Hierarchical Statistical Modeling

Peter F. Craigmile

http://www.stat.osu.edu/~pfc/

Statistics for Environmental Evaluation
Quantifying Environmental Risk and Resilience Workshop
Glasgow, Scotland
5-8 January 2016

Acknowledgments

• Thank you to NERC and The School of Mathematics and Statistics at the University of Glasgow.
• Some of my research is funded by the US National Science Foundation, NSF SES-1424481 and NSF DMS-1407604.
• This lecture discusses an application that was part of a research project between The Ohio State University and Battelle, supported by the American Chemistry Council’s (ACC) Long-Range Research Initiative (Collaborators: Catherine Calder, Noel Cressie, Hongfei Li, Rajib Paul, Thomas Santner, and Jian Zhang).

Hierarchical modeling

• Hierarchical modeling has been increasingly recognized as a powerful approach for analyzing complex phenomena.
  – Can synthesize data from different sources.
  – Can accommodate complicated dependence structures.
  – Handle irregular features of data such as missingness and censoring.
  – Incorporate scientifically based process information.

However, despite the inherent elegance of the approach, effective derivation of inferences from these models is by no means a trivial task.

What is a hierarchical statistical model?

• Hierarchical statistical models are a way to build more complicated models by connecting together a number of simpler ones.
  – The simpler models are defined in terms of conditional distributions.
  – Hierarchical models can help us to think about how the data actually arises.
Motivating example: Calibrating breathalyzers

[Adapted from Dean and Voss, 1999, Chapter 17, p.639, Ex. 1]

Solutions of alcohol are used for calibrating Breathalyzers. The data below show the alcohol concentrations in (mg/ml) of samples of alcohol solutions taken from six bottles of alcohol solution randomly selected from a large batch. Concentrations are determined by gas chromatography.

<table>
<thead>
<tr>
<th>Bottle</th>
<th>Concentration (mg/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.4357 1.4348 1.4336 1.4309</td>
</tr>
<tr>
<td>2</td>
<td>1.4244 1.4232 1.4213 1.4256</td>
</tr>
<tr>
<td>3</td>
<td>1.4153 1.4137 1.4176 1.4164</td>
</tr>
<tr>
<td>4</td>
<td>1.4331 1.4325 1.4312 1.4297</td>
</tr>
<tr>
<td>5</td>
<td>1.4252 1.4261 1.4293 1.4272</td>
</tr>
<tr>
<td>6</td>
<td>1.4179 1.4217 1.4191 1.4204</td>
</tr>
</tbody>
</table>

The one-way random effects model

- Consider the situation where we draw treatment effects from a random population; let $T$ denote the random factor.
  - We obtain a random samples from $T$. Let $T_i (i = 1, \ldots, a)$ denote these random variables.

- The one-way random effects model is given by
  \[ Y_{ij} = \mu + T_i + \epsilon_{ij}, \]
  for $i = 1, \ldots, a$, and $j = 1, \ldots, n_i$, where
  - $\{T_i\}$ are a set of independent $N(0, \sigma_T^2)$ RVs,
  - $\{\epsilon_{ij}\}$ are a set of independent $N(0, \sigma^2)$ RVs, and
  - $\epsilon_{ij}$ and $T_i$ are independent of one another.

Depicting the model graphically
Rewritten as a hierarchical model...

- Here is how we write our random effects model as a hierarchical statistical model:

1. Let \( T_1, \ldots, T_a \) be an independent set of \( \mathcal{N}(0, \sigma_T^2) \) RVs.
2. Conditional on each \( T_i \) let \( \{Y_{ij} : t = 1, \ldots, n_i\} \) be a set of independent RVs with
   \[ Y_{ij}|T_i \sim \mathcal{N}(\mu + T_i, \sigma^2), \]
   for each \( t = 1, \ldots, n_i \).

Are the random effects parameters?

Bayesian hierarchical modeling

- Suppose more generally that the vector of parameters in our statistical model is \( \theta \).
- Given data \( y \), our likelihood for the data is given by \( f(y|\theta) \).
- In the Bayesian paradigm we specify a prior distribution for our parameters, \( \pi(\theta) \).
- Then using Bayes theorem we have that the posterior distribution of the parameters given the data is
  \[ \pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}, \]
  where \( f(y) \) is the marginal density for the data.
- All Bayesian inference is based on the posterior distribution.

Arsenic study - the game plan

- We provide a behind-the-scenes look at a Bayesian hierarchical analysis.
- We show how we carried out our model building, fitting, and checking (which were complicated by issues related to the quality and quantity of available data).
- Strategies will be presented for compartmentalizing model components and inputs in order to allow efficient assessment of the impact of modeling assumptions on inferences.
Bayesian hierarchical analysis of human exposure pathways

- A pathways analysis refers to the study of the relationship between levels of toxicants in environmental media (e.g., air, water, dust, food) and levels of personal exposure.
- We use the Arizona (AZ) data from Phase I of the National Human Exposure Assessment Survey (NHEXAS), carried out in the 1990s by USEPA, CDC, and FDA.
(There were two other studies: Region 5 and Maryland)

Arsenic risk

- We focus on modeling pathways of exposure to arsenic (As).
  - Acute exposure to As:
    Associated with irritation of the gastrointestinal and respiratory tracts
  - Chronic exposure: Related to melanosis, hyperpigmentation, depigmentation, hyperkeratosis, and skin cancer.
    May also affect the nervous, cardiovascular, and haematopoietic system.
    [WHO, 1981]

Review

- Clayton et al. [2002] used a structural-equations-modeling approach to explore pathways of exposure to As and lead using NHEXAS Region 5 data.
- Also using the As NHEXAS Region 5 data, McMillan et al. [2006] and Cressie et al. [2007] developed Bayesian hierarchical pathways models.
  - Able to accommodate measurement error, as well as missing and censored (below a specified minimum detection limit (MDL)) observations, in a coherent manner.
The pathways model

17

Our approach

• One of the primary aims of NHEXAS was to identify geographical variation in exposure to hazardous chemicals [Pellizzari et al., 1995].

• Cressie et al. [2007] and Calder et al. [2008] linked Soil As concentrations in NHEXAS Region 5 to background concentrations of As in both topsoil and stream sediments.

• In our analysis: we link the three outdoor media (Soil, Water, and Outdoor air) As concentrations to background concentrations of As.

Phase I NHEXAS AZ data

• Collected using a three-stage, population-based sampling design.

• We study a subset of 179 individuals for whom biomarker and environmental variables were collected over a seven-day period.

• There are data quality issues.

<table>
<thead>
<tr>
<th>Media</th>
<th>#Missing</th>
<th>#&lt;MDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine</td>
<td>12</td>
<td>55</td>
</tr>
<tr>
<td>Personal Air</td>
<td>168</td>
<td>4</td>
</tr>
<tr>
<td>Indoor Air</td>
<td>51</td>
<td>90</td>
</tr>
<tr>
<td>Outdoor Air</td>
<td>58</td>
<td>82</td>
</tr>
<tr>
<td>Sill Dust</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>Soil</td>
<td>110</td>
<td>0</td>
</tr>
<tr>
<td>Food</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Beverage</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Water</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

NHEXAS exploratory data analysis

response
Global water data

(From Water Quality Division of the AZ Dept. of Environmental Quality).

- 10,688 As concentration measurements (units of mg/l) from treated water samples collected at 1,161 public water systems (PWSs)
- Due to security concerns, we were not provided the PWS locations.
  - Instead, we were given number of individuals in each of the 15 AZ counties served by each PWS.
  - Each PWS may serve more than one county and each county can be served by more than one PWS.

Global soil data

- Two sources of supplemental information about the levels of As in soils across AZ from U.S. Geological Survey:
  1. (USSoils) Point-referenced measurements of As in the surface layer from samples of soil, sand, silt, and alluvial deposits collected in undisturbed regions.
  2. (NGS) Stream sediment measurements of As, associated with watersheds, or hydrologically similar regions [Seaber et al., 1987]. Our analysis includes all stream sediment data that were taken in the 84 watersheds that contain at least part of the state of AZ.
- The units of both are µg/m³.
Global soil EDA

- **Interagency Monitoring of Protected Visual Environments Network** was established in 1985.
- Designed to provide long-term air-quality records for U.S. national parks and wilderness areas.
- Readings (units of $\mu g/m^3$) are collected on an every 3-4 day schedule, and certain observations are flagged as being below a MDL.

Global air data

- Interagency Monitoring of Protected Visual Environments Network was established in 1985.
- Designed to provide long-term air-quality records for U.S. national parks and wilderness areas.
- Readings (units of $\mu g/m^3$) are collected on an every 3-4 day schedule, and certain observations are flagged as being below a MDL.

Site-by-site time series

The local-environment-to-biomarker (LEB) model

- Consists of three parts:
  1. **Data model**: Each observed measurement for medium $j$ and individual $i$ is observed conditionally independent given the true process $X_{ij}^M$ and a known measurement precision. Assume same model for missing values, but censoring model for values below the MDL.
  2. **Process model**: see next slide
  3. **Prior model**: we define distributions for the other parameters in the model.
LEB process model

• We link elements of $X^M$ by choosing subsets of the media, $\{S^M_j\}$, that are defined by a directed acyclic graph (DAG) [e.g., Lauritzen, 1996].

• These subsets define the joint distribution of $X^M$ as follows:

$$[X^M | \beta^M, \mu^M, \tau] = d \prod_{i=1}^{N^M} \prod_{j=1}^{N^M} N \left( \mu^M_j + \sum_{k \in S^M_j} \beta^M_m X^M_{ik} ; 1/\tau^M_j \right).$$

• We assume $\mu^M_j$ is constant across the state.

• In our case, the DAG was defined by considering partitions $M_k$ of the set of $N^M$ media and defining selector sets $\{S^M_j\}$ that satisfy

$$S^M_j = \emptyset, \text{ if } j \in M_1; \quad S^M_j \subseteq \bigcup_{k=1}^{\ell-1} M_k, \text{ if } j \in M_\ell, \text{ when } \ell > 1.$$

The global-to-local-environment (GLE) models

• We consider alternative versions of the model for $\{X^M_{ij}\}$ for medium $j$ that draw on supplemental data sources and knowledge about the background spatial processes driving the variation in As levels in that medium.

Global water model

• Our data model respects that:

  – Not all of the PWSs in AZ have an observation available, while some have multiple observations.

  – Certain observations were below an MDL and others are missing.

• The global-water process defined at each PWS is conditionally independent given PWS-specific mean parameters and an unknown precision parameter.

  – Each PWS mean is drawn from a normal prior, with a common mean.

• We link the global-water process to the process on the Water medium in the LEB model, via a mixture model that captures the uncertainty in which individuals are served by each PWS.

Global soil model

• The data and process models for the log As concentrations in topsoil and stream sediments are taken from Calder et al. [2008].

• Then we link these media to the NHEXAS soil-related media processes in the LEB model using a mixture model [not in Calder et al., 2008].

  – This mixture model captures the uncertainty in locations of the NHEXAS individuals across AZ and the uncertainty due to our not having precise geographic information for the NHEXAS individuals (NHEXAS only provides county of residence).

• Our approach accommodates the spatially misaligned sampling schemes of NHEXAS, USSoils, and NGS.
Global air model
- We are hesitant to fit any more complicated model.
- Letting $m(A)$ be the index corresponding to the Outdoor Air medium in the LEB model we assume
  \[ \mu_{m(A)} \sim N(-6, 4), \]
on the log ppm scale.
- Has a larger standard deviation than is suggested by the data, to capture the change-of-support of moving from the IMPROVE monitors to the monitors used outside the home in the NHEXAS study.

Implementing the Markov chain Monte Carlo algorithm
- The posterior distributions for parameters are not available in closed form.
- Instead, use a Markov chain Monte Carlo algorithm (MCMC) sampling algorithm that sequentially samples from the full conditional posterior distributions of model parameters.
  - Since some of these full conditional distributions not available in closed form, we employ independence or symmetric-random-walk Metropolis steps.

Model fitting and checking
- Data management
- Implementing the Markov chain Monte Carlo algorithm
- Stepwise fitting of the model
  - The LEB model
  - The LEB model + global water + global soil + global air
  - Residual analysis
  - Further sensitivity analyses
  - Additional model comparisons

Comparing the slope parameters
Summary of scientific results

- Clayton et al. [2002] pathways model in NHEXAS Region 5 is applicable to the analysis of As exposure pathways in NHEXAS AZ.
- Except for the relationship of Urine with Soil, any declared relationship between media was positive.
- The incorporation of GLE models did little to affect these relationships, but did move the negative relationship of Urine with Soil to be closer to zero, and move the posterior distributions of the parameters of processes for subjects with missing or below MDL values.
- GLE models help us in assessing spatial variation in exposure across AZ.
  (global data have better spatial coverage than the NHEXAS AZ data).

Extra lessons learned

- Within any application, great care and organization is required.
- If model fitting not implemented carefully, inferences will be suspect.
- Derived datasets must be reproducible.
- An MCMC algorithm should be coded and tested in “bite-size pieces”.
  – Not sufficient to fit a large model all at once.
- Instead, components of models should be combined in a systematic fashion.
- If an analysis produces an interesting result, it is necessary to understand which components of the model and of the data are driving the result.

References