STAT 535: Chapter 17: Normal Hierarchical Models with Predictors

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

Spring 2022

Hierarchical Models with Predictors

- Hierarchical regression models allow us to account for the grouping structure present in hierarchical data, but allow involve predictor variables that can help us improve predictive accuracy.
- We will focus on the Cherry Blossom Road Race data we have seen before.
- In Chapter 15, we used a "complete-data" normal regression model that related net race time and age, which did not account for the grouping structure, and it seemed to not reflect the true nature of the way race time depends on age.
- We now present a couple of more complicated models that account for the hierarchical structure of the data.

Hierarchical Model with Varying Intercepts

- This model allows each group (each runner, in this example) to have its own intercept β_{0j}, $j = 1, \ldots, n$:

$$Y_{ij} | \beta_{0j}, \beta_1, \sigma_y \sim N(\mu_{ij}, \sigma_y^2) \quad \text{where} \quad \mu_{ij} = \beta_{0j} + \beta_1 X_{ij}.$$

- In this model, we assume the slopes of the group-specific regression equations are the same across groups: β_1.
- In our example, the regression lines for the various runners, when plotted on a graph, are parallel lines.
- The intercepts of the runner-specific regression lines are different, reflecting the fact that some runners are faster overall and some runners are slower.
- The rate at which runners' expected times change as they age is assumed to be the same across runners (may not reflect reality).

Parameters in this Model

- β_{0j} is the group-specific intercept for runner j
- β_1 is the global coefficient of age
- σ_y is the measure of within-group variability.
- σ_y measures how spread out are the race times for a runner above or below his/her true regression line.
- This spread of "error terms" is assumed to be the same for all runners.

Layer 2 of Model with Varying Intercepts

- The second layer of the model gives the distribution of the β_{0j}'s:

$$\beta_{0j}|\beta_0, \sigma_0 \overset{ind}{\sim} N(\beta_0, \sigma_0^2)$$

- β_0 is the global average intercept across all runners.
- σ_0 is between-group variability in intercepts β_{0j}, which measures the amount of variation in baseline speeds from runner to runner (how vertically separated the runner-specific regression lines are on the graph).

Model with All Priors

- Finally, we put priors on the parameters in the usual way:

$$
\begin{aligned}
Y_{ij}|\beta_{0j}, \beta_1, \sigma_y &\sim N(\mu_{ij}, \sigma_y^2) \\
\text{with} \quad \mu_{ij} = \beta_{0j} + \beta_1 X_{ij} & \quad \text{(regression model} \\
& \quad \text{ WITHIN runner } j) \\
\beta_{0j}|\beta_0, \sigma_0 &\overset{ind}{\sim} N(\beta_0, \sigma_0^2) \\
\text{(variability in baseline speeds} & \quad \text{ BETWEEN runners)} \\
& \quad \text{(priors on global parameters:)} \\
\beta_{0c} &\sim N(m_0, s_0^2) \\
\beta_1 &\sim N(m_1, s_1^2) \\
\sigma_y &\sim \text{Exp}(l_y) \\
\sigma_0 &\sim \text{Exp}(l_0).
\end{aligned}
$$

Estimating the Model

- We can simulate from the posteriors and estimate the model using the `stan_glmer` function in the `rstanarm` package (see R example).
- Note the 80% credible interval for β_1 is $(1.02, 1.58)$.
- All positive values in the credible interval, which implies that runners slow down on average as they age.
- In the complete pooling model, recall the credible interval for β_1 included 0, which didn't make sense.

Variation among Runners

- We can also display the variation in intercepts among the runners.
- For example, compare runners 4 and 5 via the credible intervals for their β_{0j} values, and plots of posterior draws of their estimated regression lines.
- We see runner 4 is slower than runner 5.
- We can plot the runner-specific models for all 36 runners (see R plot).

Examining Sources of Variability

- Comparing σ_y and σ_0 tells us about how much of the overall variation in race times is due to differences **between runners** (as opposed to differences in race times within the same runner).
- Our estimate of σ_0 is 13.3 and our estimate of σ_y is 5.25.
- $\dfrac{13.3^2}{13.3^2 + 5.25^2} = 0.867$, so 86.7% of the variation in race times is due to variation **between runners**.

Hierarchical Model with Varying Intercepts and Slopes

- The model we just fit assumed the slopes of each runner's regression line (which measures the rate at which race time changes with age, on average) is the same for each runner.
- This likely does not reflect reality: Some runners worsen quickly as they age, others worsen gradually as they age, and some even improve with age! (See R plots)
- The **Varying Intercepts and Slopes Model** allows each runner to have a different intercept β_{0j} and a different slope β_{1j}.

Varying Intercepts and Slopes Model

The formal model is a bit complicated:

$$
\begin{aligned}
Y_{ij} | \beta_{0j}, \beta_{1j}, \sigma_y &\sim N(\mu_{ij}, \sigma_y^2) \quad \text{where} \quad \mu_{ij} = \beta_{0j} + \beta_{1j} X_{ij} \\
\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} | \beta_0, \beta_1, \sigma_0, \sigma_1 &\sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \Sigma \right) \\
\beta_{0c} &\sim N(100, 10^2) \\
\beta_1 &\sim N(2.5, 1^2) \\
\sigma_y &\sim \text{Exp}(0.072) \\
\Sigma &\sim \text{(decomposition of covariance)}.
\end{aligned}
$$

Covariance Matrix of the β's

- The values in the covariance matrix Σ measure the variances and the covariance between the regression parameters β_0 and β_1.
- If β_0 and β_1 have a strong correlation, then runners who are especially fast (low β_0) or slow (high β_0) tend to have a strong effect of age on race time (very negative or very positive β_1).
- The precise interpretation of such a correlation would depend on the sign of β_1.

Variance Components of the β's

- The variance components in Σ measure the relative proportion of this variability between groups that's due to differing intercepts vs differing slopes:

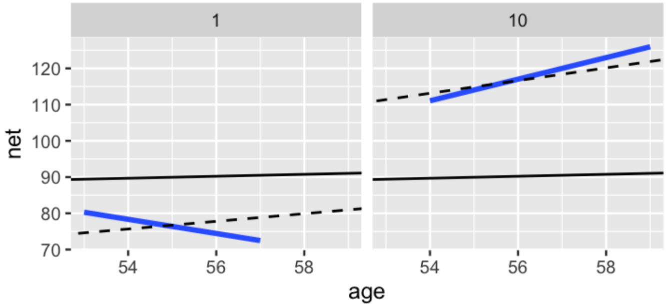
$$\pi_0 = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2} \quad \text{vs} \quad \pi_1 = \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2}$$

- If the first fraction is large, that means that most of the variation among the runner's regression lines is due to the differences in intercepts.
- If the second fraction is large, that means that most of the variation among the runner's regression lines is due to the differences in slopes (aging trends).

Posterior Simulation

- The posterior analysis is again done using `stan_glmer`.
- There are 78 parameters, so it is slow.
- The posterior median model is similar to the one for the random intercepts model: $\hat{Y} = 18.5 + 1.32 \times \text{age}$
- The advantage of this model is seen when we examine the runner-specific models with the different β_{0j} and β_{1j} parameters (see R examples and plots of runner-specific lines).

Plots for Two Example Runners



Shrinkage in this Model

- We see (examine the plots just for runners 1 and 10) that the runner's trend line from the no-pooling model (blue line) is shrunk toward the OVERALL regression line from the complete-pooling model (solid black line) to produce the runner's estimated regression line from the partial-pooling model (dashed black line).
- The assumption is that information from the other runners (which is carried in the overall complete-pooled regression line) should inform the estimated regression line for runner j.
- The hierarchical model assumes the data for a single runner (especially if there are few data points for that runner) don't tell the whole story about that runner's true trend line.
- Information from the broader population of runners should also play a role.
- This shrinkage, this balancing of information from two sources, is similar to the Bayesian paradigm of balancing the information between the observed data and the prior.

Model Selection

- So we have our choice of models: (1) complete pooling; (2) no pooling; (3) varying intercepts; and (4) varying intercepts and slopes.
- We can use our intuition to help decide which model we should use, but we can formally check the models' fit using `pp_check`.
- The `prediction_summary` output and the ELPD values can help us compare prediction accuracy for the competing models.
- See R example: What do the criteria say about the choice between "varying intercepts" and "varying intercepts and slopes" models?

Posterior Prediction of Race Time for a New Individual

- If we want to predict the race time for a new individual of a certain age, we can use the `posterior_predict` function with our chosen model.
- We could also predict the race time for someone in our sample, at a different age than we have already observed data for that person.
- For example, consider predicting the race time at age 61 for: runner 1; runner 10; and a new runner, Miles.
- Since we have no previous data on Miles, the prediction of Miles's age-61 race time will be much less precise (see R plots).
- Section 17.7 has an interesting example on the `spotify` data set in which they use a model to predict a song's danceability using its genre and its "valence" (mood). Read about it on your own!