STA 610L: Module 1.3

Introduction to hierarchical models

Dr. Olanrewaju Michael Akande

Introduction to hierarchical models

The terminology hierarchical model is quite general and covers a very wide range of models.

A hierarchical model can refer to the simple use of a prior distribution for a relatively simple model, or to a highly organized data hierarchy (students nested in classes, nested in schools, nested in states, nested in countries).

Hierarchical models are thus usually richer and more flexible than the usual standard models.

For grouped data for example, we may want to estimate the relationship between a response variable and certain predictors collected across all the groups, where observations in the same group are more alike than those in different groups.

In that case, we may want to do inference in a way that takes advantage of the relationship between observations in the same group, but we may also look to borrow information across groups.

Hierarchical modeling lets us do exactly that.

Introduction to hierarchical models

Hierarchical models are often used in many commonly-encountered settings such as when

- members of a "cluster" share more similarities with each other than with members of other clusters, violating the typical independence assumption of generalized linear models
    - examples of clusters include members of a family, or students in a class;
- hypotheses of interest include context-dependent associations, often across a large number of settings
    - e.g., does success of a new mode of instruction depend on the individual teacher;
- it is necessary to borrow information across groups in order to stabilize estimates or to obtain estimates with desirable properties
    - e.g., we want to make state-specific estimates of election candidate preference by country of origin, but some states may have few immigrants from a given country.

Hypothetical school testing example

Suppose we wish to estimate the distribution of test scores for students at J different high schools.

In each school j, where $j = 1, \ldots, J$, suppose we test a random sample of n_j students.

Let y_{ij} be the test score for the ith student in school j, with $i = 1, \ldots, n_j$.

Hypothetical school testing example

Option I: estimation can be done separately in each group, where we assume

$$y_{ij}|\mu_j, \sigma_j^2 \sim N\left(\mu_j, \sigma_j^2\right)$$

where for each school j, μ_j is the school-wide average test score, and σ_j^2 is the school-wide variance of individual test scores.

We can do classical inference for each school based on large sample 95% CI: $\bar{y}_j \pm 1.96\sqrt{s_j^2/n_j}$, where $\bar{y}_j$ is the sample average in school j, and s_j^2 is the sample variance in school j.

Clearly, we can overfit the data within schools, for example, what if we only have 4 students from one of the schools?

Hypothetical school testing example

Option II: alternatively, we might believe that $\mu_j = \mu$ for all j; that is, all schools have the same mean. This is the assumption (null hypothesis) in ANOVA models for example.

Option I ignores that the μ_j's should be reasonably similar, whereas option II ignores any differences between them.

It would be nice to find a compromise!

This is what we are able to do with hierarchical modeling.

Hierarchical model

Once again, suppose

$$y_{ij}|\mu_j, \sigma_j^2 \sim N\left(\mu_j, \sigma_j^2\right); \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, J.$$

We can assume that the μ_j's are drawn from a distribution based on the following: conceive of the schools themselves as being a random sample from all possible school.

Suppose μ_0 is the overall mean of all school's average scores (a mean of the means), and τ^2 is the variance of all school's average scores (a variance of the means).

Hierarchical model

Then, we can think of each μ_j as being drawn from a distribution, e.g.,

$$\mu_j|\mu_0, \tau^2 \sim N\left(\mu_0, \tau^2\right),$$

which gives us one more level, resulting in a hierarchical specification.

Usually, μ_0 and τ^2 will also be unknown so that we need to estimate them (usually MLE or Bayesian methods).

We will revisit estimation soon.

Hierarchical model: school testing example

Back to our example, it turns out that the multilevel estimate is

$$\hat{\mu}_j \approx \frac{\frac{n_j}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu_0}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}},$$

but since the unknown parameters have to be estimated, the classical estimate is

$$\hat{\mu}_j \approx \frac{\frac{n_j}{s_j^2}\bar{y}_j + \frac{1}{\hat{\tau}^2}\bar{y}_{\text{all}}}{\frac{n_j}{s_j^2} + \frac{1}{\hat{\tau}^2}},$$

where $\bar{y}_{\text{all}}$ is the completely pooled estimate (the overall sample mean of all test scores).

Hierarchical model: implications

Our estimate for each μ_j is a weighted average of $\bar{y}_j$ and μ_0, ensuring that we are borrowing information across all levels through μ_0 and τ^2.

The weights for the weighted average is determined by relative precisions (the inverse of variance is often referred to as precision) from the data and from the second level model.

Suppose all $\sigma_j^2 \approx \sigma^2$. Then the schools with smaller n_j have estimated μ_j closer to μ_0 than schools with larger n_j.

Thus, the hierarchical model shrinks estimates with high variance towards the grand mean.

We seek to specify models like this in many different contexts, for many reasons, including the idea of "shrinkage".

We will do this over and over throughout the course.

Hierarchical model: demo

For some intuition behind hierarchical models, we'll check out this neat tutorial by Michael Freeman at University of Washington.

What's next?

Move on to the readings for the next module!