STAT 535: Chapter 15: Hierarchical Models and Pooling

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

Hierarchical (Grouped) Data

- Hierarchical models are used for data with some type of **grouping** structure.
- Examples from textbook:
  - a sampled group of schools and data Y on multiple individual students within each school (this is called **clustered data**)
  - a sampled group of labs and data Y from multiple individual experiments within each lab (this is also **clustered data**)
  - a sampled group of people on whom we make multiple individual observations of some variable Y over time (this is called **longitudinal data**)
- Such data are not independent: Observations within each school, or within each lab, or for each person, are likely to be similar to each to other, so they will be correlated.
- If we ignore the fact that the data are grouped, then our estimates and inferences can be misleading.

Terminology

- Sometimes hierarchical models are called multilevel models, mixed effects models, or random effects models (these last terms depend on the complexity of the situation).
- Panel data are longitudinal data in which the set of subjects is the same at each time point (in general, with longitudinal data the set of subjects at each time point may or may not be the same).

Example of Grouped Data

- Consider the `cherry_blossom_sample` data set in the textbook. It has race times (in minutes) for a set of runners (aged in their 50s or 60s) over a series of years.
- Note that some/most runners appear multiple times in the data set (since they raced in multiple years), so the data are grouped within runner.
- Race times coming from the same runner will NOT be independent; they will be correlated.

Visualizing the Cherry Blossom Race Data

- Side-by-side boxplots show the distribution of race times for 36 runners who ran in multiple Cherry Blossom races.
- Runner 10 was slow; runner 29 was fast and consistent; runner 17's performance varied a lot.
- Question of interest: What is the relationship between a runner's age and race time?

Complete Pooling

- Suppose we **pool** all the data together (even though some data values are coming from the same runner) and look at a scatterplot of race time against age (see R).
- Looks like a fairly weak relationship based on this plot.
- Let's do a simple linear regression of $Y =$ net race time against $X =$ age (see R example):
- Conclusions: Age does not seem to be a significant predictor of race time. Does this make sense?

Looking Further

- Let's look at the regression line using the posterior median values of β_0 and β_1.
- And let's plot (in gray) all of the regression lines if we regressed time vs. age **separately** for each runner.
- The line from the pooled regression is almost flat, but the lines for the individual runners are steeper, showing that race times worsen as the runners age.
- Let's examine how poorly the overall line approximates this aging trend for three specific runners 1, 20, and 22 (see R plot).
- Note for these three runners, the aging trend is very different.

Drawbacks of the Complete Pooling Model

- Why does the Complete Pooling Model not work well?
- It assumed the data are all independent, when data values coming from the same person are correlated.
- It assumes the aging trend is the same across all runners, but in reality it may be different for different runners.
- The unfortunate consequence is that we get misleading conclusions about the regression relationship between Y and X itself and its significance.

The No-Pooling Model

- Another approach is not to pool at all, and to fit separate regressions to each runner in the data set.

$$Y_{ij}|\beta_{0j}, \beta_{1j}, \sigma \sim N\left(\mu_{ij}, \sigma^2\right) \quad \text{with} \quad \mu_{ij} = \beta_{0j} + \beta_{1j} X_{ij}.$$

- Each runner ($j = 1, \ldots, n$) is allowed to have his/her own intercept AND slope.
- Note this model is much more complicated! Instead of 2 regression coefficients, we have $2n$ coefficients.
- Based on an R plot for 3 example runners, it looks great!

Drawback to No-Pooling Model

- So what's wrong?
- These models are useless for other runners besides the runner the model was fit for.
- They can't be used to predict running times for a new runner.
- Also, can they be used to make a general statement about how age affects race time in the population?
- No, because the slopes here are different for each runner.

Drawback to No-Pooling Model

- With the no-pooling model: We cannot reliably generalize or apply the group-specific (runner-specific, in our example) models to groups (runners) outside those in our sample.
- This no-pooling approach assumes that one group doesn't contain relevant information about another, which ignores potentially valuable information.

Other examples of Hierarchical Data

- Recall the examples of multilevel data mentioned earlier:
- Imagine students in several schools taking the same achievement test.
- The test scores of students in the same school may be correlated.
- The schools would be the groups, so we could consider school-specific models, like the runner-specific models we just looked at (with the same drawbacks).
- The hierarchical structure could get complicated: Imagine students grouped within classrooms, grouped with schools, grouped within districts, grouped within states!
- Usually we keep the number of levels around 2 or 3 at the most.

Why Use Hierarchical Data?

- Often there are practical advantages to collecting hierarchical data.
- Suppose we need 30 measurements of pH levels of rain in a location.
- Should we take 1 measurement from each of 30 rainfalls? (Independent data)
- Or should we take 6 measurements from each of 5 rainfalls? (Correlated within rainfall)
- Data can be grouped by spatial location too: Imagine taking multiple (say, 10) soil measurements at each of 8 fields (correlated within field).
- Easier than taking 1 soil measurement at 80 different fields (independent).
- The key thing is to allow your statistical model to account for the correlation, and not ignore it and pretend the data are independent when they're not.

A Happy Medium: Partial Pooling

- Hierarchical models use partial pooling: Their results tend to be partly between these two extremes.
- Idea: Each group is unique, so we should retain group information in the model.
- But we can **borrow information** across groups to better estimate the parameters of interest.
- Hierarchical models allow us to assess both *within-group variability* (how similar are observations within a group?)
- . . . and *between-group variability* (how different are the various groups?)
- The Bayesian framework is well-suited to work with hierarchical models.