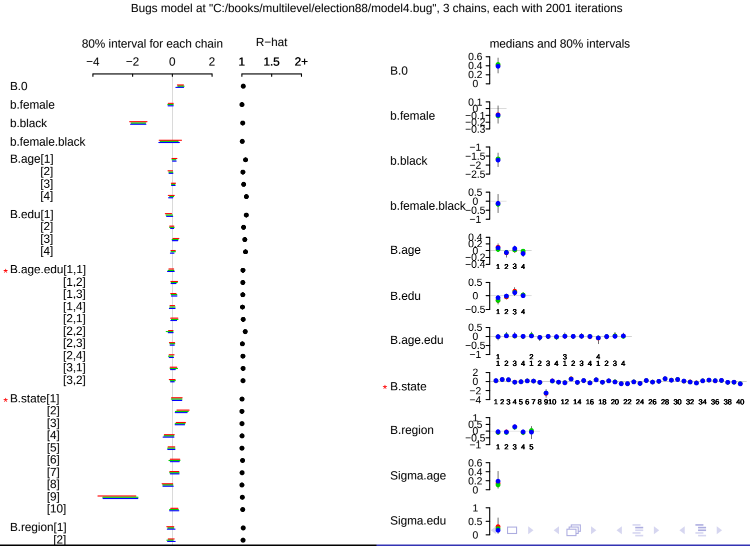
Displaying and summarizing inferences

- Displaying parameters in groups rather than as a long list
- Analysis of variance
- Average predictive effects
- R^2 and partial pooling factors

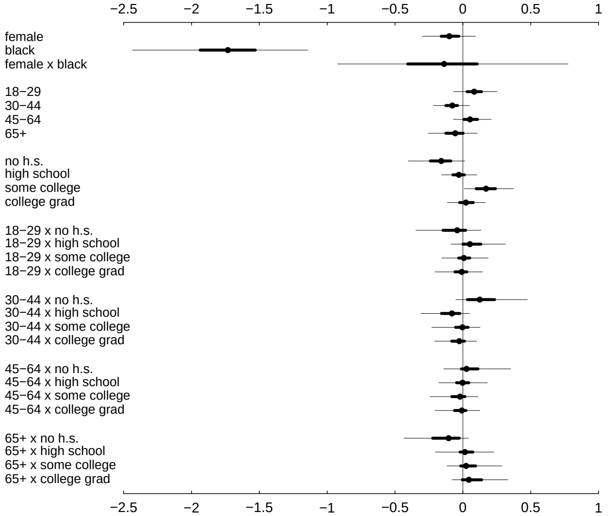
Raw display of inference

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
B.0	0.402	0.147	0.044	0.326	0.413	0.499	0.652	1.024	110
b.female	-0.094	0.102	-0.283	-0.162	-0.095	-0.034	0.107	1.001	1000
b.black	-1.701	0.305	-2.323	-1.910	-1.691	-1.486	-1.152	1.014	500
b.female.black	-0.143	0.393	-0.834	-0.383	-0.155	0.104	0.620	1.007	1000
B.age[1]	0.084	0.088	-0.053	0.012	0.075	0.140	0.277	1.062	45
B.age[2]	-0.072	0.087	-0.260	-0.121	-0.054	-0.006	0.052	1.017	190
B.age[3]	0.044	0.077	-0.105	-0.007	0.038	0.095	0.203	1.029	130
B.age[4]	-0.057	0.096	-0.265	-0.115	-0.052	0.001	0.133	1.076	32
B.edu[1]	-0.148	0.131	-0.436	-0.241	-0.137	-0.044	0.053	1.074	31
B.edu[2]	-0.022	0.082	-0.182	-0.069	-0.021	0.025	0.152	1.028	160
B.edu[3]	0.148	0.112	-0.032	0.065	0.142	0.228	0.370	1.049	45
B.edu[4]	0.023	0.090	-0.170	-0.030	0.015	0.074	0.224	1.061	37
B.age.edu[1,1]	-0.044	0.133	-0.363	-0.104	-0.019	0.025	0.170	1.018	1000
B.age.edu[1,2]	0.059	0.123	-0.153	-0.011	0.032	0.118	0.353	1.016	580
B.age.edu[1,3]	0.049	0.124	-0.146	-0.023	0.022	0.104	0.349	1.015	280
B.age.edu[1,4]	0.001	0.116	-0.237	-0.061	0.000	0.052	0.280	1.010	1000
B.age.edu[2,1]	0.066	0.152	-0.208	-0.008	0.032	0.124	0.449	1.022	140
B.age.edu[2,2]	-0.081	0.127	-0.407	-0.137	-0.055	0.001	0.094	1.057	120
B.age.edu[2,3]	-0.004	0.102	-0.226	-0.048	0.000	0.041	0.215	1.008	940
B.age.edu[2,4]	-0.042	0.108	-0.282	-0.100	-0.026	0.014	0.157	1.017	170
B.age.edu[3,1]	0.034	0.135	-0.215	-0.030	0.009	0.091	0.361	1.021	230
B.age.edu[3,2]	0.007	0.102	-0.213	-0.039	0.003	0.052	0.220	1.019	610
B.age.edu[3,3]	0.033	0.130	-0.215	-0.029	0.009	0.076	0.410	1.080	61
B.age.edu[3,4]	-0.009	0.109	-0.236	-0.064	-0.005	0.043	0.214	1.024	150
B.age.edu[4,1]	-0.141	0.190	-0.672	-0.224	-0.086	-0.003	0.100	1.036	270
B.age.edu[4,2]	-0.014	0.119	-0.280	-0.059	-0.008	0.033	0.239	1.017	240
B.age.edu[4,3]	0.046	0.132	-0.192	-0.024	0.019	0.108	0.332	1.010	210
B.age.edu[4,4]	0.042	0.142	-0.193	-0.022	0.016	0.095	0.377	1.015	160

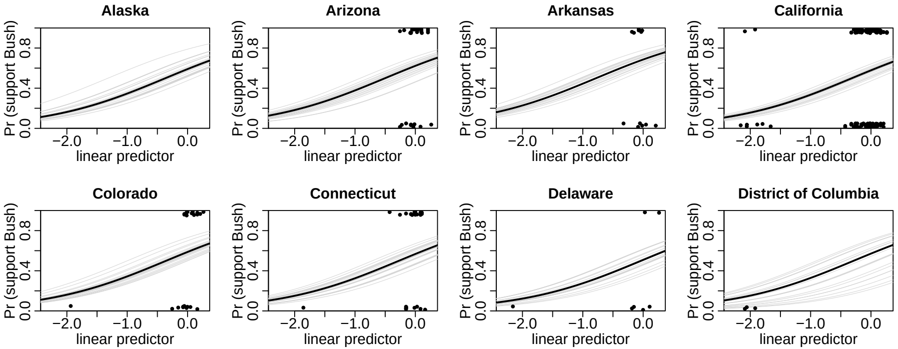
Raw graphical display



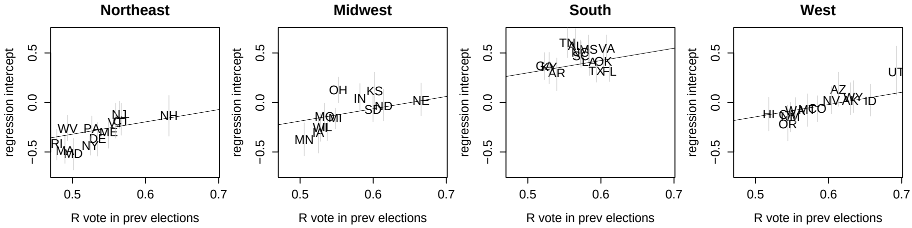
Better graphical display 1: demographics



Better graphical display 2: within states



Better graphical display 3: between states



Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
    Estimating and comparing variance components
    Not testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
  - Estimating and comparing variance components
  - Not testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
  - Estimating and comparing variance components
  - *Not* testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
  - Estimating and comparing variance components
  - *Not* testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
  - Estimating and comparing variance components
  - *Not* testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
  - Estimating and comparing variance components
  - *Not* testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

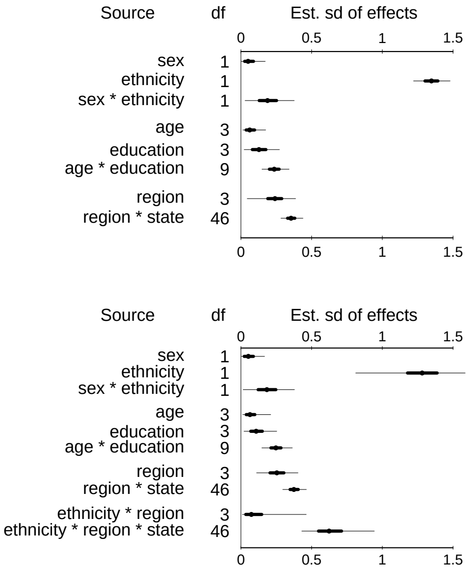
Anova and multilevel models

- Each row of the Anova table is a variance component
- Goal
  - How important is each source of variation?
  - Estimating and comparing variance components
  - *Not* testing if a variance component equals 0
- Multilevel regression solves classical Anova problems

Raw display of inference

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
B.0	0.402	0.147	0.044	0.326	0.413	0.499	0.652	1.024	110
b.female	-0.094	0.102	-0.283	-0.162	-0.095	-0.034	0.107	1.001	1000
b.black	-1.701	0.305	-2.323	-1.910	-1.691	-1.486	-1.152	1.014	500
b.female.black	-0.143	0.393	-0.834	-0.383	-0.155	0.104	0.620	1.007	1000
B.age[1]	0.084	0.088	-0.053	0.012	0.075	0.140	0.277	1.062	45
B.age[2]	-0.072	0.087	-0.260	-0.121	-0.054	-0.006	0.052	1.017	190
B.age[3]	0.044	0.077	-0.105	-0.007	0.038	0.095	0.203	1.029	130
B.age[4]	-0.057	0.096	-0.265	-0.115	-0.052	0.001	0.133	1.076	32
B.edu[1]	-0.148	0.131	-0.436	-0.241	-0.137	-0.044	0.053	1.074	31
B.edu[2]	-0.022	0.082	-0.182	-0.069	-0.021	0.025	0.152	1.028	160
B.edu[3]	0.148	0.112	-0.032	0.065	0.142	0.228	0.370	1.049	45
B.edu[4]	0.023	0.090	-0.170	-0.030	0.015	0.074	0.224	1.061	37
B.age.edu[1,1]	-0.044	0.133	-0.363	-0.104	-0.019	0.025	0.170	1.018	1000
B.age.edu[1,2]	0.059	0.123	-0.153	-0.011	0.032	0.118	0.353	1.016	580
B.age.edu[1,3]	0.049	0.124	-0.146	-0.023	0.022	0.104	0.349	1.015	280
B.age.edu[1,4]	0.001	0.116	-0.237	-0.061	0.000	0.052	0.280	1.010	1000
B.age.edu[2,1]	0.066	0.152	-0.208	-0.008	0.032	0.124	0.449	1.022	140
B.age.edu[2,2]	-0.081	0.127	-0.407	-0.137	-0.055	0.001	0.094	1.057	120
B.age.edu[2,3]	-0.004	0.102	-0.226	-0.048	0.000	0.041	0.215	1.008	940
B.age.edu[2,4]	-0.042	0.108	-0.282	-0.100	-0.026	0.014	0.157	1.017	170
B.age.edu[3,1]	0.034	0.135	-0.215	-0.030	0.009	0.091	0.361	1.021	230
B.age.edu[3,2]	0.007	0.102	-0.213	-0.039	0.003	0.052	0.220	1.019	610
B.age.edu[3,3]	0.033	0.130	-0.215	-0.029	0.009	0.076	0.410	1.080	61
B.age.edu[3,4]	-0.009	0.109	-0.236	-0.064	-0.005	0.043	0.214	1.024	150
B.age.edu[4,1]	-0.141	0.190	-0.672	-0.224	-0.086	-0.003	0.100	1.036	270
B.age.edu[4,2]	-0.014	0.119	-0.280	-0.059	-0.008	0.033	0.239	1.017	240
B.age.edu[4,3]	0.046	0.132	-0.192	-0.024	0.019	0.108	0.332	1.010	210
B.age.edu[4,4]	0.042	0.142	-0.193	-0.022	0.016	0.095	0.377	1.015	160

Bayesian Anova



Fixed and random effects?

- What are "fixed" and "random" effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1980)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are estimated using shrinkage (Snijders, 1999)

Fixed and random effects?

- What are "fixed" and "random" effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1960)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are esitmated using shrinkage (Snijders, 1999)

Fixed and random effects?

- What are "fixed" and "random" effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1960)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are esitmated using shrinkage (Snijders, 1999)

Fixed and random effects?

- What are "fixed" and "random" effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1960)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are esitmated using shrinkage (Snijders, 1999)

Fixed and random effects?

- What are “fixed” and “random” effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1960)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are esitmated using shrinkage (Snijders, 1999)

Fixed and random effects?

- What are "fixed" and "random" effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1960)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are esitmated using shrinkage (Snijders, 1999)

Fixed and random effects?

- What are "fixed" and "random" effects?
- Five incompatible definitions:
  1. Fixed effects are constant across individuals; random effects vary (Leeuw, 1998)
  2. Effects are fixed if they are interesting in themselves, random if you care about the population (Searle, 1992)
  3. Fixed effects are the entire population, random are a small sample from a larger population (Tukey, 1960)
  4. Random effects are realized values of a random variable (LaMotte, 1983)
  5. Fixed effects are estimated using least squares, random effects are esitmated using shrinkage (Snijders, 1999)

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then find they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

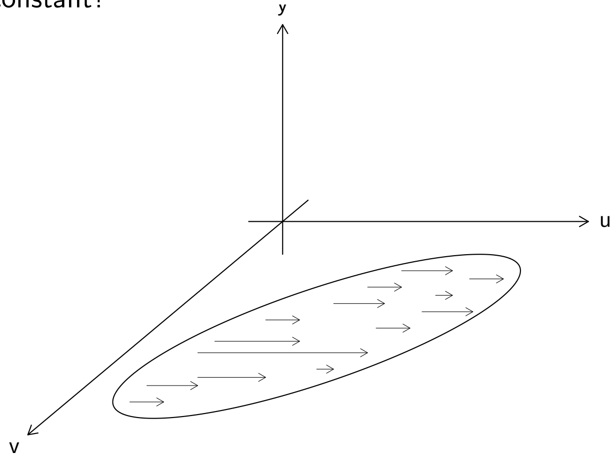
- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

How to think about fixed and random effects

- Ideally, allow all coefficients to vary by group
  - Main limitation: complicated models can be overwhelming
- Bayesian multilevel modeling
  - Simultaneously estimate population parameters and individual coefficients
  - Suppose you are estimating a finite set of effects, then told they are a sample from a larger population
  - No need to change the model
  - But estimand of interest might change!
- Separation of modeling, inference, and decision analysis

Average predictive effects

- What is $E(y \mid x_1 = \text{high}) - \text{E}(\text{y} \mid \text{x}_1 = \text{low})$, with all other x's held constant?



Average predictive effects

- What is $E(y \mid x_1 = \text{high}) - \text{E}(\text{y} \mid \text{x}_1 = \text{low})$, with all other x's held constant?
- In general, difference can depend on x
- Average over distribution of x in the data
  - You can't just use a central value of x
- Compute APE for each input variable x
- Multilevel factors are categorical input variables

Average predictive effects

- What is $E(y \,|\, x_1 = \mathrm{high}) - \mathrm{E}(\mathrm{y} \,|\, \mathrm{x}_1 = \mathrm{low})$, with all other x's held constant?
- In general, difference can depend on x
- Average over distribution of x in the data
  - You can't just use a central value of x
- Compute APE for each input variable x
- Multilevel factors are categorical input variables

Average predictive effects

- What is $E(y \mid x_1 = \text{high}) - \mathrm{E}(\mathrm{y} \mid \mathrm{x}_1 = \mathrm{low})$, with all other x's held constant?
- In general, difference can depend on x
- Average over distribution of x in the data
  - You can't just use a central value of x
- Compute APE for each input variable x
- Multilevel factors are categorical input variables

Average predictive effects

- What is $E(y \,|\, x_1 = \mathrm{high}) - \mathrm{E(y \,|\, x_1 = low)}$, with all other x's held constant?
- In general, difference can depend on x
- Average over distribution of x in the data
  - You can't just use a central value of x
- Compute APE for each input variable x
- Multilevel factors are categorical input variables

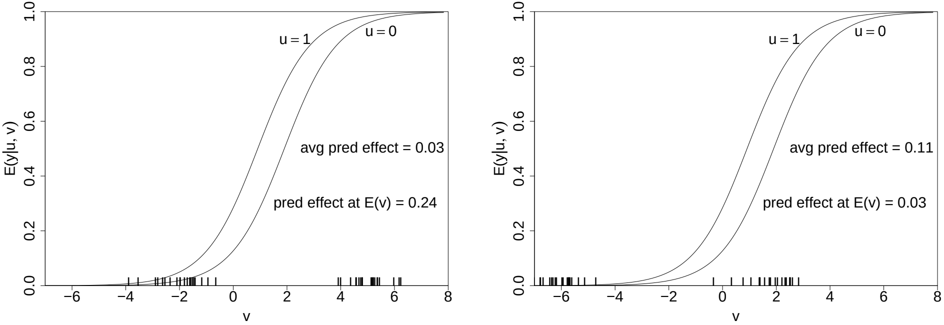
Average predictive effects

- What is $E(y \mid x_1 = \text{high}) - \mathrm{E}(\mathrm{y} \mid \mathrm{x}_1 = \text{low})$, with all other x's held constant?
- In general, difference can depend on x
- Average over distribution of x in the data
  - You can't just use a central value of x
- Compute APE for each input variable x
- Multilevel factors are categorical input variables

Average predictive effects

- What is $E(y \,|\, x_1 = \text{high}) - \text{E}(\text{y} \,|\, \text{x}_1 = \text{low})$, with all other x's held constant?
- In general, difference can depend on x
- Average over distribution of x in the data
  - You can't just use a central value of x
- Compute APE for each input variable x
- Multilevel factors are categorical input variables

APE: why you can't just use a central value of x



Framework for average predictive effects

- Regresion model, $E(y|x, \theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age × sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant

Framework for average predictive effects

- Regresion model, $E(y|x,\theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant

Framework for average predictive effects

- Regresion model, $E(y|x,\theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant

Framework for average predictive effects

- Regresion model, $E(y|x, \theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant
  - Scalar input u: the "input of interest"
    Vector v: all other inputs

Framework for average predictive effects

- Regresion model, $E(y|x, \theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant
  - Scalar input u: the "input of interest"
  - Vector v: all other inputs

Framework for average predictive effects

- Regresion model, $E(y|x, \theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant
  - Scalar input u: the "input of interest"
  - Vector v: all other inputs

Framework for average predictive effects

- Regresion model, $E(y|x, \theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant
  - Scalar input u: the "input of interest"
  - Vector v: all other inputs

Framework for average predictive effects

- Regresion model, $E(y|x, \theta)$
- Predictors come from "input variables"
  - Example: regression on age, sex, age $\times$ sex, and age^2
  - 5 linear *predictors* (including the constant term)
  - But only 4 *inputs*
- Compute APE for each input variable, one at a time, with all others held constant
  - Scalar input u: the "input of interest"
  - Vector v: all other inputs

Defining predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Average over:
  - The transition, $u^{(1)} \rightarrow u^{(2)}$
  - The other inputs, v
  - The regression coefficients, θ

Defining predictive effects

- predictive effect:
  $$\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$$
- Average over:
  - The transition, $u^{(1)} \rightarrow u^{(2)}$
  - The other inputs, v
  - The regression coefficients, θ

Defining predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Average over:
  - The transition, $u^{(1)} \rightarrow u^{(2)}$
  - The other inputs, v
  - The regression coefficients, θ

Defining predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Average over:
  - The transition, $u^{(1)} \rightarrow u^{(2)}$
  - The other inputs, v
  - The regression coefficients, θ

Defining predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Average over:
  - The transition, $u^{(1)} \rightarrow u^{(2)}$
  - The other inputs, v
  - The regression coefficients, θ

Defining predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Average over:
  - The transition, $u^{(1)} \rightarrow u^{(2)}$
  - The other inputs, v
  - The regression coefficients, θ

Average predictive effects for binary inputs

- predictive effect:
  $\delta_u(u^{(1)} \to u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Binary input u:
  - predictive effect: $\delta_u(0 \to 1, v, \theta) = E(y|1, v, \theta) - E(y|0, v, \theta)$
  - Average over $v_1, \ldots, v_n$ in the data (or weighted average if desired)
  - Average over θ from inferential simulations
  - Standard error of APE from uncertainty in θ

Average predictive effects for binary inputs

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Binary input u:
  - predictive effect: $\delta_u(0 \rightarrow 1, v, \theta) = E(y|1, v, \theta) - E(y|0, v, \theta)$
  - Average over $v_1, \ldots, v_n$ in the data (or weighted average if desired)
  - Average over θ from inferential simulations
  - Standard error of APE from uncertainty in θ

Average predictive effects for binary inputs

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Binary input u:
  - predictive effect: $\delta_u(0 \rightarrow 1, v, \theta) = E(y|1, v, \theta) - E(y|0, v, \theta)$
  - Average over $v_1, \ldots, v_n$ in the data (or weighted average if desired)
  - Average over θ from inferential simulations
  - Standard error of APE from uncertainty in θ

Average predictive effects for binary inputs

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta)-E(y|u^{(1)},v,\theta)}{u^{(2)}-u^{(1)}}$
- Binary input u:
  - predictive effect: $\delta_u(0 \rightarrow 1, v, \theta) = E(y|1, v, \theta) - E(y|0, v, \theta)$
  - Average over $v_1, \ldots, v_n$ in the data (or weighted average if desired)
  - Average over θ from inferential simulations
  - Standard error of APE from uncertainty in θ

Average predictive effects for binary inputs

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Binary input u:
  - predictive effect: $\delta_u(0 \rightarrow 1, v, \theta) = E(y|1, v, \theta) - E(y|0, v, \theta)$
  - Average over $v_1, \ldots, v_n$ in the data (or weighted average if desired)
  - Average over θ from inferential simulations
  - Standard error of APE from uncertainty in θ

Average predictive effects for binary inputs

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Binary input u:
  - predictive effect: $\delta_u(0 \rightarrow 1, v, \theta) = E(y|1, v, \theta) - E(y|0, v, \theta)$
  - Average over $v_1, \ldots, v_n$ in the data (or weighted average if desired)
  - Average over θ from inferential simulations
  - Standard error of APE from uncertainty in θ

Scenarios for average predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Continuous inputs
- Unordered discrete inputs
- Variance components
- Interactions
- Inputs that are not always active

Scenarios for average predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Continuous inputs
- Unordered discrete inputs
- Variance components
- Interactions
- Inputs that are not always active

Scenarios for average predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Continuous inputs
- Unordered discrete inputs
- Variance components
- Interactions
- Inputs that are not always active

Scenarios for average predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Continuous inputs
- Unordered discrete inputs
- Variance components
- Interactions
- Inputs that are not always active

Scenarios for average predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Continuous inputs
- Unordered discrete inputs
- Variance components
- Interactions
- Inputs that are not always active

Scenarios for average predictive effects

- predictive effect:
  $\delta_u(u^{(1)} \rightarrow u^{(2)}, v, \theta) = \frac{E(y|u^{(2)},v,\theta) - E(y|u^{(1)},v,\theta)}{u^{(2)} - u^{(1)}}$
- Continuous inputs
- Unordered discrete inputs
- Variance components
- Interactions
- Inputs that are not always active

R^2 for multilevel models

- How much of the variance is "explained" by the model?
- Separate R^2 for each level
- Classical $R^2 = 1 - \frac{\text{variance of the residuals}}{\text{variance of the data}}$
- Multilevel model:
  at each level, k units: $\theta_k = (X\beta)_k + \epsilon_k$
- At each level: $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$

R^2 for multilevel models

- How much of the variance is "explained" by the model?
- Separate R^2 for each level
- Classical $R^2 = 1 - \frac{\text{variance of the residuals}}{\text{variance of the data}}$
- Multilevel model:
  at each level, k units: $\theta_k = (X\beta)_k + \epsilon_k$
- At each level: $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$

R^2 for multilevel models

- How much of the variance is "explained" by the model?
- Separate R^2 for each level
- Classical $R^2 = 1 - \frac{\text{variance of the residuals}}{\text{variance of the data}}$
- Multilevel model:
  at each level, k units: $\theta_k = (X\beta)_k + \epsilon_k$
- At each level: $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$

R^2 for multilevel models

- How much of the variance is "explained" by the model?
- Separate R^2 for each level
- Classical $R^2 = 1 - \frac{\text{variance of the residuals}}{\text{variance of the data}}$
- Multilevel model:
  at each level, k units: $\theta_k = (X\beta)_k + \epsilon_k$
- At each level: $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$

R^2 for multilevel models

- How much of the variance is "explained" by the model?
- Separate R^2 for each level
- Classical $R^2 = 1 - \frac{\text{variance of the residuals}}{\text{variance of the data}}$
- Multilevel model:
  at each level, k units: $\theta_k = (X\beta)_k + \epsilon_k$
- At each level: $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$

R^2 for multilevel models

- How much of the variance is "explained" by the model?
- Separate R^2 for each level
- Classical $R^2 = 1 - \frac{\text{variance of the residuals}}{\text{variance of the data}}$
- Multilevel model:
  at each level, k units: $\theta_k = (X\beta)_k + \epsilon_k$
- At each level: $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$

Bayesian R^2

- At each level
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$
- Numerator and denominator estimated by their posterior means
- Posterior distribution automatically accounts for uncertainty
- Bayesian generalization of classical "adjusted R^2"

Bayesian R^2

- At each level
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$
- Numerator and denominator estimated by their posterior means
- Posterior distribution automatically accounts for uncertainty
- Bayesian generalization of classical "adjusted R^2"

Bayesian R^2

- At each level
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$
- Numerator and denominator estimated by their posterior means
- Posterior distribution automatically accounts for uncertainty
- Bayesian generalization of classical "adjusted R^2"

Bayesian R^2

- At each level
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$
- Numerator and denominator estimated by their posterior means
- Posterior distribution automatically accounts for uncertainty
- Bayesian generalization of classical "adjusted R^2"

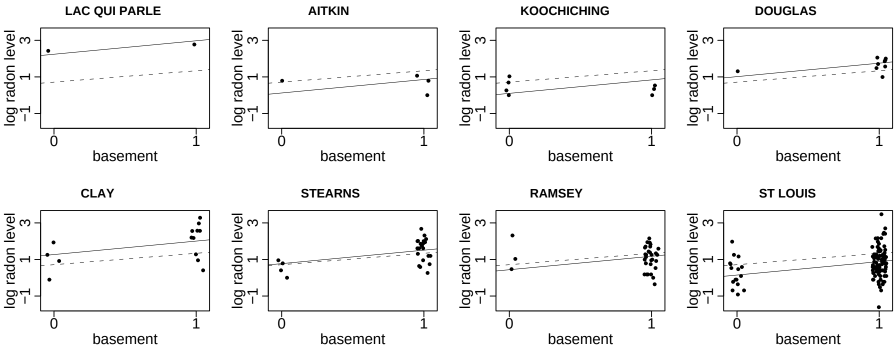
Bayesian R^2

- At each level
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$
- Numerator and denominator estimated by their posterior means
- Posterior distribution automatically accounts for uncertainty
- Bayesian generalization of classical "adjusted R^2"

Bayesian R^2

- At each level
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $R^2 = 1 - \frac{\text{variance among the } (X\beta)_k\text{'s}}{\text{variance among the } \epsilon_k\text{'s}}$
- Numerator and denominator estimated by their posterior means
- Posterior distribution automatically accounts for uncertainty
- Bayesian generalization of classical "adjusted R^2"

Example of partial pooling



Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
  - Can't simply compare to the "complete pooling" and "no pooling" estimates
    "No pooling" estimate doesn't always exist!
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
  - Can't simply compare to the "complete pooling" and "no pooling" estimates
  - "No pooling" estimate doesn't always exist!
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
  - Can't simply compare to the "complete pooling" and "no pooling" estimates
  - "No pooling" estimate doesn't always exist!
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
  - Can't simply compare to the "complete pooling" and "no pooling" estimates
  - "No pooling" estimate doesn't always exist!
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
  - Can't simply compare to the "complete pooling" and "no pooling" estimates
  - "No pooling" estimate doesn't always exist!
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Partial pooling factors

- At each level of the model:
  - $\theta_k = (X\beta)_k + \epsilon_k$
  - $\lambda = 0$ if no pooling of ϵ's
  - $\lambda = 0$ if complete pooling of ϵ's to 0
- Multilevel generalization of Bayesian pooling factor
  - Can't simply compare to the "complete pooling" and "no pooling" estimates
  - "No pooling" estimate doesn't always exist!
- At each level, our pooling factor is defined based on the mean and variance of the ϵ_k's

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance

- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null hypothesis testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Understanding sources of variation

- Graphs, not tables, of parameter estimates
- In display, use the grouping info
- Analysis of variance
  - Summarize the scale of each batch of predictors
  - Go beyond classical null-hypothesis-testing framework
- Average predictive effects for models with nonlinearity and interactions
- Generalization of R^2 (explained variance), defined at each level of the model
- Partial pooling factor, defined at each level

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building

Multilevel models even when # groups is small

- Forget about "fixed and random effects"; think about "finite-pop and superpop estimands" instead
- Always use the multilevel model, but estimand of interest depends on context
- For small # groups: use the new half-t prior dist for variance parameters
- Challenges:
  - Multivariate models (for example, varying-intercept, varying-slope models)
  - Models with deep interaction structures
  - Automatic graphical display
  - Model building