

Why Using Multilevel Models

Song S. Qian

Nicholas School of the Environment and Earth Sciences
Duke University

October 8, 2008

- 1 Introduction
- 2 Effects of Urbanization on Stream Ecosystems
- 3 Partial Pooling Data
 - Group Level Predictor
- 4 Why Multilevel Modeling

What is Multilevel Modeling?

- Multilevel structure in data

What is Multilevel Modeling?

- Multilevel structure in data
 - Data are grouped – by species, by community, by ecosystem

What is Multilevel Modeling?

- Multilevel structure in data
 - Data are grouped – by species, by community, by ecosystem
 - Information about the subject has “levels” – on individual, on species, on community, etc

What is Multilevel Modeling?

- Multilevel structure in data
 - Data are grouped – by species, by community, by ecosystem
 - Information about the subject has “levels” – on individual, on species, on community, etc
 - Lakes in Finland – observations from a given lake, lakes in a group

What is Multilevel Modeling?

- Multilevel structure in data
 - Data are grouped – by species, by community, by ecosystem
 - Information about the subject has “levels” – on individual, on species, on community, etc
 - Lakes in Finland – observations from a given lake, lakes in a group
- Regression approaches – varying intercepts and/or varying slopes by using factor predictor(s)

Some Notations

- Varying intercept model: $y_i = \beta_{0j[i]} + \beta_1 x_i + \epsilon_i$

Some Notations

- Varying intercept model: $y_i = \beta_{0j[i]} + \beta_1 x_i + \epsilon_i$
 - i – index for observations

Some Notations

- Varying intercept model: $y_i = \beta_{0j[i]} + \beta_1 x_i + \epsilon_i$
 - i – index for observations
 - $j[i]$ – indicating that the i th observation belongs to j th group

Some Notations

- Varying intercept model: $y_i = \beta_{0j[i]} + \beta_1 x_i + \epsilon_i$
 - i – index for observations
 - $j[i]$ – indicating that the i th observation belongs to j th group
 - $\beta_{0j[i]}$ – intercept varies by group

Some Notations

- Varying intercept model: $y_i = \beta_{0j[i]} + \beta_1 x_i + \epsilon_i$
 - i – index for observations
 - $j[i]$ – indicating that the i th observation belongs to j th group
 - $\beta_{0j[i]}$ – intercept varies by group
- Varying slope model: $y_i = \beta_0 + \beta_{1j[i]} x_i + \epsilon_i$

Some Notations

- Varying intercept model: $y_i = \beta_{0j[i]} + \beta_1 x_i + \epsilon_i$
 - i – index for observations
 - $j[i]$ – indicating that the i th observation belongs to j th group
 - $\beta_{0j[i]}$ – intercept varies by group
- Varying slope model: $y_i = \beta_0 + \beta_{1j[i]} x_i + \epsilon_i$
- Varying intercept, varying slope model:
 $y_i = \beta_{0j[i]} + \beta_{1j[i]} x_i + \epsilon_i$

Data Structure

- Individual level data, e.g., individual lake data

Data Structure

- Individual level data, e.g., individual lake data
- Group level data, e.g., climate of a region

Data Structure

- Individual level data, e.g., individual lake data
- Group level data, e.g., climate of a region
- Often we have separate data matrices for individual level data

Data Structure

- Individual level data, e.g., individual lake data
- Group level data, e.g., climate of a region
- Often we have separate data matrices for individual level data
- Often we model lakes or streams in different regions separately

Data Structure

- Individual level data, e.g., individual lake data
- Group level data, e.g., climate of a region
- Often we have separate data matrices for individual level data
- Often we model lakes or streams in different regions separately
- Group level information is never a big part

Why Multilevel Models?

- Modeling group level regression coefficients

Why Multilevel Models?

- Modeling group level regression coefficients
- Modeling variation among individual level regression coefficients

Why Multilevel Models?

- Modeling group level regression coefficients
- Modeling variation among individual level regression coefficients
- Estimating regression coefficients for a particular group with limited sample size

Effects of Urbanization on Stream Ecosystems

- USGS' National Water Quality Assessment Program – EUSE Topic

Effects of Urbanization on Stream Ecosystems

- USGS' National Water Quality Assessment Program – EUSE Topic
- 9 sites selected to study EUSE

Effects of Urbanization on Stream Ecosystems

- USGS' National Water Quality Assessment Program – EUSE Topic
- 9 sites selected to study EUSE
- Urbanization is measured by the National Urbanization Intensity Index

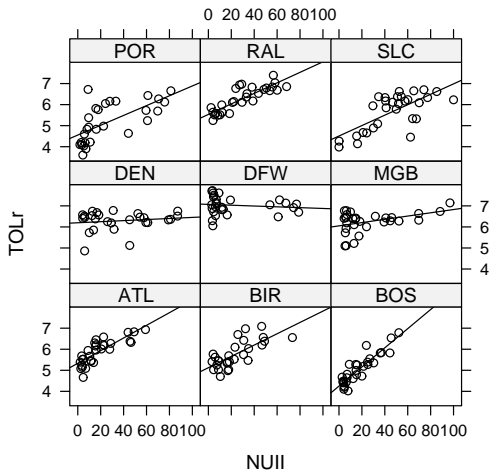
Effects of Urbanization on Stream Ecosystems

- USGS' National Water Quality Assessment Program – EUSE Topic
- 9 sites selected to study EUSE
- Urbanization is measured by the National Urbanization Intensity Index
- Ecosystem response is measured by species composition of benthic algae and macroinvertebrates, and composition of fish species.

Effects of Urbanization on Stream Ecosystems

- USGS' National Water Quality Assessment Program – EUSE Topic
- 9 sites selected to study EUSE
- Urbanization is measured by the National Urbanization Intensity Index
- Ecosystem response is measured by species composition of benthic algae and macroinvertebrates, and composition of fish species.
- Community diversity is measured by various aggregated variables

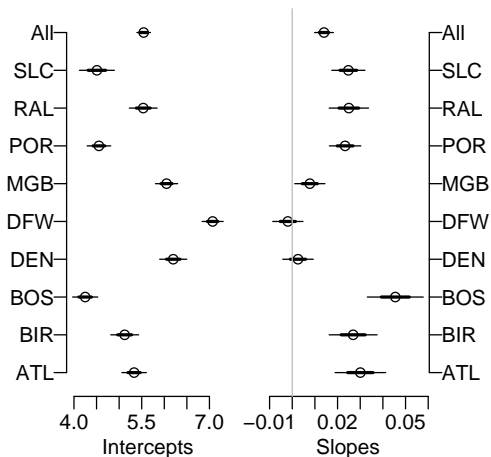
Data



Fitting a Simple Linear Regression Model

Models fitted to data from each region separately:

$Y = \beta_0 + \beta_1 X + \varepsilon$, using `Region` as a dummy variable:



Are the 9 Regions Entirely Different?

- When fitting a regression model using `Region` as a factor predictor, we fit 9 separate regression models (no pooling)

Are the 9 Regions Entirely Different?

- When fitting a regression model using `Region` as a factor predictor, we fit 9 separate regression models (no pooling)
- We can combine data from all 9 sites to fit a single model (complete pooling)

Are the 9 Regions Entirely Different?

- When fitting a regression model using `Region` as a factor predictor, we fit 9 separate regression models (no pooling)
- We can combine data from all 9 sites to fit a single model (complete pooling)
- Neither approaches are effective:

Are the 9 Regions Entirely Different?

- When fitting a regression model using `Region` as a factor predictor, we fit 9 separate regression models (no pooling)
- We can combine data from all 9 sites to fit a single model (complete pooling)
- Neither approaches are effective:
 - We are studying the same effects in 9 regions, they have common features

Are the 9 Regions Entirely Different?

- When fitting a regression model using `Region` as a factor predictor, we fit 9 separate regression models (no pooling)
- We can combine data from all 9 sites to fit a single model (complete pooling)
- Neither approaches are effective:
 - We are studying the same effects in 9 regions, they have common features
 - The regions are different in many ways

Are the 9 Regions Entirely Different?

- When fitting a regression model using `Region` as a factor predictor, we fit 9 separate regression models (no pooling)
- We can combine data from all 9 sites to fit a single model (complete pooling)
- Neither approaches are effective:
 - We are studying the same effects in 9 regions, they have common features
 - The regions are different in many ways
- Can we partially pool the data?

Multilevel Model—the General Idea

- The 9 regions have common features: 30 sub-watersheds within each of the 9 regions were selected to represent a gradient of urbanization, ecosystem responses are measured using the same metric

Multilevel Model—the General Idea

- The 9 regions have common features: 30 sub-watersheds within each of the 9 regions were selected to represent a gradient of urbanization, ecosystem responses are measured using the same metric
- The 9 regions are different: different climate, different eco-region, different history of urbanization

Multilevel Model—the General Idea

- The 9 regions have common features: 30 sub-watersheds within each of the 9 regions were selected to represent a gradient of urbanization, ecosystem responses are measured using the same metric
- The 9 regions are different: different climate, different eco-region, different history of urbanization
- How to model the common feature and distinguish region-specific characteristics?

Multilevel Model—the General Idea

- The 9 regions have common features: 30 sub-watersheds within each of the 9 regions were selected to represent a gradient of urbanization, ecosystem responses are measured using the same metric
- The 9 regions are different: different climate, different eco-region, different history of urbanization
- How to model the common feature and distinguish region-specific characteristics?
- A hyper distribution on model parameters

Multilevel Model – A Simplified Example

- Suppose the 9 regions have similar urbanization index, we want to estimate the tolerance of macroinvertebrates for each of the 9 regions

Multilevel Model – A Simplified Example

- Suppose the 9 regions have similar urbanization index, we want to estimate the tolerance of macroinvertebrates for each of the 9 regions
- Suppose the measure we used can be model by a normal distribution

Multilevel Model – A Simplified Example

- Suppose the 9 regions have similar urbanization index, we want to estimate the tolerance of macroinvertebrates for each of the 9 regions
- Suppose the measure we used can be model by a normal distribution
- No pooling – we estimate the mean and variance of the diversity measure separately (9 means and 9 variances)

Multilevel Model – A Simplified Example

- Suppose the 9 regions have similar urbanization index, we want to estimate the tolerance of macroinvertebrates for each of the 9 regions
- Suppose the measure we used can be model by a normal distribution
- No pooling – we estimate the mean and variance of the diversity measure separately (9 means and 9 variances)
- Complete pooling – we estimate the combined mean and variance

Multilevel Model – A Simplified Example

- Suppose the 9 regions have similar urbanization index, we want to estimate the tolerance of macroinvertebrates for each of the 9 regions
- Suppose the measure we used can be model by a normal distribution
- No pooling – we estimate the mean and variance of the diversity measure separately (9 means and 9 variances)
- Complete pooling – we estimate the combined mean and variance
- Partial pooling – we consider site means are related and model them as random variables from the same (hyper) distribution

Multilevel Model – A Simplified Example

- No pooling: $y_i \sim N(\mu_{j[i]}^N, \sigma^N)$

Multilevel Model – A Simplified Example

- No pooling: $y_i \sim N(\mu_{j[i]}^N, \sigma^N)$
- Complete pooling: $y_i \sim N(\mu^c, \sigma^c)$

Multilevel Model – A Simplified Example

- No pooling: $y_i \sim N(\mu_{j[i]}^N, \sigma^N)$
- Complete pooling: $y_i \sim N(\mu^c, \sigma^c)$
- Partial pooling:

$$y_i \sim N(\mu_{j[i]}^p, \sigma)$$
$$\mu_j^p \sim N(\mu, \tau)$$

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{\cdot j}$

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$
- Partial pooling:

$$\hat{\mu}_j^p = \frac{\frac{n_j}{\sigma^2} \bar{y}_{.j} + \frac{1}{\tau^2} \bar{y}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$
- Partial pooling:

$$\hat{\mu}_j^p = \frac{\frac{n_j}{\sigma^2} \bar{y}_{.j} + \frac{1}{\tau^2} \bar{y}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

- a weighted average of city specific sample mean and overall sample mean ($\hat{\mu}_j^p$ is closer to \bar{y} than $\bar{y}_{.j}$ is)

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$
- Partial pooling:

$$\hat{\mu}_j^p = \frac{\frac{n_j}{\sigma^2} \bar{y}_{.j} + \frac{1}{\tau^2} \bar{y}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

- a weighted average of city specific sample mean and overall sample mean ($\hat{\mu}_j^p$ is closer to \bar{y} than $\bar{y}_{.j}$ is)
 - If n_j is large, $\hat{\mu}_j$ will be close to $\bar{y}_{.j}$

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$
- Partial pooling:

$$\hat{\mu}_j^p = \frac{\frac{n_j}{\sigma^2} \bar{y}_{.j} + \frac{1}{\tau^2} \bar{y}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

- a weighted average of city specific sample mean and overall sample mean ($\hat{\mu}_j^p$ is closer to \bar{y} than $\bar{y}_{.j}$ is)
 - If n_j is large, $\hat{\mu}_j$ will be close to $\bar{y}_{.j}$
 - If n_j is small, $\hat{\mu}_j$ will be close to \bar{y}

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$
- Partial pooling:

$$\hat{\mu}_j^p = \frac{\frac{n_j}{\sigma^2} \bar{y}_{.j} + \frac{1}{\tau^2} \bar{y}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

- a weighted average of city specific sample mean and overall sample mean ($\hat{\mu}_j^p$ is closer to \bar{y} than $\bar{y}_{.j}$ is)
 - If n_j is large, $\hat{\mu}_j$ will be close to $\bar{y}_{.j}$
 - If n_j is small, $\hat{\mu}_j$ will be close to \bar{y}
 - If $n_j = 0$, $\hat{\mu}_i = \bar{y}$

Multilevel Model – A Simplified Example

- No pooling: $\hat{\mu}_j^N = \bar{y}_{.j}$
- Complete pooling: $\hat{\mu}^c = \bar{y}$
- Partial pooling:

$$\hat{\mu}_j^p = \frac{\frac{n_j}{\sigma^2} \bar{y}_{.j} + \frac{1}{\tau^2} \bar{y}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

- a weighted average of city specific sample mean and overall sample mean ($\hat{\mu}_j^p$ is closer to \bar{y} than $\bar{y}_{.j}$ is)
 - If n_j is large, $\hat{\mu}_j$ will be close to $\bar{y}_{.j}$
 - If n_j is small, $\hat{\mu}_j$ will be close to \bar{y}
 - If $n_j = 0$, $\hat{\mu}_i = \bar{y}$
 - If τ^2 is high (or σ^2 is small), $\hat{\mu}_j$ will be close to $\bar{y}_{.j}$, and vice versa.

Multilevel Model – Shrinkage

- The pulling of $\hat{\mu}_j^P$ towards \bar{y} is known as shrinkage

Multilevel Model – Shrinkage

- The pulling of $\hat{\mu}_j^P$ towards \bar{y} is known as shrinkage
- Shrinking towards the overall mean is a form of information discounting

Multilevel Model – Shrinkage

- The pulling of $\hat{\mu}_j^P$ towards \bar{y} is known as shrinkage
- Shrinking towards the overall mean is a form of information discounting
- Many studies shown that partial pooling performs better than no or complete pooling

Multilevel Model – Linear Regression

- $y_i = \beta_{0j[i]} + \beta_{1j[i]}x_i + \varepsilon_i$

Multilevel Model – Linear Regression

- $y_i = \beta_{0j[i]} + \beta_{1j[i]}x_i + \varepsilon_i$
- $\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{\beta_0} \\ \mu_{\beta_1} \end{pmatrix}, \Sigma_{\beta} \right)$

Multilevel Model – Linear Regression

- $y_i = \beta_{0j[i]} + \beta_{1j[i]}x_i + \varepsilon_i$
- $\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{\beta_0} \\ \mu_{\beta_1} \end{pmatrix}, \Sigma_{\beta} \right)$
- The partial pooling estimated regression coefficients are weighted average of the estimates from no pooling and complete pooling

Group Level Predictors

- What can explain the changes in the slope?

Group Level Predictors

- What can explain the changes in the slope?
 - Slope is the effect of urbanization on average tolerance

Group Level Predictors

- What can explain the changes in the slope?
 - Slope is the effect of urbanization on average tolerance
 - Can the difference reflect the site climate/environmental conditions?

Group Level Predictors

- What can explain the changes in the slope?
 - Slope is the effect of urbanization on average tolerance
 - Can the difference reflect the site climate/environmental conditions?
- What can explain the changes in the intercept?

Group Level Predictors

- What can explain the changes in the slope?
 - Slope is the effect of urbanization on average tolerance
 - Can the difference reflect the site climate/environmental conditions?
- What can explain the changes in the intercept?
 - Intercept is the tolerance at 0 urbanization

Group Level Predictors

- What can explain the changes in the slope?
 - Slope is the effect of urbanization on average tolerance
 - Can the difference reflect the site climate/environmental conditions?
- What can explain the changes in the intercept?
 - Intercept is the tolerance at 0 urbanization
 - It is reasonable to see varying intercept by site

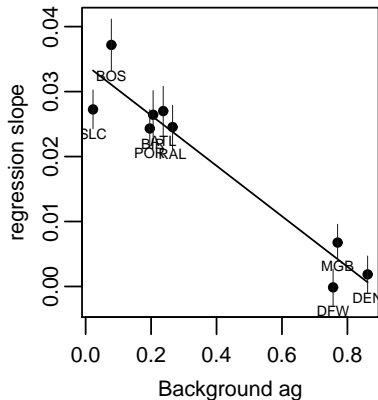
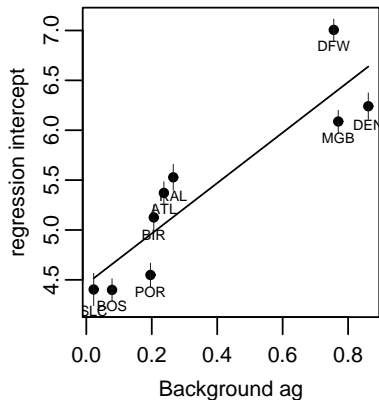
Group Level Predictors

- What can explain the changes in the slope?
 - Slope is the effect of urbanization on average tolerance
 - Can the difference reflect the site climate/environmental conditions?
- What can explain the changes in the intercept?
 - Intercept is the tolerance at 0 urbanization
 - It is reasonable to see varying intercept by site
- A careful detective work is needed.

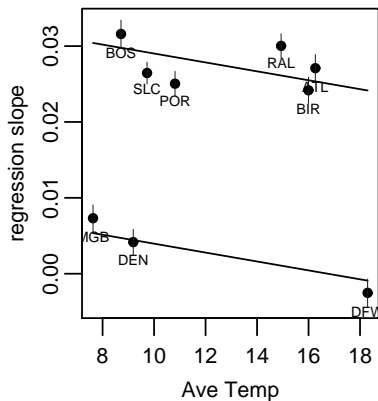
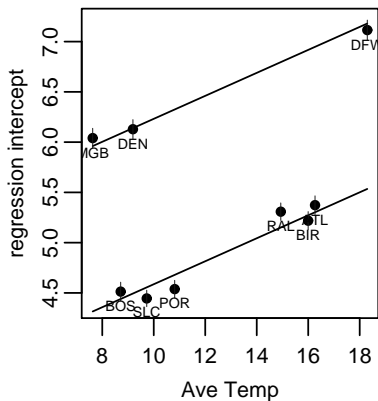
What Are the Usual Suspects?

- Temperature and its variation
- Moisture and its variation
- Elevation and relief
- Anything else?

The Agriculture effect



Temperature effect



Why multilevel is better?

- A framework for pooling data from multiple sources

Why multilevel is better?

- A framework for pooling data from multiple sources
- A framework for combining information

Why multilevel is better?

- A framework for pooling data from multiple sources
- A framework for combining information
- A mechanism for learning – combining data across multiple spatiotemporal scales to widen the inference base of the model

EUSE Applications

- EUSE – biological community response to urbanization – which response is affected by which predictor?

EUSE Applications

- EUSE – biological community response to urbanization – which response is affected by which predictor?
 - Which factor in `nuii` is the main driving force?

EUSE Applications

- EUSE – biological community response to urbanization – which response is affected by which predictor?
 - Which factor in `nuii` is the main driving force?
 - Which indicator is responding to what aspect of urbanization?

EUSE Applications

- EUSE – biological community response to urbanization – which response is affected by which predictor?
 - Which factor in `nuii` is the main driving force?
 - Which indicator is responding to what aspect of urbanization?
 - How the responses vary and why?

Acknowledgements

- Co-authors – Thomas F. Cuffney, Ibrahim Alameddine, Gerald McMahon and Kenneth H. Reckhow
- Funding – USGS-Duke cooperative agreement