Lecture 15: Gelman Hill Ch 11

Multilevel Models

For multilevel models, observations fall into groups and coefficients can vary by the group.

Assume there are J groups and $j[i]$ denotes that observation i falls into group j

$$y_i = \alpha + \beta x_i + \epsilon_i$$

complete pooling

$$\begin{aligned}
y_{[1]i} &= \alpha_1 + \beta_1 x_{[1]i} + \epsilon_i \\
y_{[2]i} &= \alpha_2 + \beta_2 x_{[2]i} + \epsilon_i \\
&. \\
&. \\
&. \\
y_{[J]i} &= \alpha_J + \beta_J x_{[J]i} + \epsilon_i
\end{aligned}$$

no pooling

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$$

partial pooling

Shrinkage Equation

Assume that the multilevel model only includes group averages, then the partial pooling estimate of the mean (or intercept) is

$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}$$

where σ_y^2 is the variance of the data and σ_α^2 is the variance of the group-level averages.

Thus the estimate value for a group is a weighted average from the data in that group and the overall data.

The weights are:

1. a function of the data variance and the number of observations in a group, and
2. function of the variance of the group level estimates.

For each scenario, large variance corresponds to a lower weight on that component and smaller variance (high precision) corresponds to higher weights

Correlation structure

A common assumption in regression models is that the observations are independent. There are a few common data types that violate this assumption and can be addressed with hierarchical models.

Repeated Measurements: repeated measurements on persons (or units), thus the data observations are clustered.

Cross Sectional Data (Longitudinal/Time series): Repeated measurements across time.

"Fixed vs. Random"

These type of models are commonly referred to as "mixed models" that include "fixed" and "random" effects.

Random Effects: the coefficients that vary (across groups) are often referred to as random effects. We will see a formal statistical distribution associated with these later on.

Fixed Effects: GH point out inconsistencies with this term. Fixed effects generally refer to coefficients that do not vary (say a parameter estimated across all groups). This could also apply to the separate models approach. The defining feature is largely a probability distribution for model.

Some general advice about when to use fixed/random effects focuses on the research goal; however, GH suggest *always* using multilevel models.

Furthermore, given the inconsistencies in the meaning of fixed/random, GH (and I) prefer using multilevel or hierarchical models.

Multilevel Modeling Pros and Cons

Classical Regression Overview

- prediction for continuous or discrete outcomes
- fitting of nonlinear relationships (using transformations and basis functions)
- inclusion of categorical predictors using indicator functions
- interactions between inputs
- GLM frameworks for non-Gaussian (normal) probability distributions

Multilevel Modeling

- Accounting for and estimating individual- and group-level variation by estimating group-level coefficients (and potentially including group-level covariates.
- Modeling variation among individual-level regression coefficients and making predictions for new individuals/groups.
- Note there is extra complexity in fitting a multilevel model and additional modeling assumptions.
- Limiting cases of multilevel models
  - very little group variation, then the multilevel model approaches the complete pooling scenario
  - very large group variation, then the multilevel model approaches the seperate model solution