IDS 702: Module 4.1

Introduction to multilevel/hierarchical models

Dr. Olanrewaju Michael Akande

IDS 702

Multilevel, clustered or grouped data

- Often data are grouped or clustered naturally, for example
  - students within schools,
  - patients within hospitals,
  - voters within counties or states, or
  - repeated measurements on same person, as is often the case in longitudinal studies.
- For such clustered data, we may want to infer or estimate the relationship between a response variable and certain predictors collected across all the groups.
- Ideally, we should do so in a way that takes advantage of the relationship between observations in the same group, but we should also look to borrow information across groups.
- Hierarchical or multilevel models provide a principled way to do so. We will start with simpler cases to elucidate the main ideas.

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 $i = 1, \ldots, n_j$.
- 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.

IDS 702

Hypothetical school testing example

- 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?
- 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).

- 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 (think maximum likelihood or Bayesian methods).

IDS 702

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, we actually have

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

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

IDS 702

Hierarchical model: school testing example

- We will only scratch the surface of hierarchical modeling. Take a look at the readings for hierarchical linear models on the website for more resources.
- If you want to take a course that explores hierarchical models in much more detail, consider taking STA 610 (after taking STA 602).
- For those interested in Bayesian inference (feel free to skip this if you are not!), it turns out that the posterior distribution of μ_j, $p(\mu_j|Y, \sigma_j^2, \mu_0, \tau^2) = N(\mu_j^\star, \nu_j^\star)$, where

$$\mu_j^\star = \frac{\frac{n_j}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu_0}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$\nu_j^\star = \frac{1}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}$$

IDS 702

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.

What's next?

Move on to the readings for the next module!

IDS 702