Understanding and summarizing hierarchical models

Patrick Breheny

April 4

Introduction

- Now that we have fit a variety of hierarchical models, let's talk a bit more about their parameters to make sure we understand them and discuss how to communicate the results of these analyses to others
- Typically, in a regression analysis, we summarize a fitted model by summarizing the regression coefficients along with their standard errors (or, in a GLM, some function of the regression coefficients such as the odds ratio)
- In hierarchical models, however, we can easily have large numbers of parameters (*e.g.*, in the radon example, we had 85 intercepts and, when we allowed varying slopes as well, 85 slopes also)

Parameters and hyperparameters

- It would be unrealistic to describe every single one of these parameters in an analysis
- However, typically there will be a much smaller number of hyperparameters describing the distribution of these group-level slopes and intercepts (and interactions, in the height-earnings example)
- Also, as we have seen, it is usually both feasible and helpful to present the varying slopes/intercepts graphically, in order to illustrate the group-level model

Uncertainty vs. Variability

- It is critical in statistics to distinguish between *uncertainty* and *variability*, but this can be more difficult in hierarchical models
- Uncertainty reflects a lack of knowledge about a parameter
- Variability, on the other hand, refers to underlying differences between groups or between individuals
- The key distinction between the two is that if we had an infinite amount of data, uncertainty will disappear, but variability will not

Example: Varying-intercept radon model

For example, let's go back to our first hierarchical model, the varying-intercept radon model:

	mean	sd	2.5%	97.5%
a[1]	1.19	0.25	0.68	1.68
a[2]	0.93	0.10	0.73	1.12
a[3]	1.48	0.26	0.96	2.00
. . .				
a[84]	1.59	0.18	1.25	1.94
a[85]	1.39	0.28	0.83	1.94
b	-0.69	0.07	-0.83	-0.55
mu	1.46	0.05	1.36	1.57
sigma.y	0.76	0.02	0.72	0.79
sigma.a	0.33	0.05	0.24	0.43

Uncertainty in the varying-intercept radon model

- The posterior standard deviation of α_1 is 0.25; this is uncertainty: it tells us that the actual mean (log) radon level in the basements of county 1 might not be 1.2, but could be as low as 0.7 or as high as 1.7
- County 1 had 4 measurements (3 basement measurements); if we had 100 measurements, this uncertainty would be lower
- Indeed, county 2 had 52 measurements (49 basement measurements), and our uncertainty about its mean basement radon level was lower (posterior standard deviation 0.10)

Variability in the varying-intercept radon model

- The posterior mean of σ_y is 0.76; this is variability: it tells us that house measurements (on the same floor, within the same county) vary from one another by about 0.8
- It is worth noting that σ_y here encompasses both house-to-house variability within a county as well as any measurement error involved in recording radon levels; it is impossible here to distinguish between the two without repeatedly measuring individual houses (which would introduce another level in our hierarchy)
- The posterior SD of σ_y is 0.02; this is uncertainty about variability: in this case, we are pretty sure exactly how much houses vary within counties
- Classroom exercise: Compare counties 2 and 7, Anoka vs. Blue Earth, in terms of uncertainty and then variability

Getting some new measurements

Suppose we go out and obtain another 400 measurements of houses in county 3; what will happen to:

- Posterior mean of α_1?
- Posterior SD of α_1?
- Posterior mean of α_3?
- Posterior SD of α_3?
- Posterior mean of β?
- Posterior SD of β?
- Posterior mean of μ?
- Posterior SD of μ?
- Posterior mean of σ_y?
- Posterior SD of σ_y?
- Posterior mean of σ_α?
- Posterior SD of σ_α?

Getting some new measurements (cont'd)

Now suppose that there are, say, 500 counties in Minnesota, and we go out and obtain a few measurements in each county for 100 more counties; what will happen to:

- Posterior mean of α_1?
- Posterior SD of α_1?
- Posterior mean of β?
- Posterior SD of β?
- Posterior mean of μ?
- Posterior SD of μ?

- Posterior mean of σ_y?
- Posterior SD of σ_y?
- Posterior mean of σ_α?
- Posterior SD of σ_α?

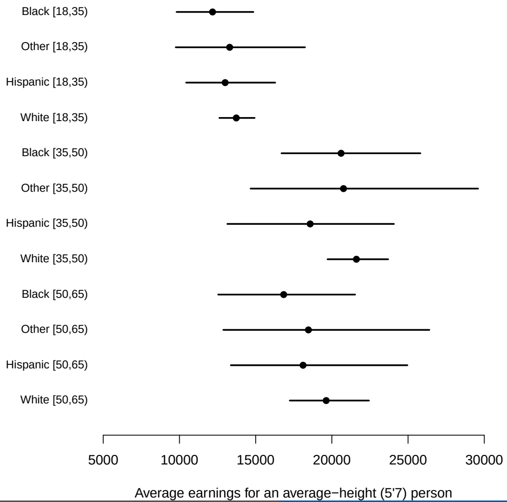
Remarks

- Note that in all of these scenarios, we would not expect any of the estimates to change (they could change, of course, but we would not expect any of them to systematically increase or decrease
- We would expect uncertainty in all parameters to decrease – sometimes by a lot, sometimes barely at all
- In particular, even if we collected an infinite amount of data in county 3, or on an infinite number of counties, uncertainty would remain about some parameters
- For all uncertainty to disappear, we would need both the number of counties sampled as well as the number of houses per county to be going to infinity

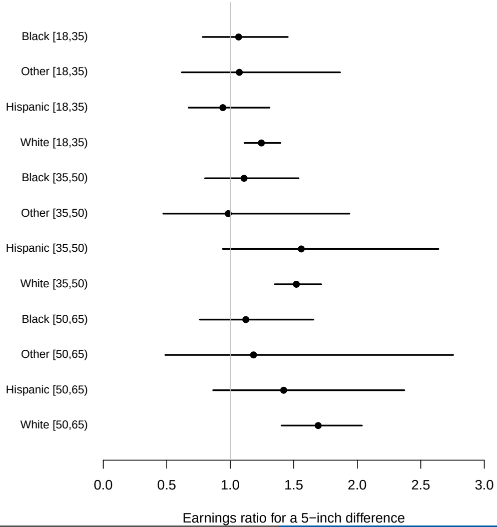
Height and earnings

- Let's consider another example and look at height and earnings
- Here, the individual estimates $\alpha_{j,k}$ and $\beta_{j,k}$ are of direct interest, and we really want to look at, report, and consider all of them (we probably have no such desire for all the counties in Minnesota)
- This means 24 parameters, each with a mean/SD/interval, which would make for a cumbersome table
- This would be a good time to consider an interval plot (sometimes called a forest plot)

Average earnings



Earnings vs. Height



Remarks

- Note that we can be quite sure of a positive association between earnings and height for whites of all ages
- We have no similar certainty for any other age/ethnic group combination, although the older and middle-aged Hispanic groups are close
- It is interesting to compare the width of our 95% PI to that of an "independent-parameters" 95% CI for Hispanics aged 50-64:

Earnings ratio
5-inch difference

	Estimate	2.5%	97.5%
Hierarchical	1.4	0.9	2.4
Independent	1.7	0.6	4.4

Remarks (cont'd)

- Even though both models have interactions and therefore allow separate slopes for each age-ethnic group combination, the width of the interval in the hierarchical model is much narrower
- The reason, of course, is that we are borrowing information across other ethnic groups of the same age (50-64) as well as across Hispanics of other ages to assist us in estimating the height-earnings relationship in older Hispanics

Height vs. Earnings: Other parameters

	mean	sd	2.5%	97.5%
mu[1]	9.75	0.22	9.29	10.23
mu[2]	0.04	0.09	-0.13	0.22
sigma.y	0.87	0.02	0.84	0.91
sigma.a[1]	0.30	0.24	0.08	0.91
sigma.a[2]	0.10	0.07	0.04	0.29
sigma.e[1]	0.12	0.08	0.04	0.33
sigma.e[2]	0.09	0.05	0.04	0.22
sigma.t[1]	0.11	0.06	0.04	0.28
sigma.t[2]	0.06	0.02	0.03	0.12

Uncertainty about μ_2

- It is interesting to note here that our uncertainty about the average slope (0.09) is much larger than our uncertainty about the slope in several of our individual age/ethnic group combinations (some of which were as low as 0.01)
- Indeed, we have tremendous uncertainty about whether the "average slope" is even positive
- Note that "average slope" here means the height-earnings relationship for the "average ethnic group" and the "average age"
- The idea of an "average ethnic group" is perhaps a little murky here, so it is worth thinking about this a little further

Random ethnic groups

- In our model, we envision randomly drawing ethnic groups from an infinite source of new ethnic groups
- This is of somewhat dubious meaning, especially since we already have a category for “other”
- Furthermore, μ_1 and μ_2 would refer to the average across all of these new ethnic groups
- Does this make sense, given that, for example, whites make up 82% of our sample and “other” makes up only 2%?

Finite-population weighted average

- We might instead wish to consider the weighted average of the four ethnic groups without regard to hypothetical new ethnic groups (what you might call a "finite-population weighted average"; we'll talk more about finite populations later)
- In our data, the percentages

Other	Hispanic	Black	White
2	6	10	82

- Assuming this is a random sample and ignoring the possibility of differential nonresponse across ethnic groups, we could use these as estimates of the fraction of the population that belongs to each group, and therefore as weights in our weighted average

Results

- We can use these weights, then, to construct the quantities of interest

  $$\omega_k = \sum_{j=1}^{4} w_j \beta_{j,k},$$

  and as usual, since we have draws of β, this gives us draws from the posterior of ω (see code for details of doing this in R; alternatively one could create ω in JAGS)
- Note than when we do this, our uncertainty about the population average is indeed lower than our uncertainty about any of the individual ethnic groups
- Furthermore, we can be sure that the population-average association between height and earnings is positive in all three age groups

Incorporating uncertainty about w

- You may be thinking that this isn't entirely sound, in that we are treating the weights $\{w_j\}$ as fixed, even though we are estimating them from the data
- The weights, too, could be incorporated into the Bayesian analysis; letting $\mathbf{e} = (E_1, E_2, E_3, E_4)$ denote the counts in each ethnicity category, a natural distribution would be

  $$\mathbf{e} \sim \text{Multinom}(\boldsymbol{\pi}, n),$$

  where, if you have not seen it before, the multinomial distribution is just the multivariate extension of the binomial distribution
- Of course, now we need to put a prior on $\boldsymbol{\pi}$

The Dirichlet distribution

- This is a good excuse to go off on a minor tangent and discuss the Dirichlet distribution, as it is useful to know about and widely used in a variety of fields that focus on the analysis of discrete data such as genetics and natural language processing
- The *Dirichlet distribution* is simply the multivariate extension of the beta distribution: $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ implies

$$p(\boldsymbol{\pi}) = \Gamma\left(\sum_j \alpha_j\right) \prod_j \frac{\pi_j^{\alpha_j}}{\Gamma(\alpha_j)}$$

Conjugacy

- The Dirichlet and multinomial distributions are easily set up in JAGS:

```
E ~ dmulti(pi, n)
pi ~ ddirch(pi0)
```

- The Dirichlet distribution is conjugate to the multinomial distribution; in particular,

$$\pi | \mathbf{e} \sim \mathrm{Dir}(\pi_0 + \mathbf{e})$$

- This is exactly analogous to the beta-binomial relationship, which is indeed just a special case of the Dirichlet-multinomial conjugate relationship
- Base `R` does not provide a `rdirichlet` function, but several R packages do, and I have put one online as well in the usual `fun.R` file

Results

- The posterior for $\boldsymbol{\pi}|\mathbf{e}$:

	50%	2.5%	97.5%
Black	0.10	0.08	0.12
Hispanic	0.06	0.05	0.07
Other	0.02	0.01	0.03
White	0.82	0.80	0.84

- As you might imagine, with such little uncertainty in the weights, the interval for the finite-population weighted average is essentially unchanged
- For example, in the 50-64 age group, the interval goes from (1.351, 1.886) to (1.350, 1.887)

Problems with σ_α in a finite population

- If the idea of randomly drawing new ethnic groups from an infinite population of different ethnic groups is of questionable meaning, then how should we interpret the variance parameter for those coefficients?
- Similarly, given that there are only 85 counties in Minnesota, we really only care about the variability among those 85 counties, and don't care about variation in new, hypothetical counties that could be in Minnesota but aren't
- To address these questions, we need to distinguish between two notions of variation among parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_J\}$ arising from a common distribution with variance σ_α

Superpopulations

- The parameter σ_α is the standard deviation of the *superpopulation*, the infinite population from which the α_j's are drawn according to the model
- The concept of a superpopulation is obviously critical to making statements about new groups that are not in the sample
- However, the superpopulation is an important ingredient in the model even if the actual population $\{\alpha_j\}$ is finite, as it controls the way in which information is borrowed across groups

Finite-population standard deviation

- The finite-population standard deviation, on the other hand, is concerned only with variation among the existing groups:

$$s_\alpha = \sqrt{\frac{1}{J}\sum_j (\alpha_j - \bar{\alpha})^2},$$

  where $\bar{\alpha}$ is the sample mean of the α_j's

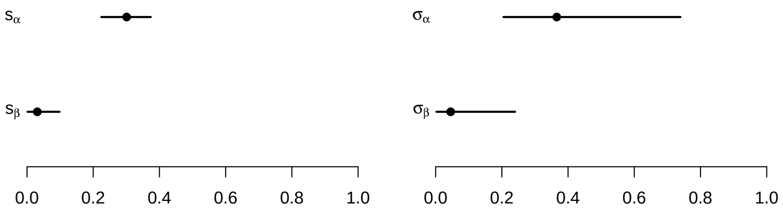
- As usual, since we have draws from the posterior of $\{\alpha_j\}$, it is straightforward to obtain draws from the posterior of s_α

Example: Radon

- For example, the posterior for s_α in the varying-intercept radon example has median 0.32 and 95% posterior interval (0.25, 0.40)
- Compare this with the posterior for σ_α, which had median 0.33 and 95% interval (0.25, 0.42)
- In this example, with a lot of counties in our sample, there really isn't much difference between the two quantities
- However, note that the posterior interval is a bit narrower when we don't have to consider the possibility of new counties

Superpopulation vs. finite-population: Flight data

This phenomenon is more pronounced with fewer groups, as in the flight simulator study:



Superpopulation vs. finite population for μ

- Recall that in the flight study, our posterior for μ had median 0.44, with 95% interval (0.13, 0.76)
- Here, μ had an interpretation as the average recovery rate across the superpopulations of both training groups and scenarios
- Alternatively, we could consider the average recovery rate across the finite population of training groups, $\mu + \bar{\beta}$, which has a posterior median of 0.44 and 95% PI (0.14, 0.75)
- Finally, we could consider the average recovery rate across the finite population of scenarios and training groups, $\mu + \bar{\alpha} + \bar{\beta}$, which has a posterior median of 0.44 and a 95% PI of (0.37, 0.51)
- Note that this is even narrower than the 95% t-intercal of (0.32, 0.56)

Fixed and random effects

- As Gelman and Hill state, “much of the statistical literature on fixed and random effects can be fruitfully re-expressed in terms of finite-population and superpopulation inferences”
- In some contexts (ethnic groups, training groups), the finite-population interpretation is more meaningful
- In others (scenarios (probably), subjects (definitely)) the interest lies in the superpopulation
- The shortcoming of the fixed/random effect terminology is that it conflates the hierarchical modeling with the population – it may very well be of interest, however, to make inferences concerning the finite population even though we have used a hierarchical model