Further extensions: Non-nested models and generalized linear models

Patrick Breheny

April 2

Flight simulator study

- Today we will consider some extensions involving the application of hierarchical models to problems outside the "repeated measurements on units" structure
- First, we consider a study from the field of aviation involving what are known as human factors
- In the study, which took place in a flight simulator, pilots were exposed to what is known as an "aircraft upset", the technical term for a loss of aircraft control
- Ideally, the pilots would recover from the upset and manage to land the plane safely, but sometimes they would be unable to recover and the plane would crash

Training levels and scenarios

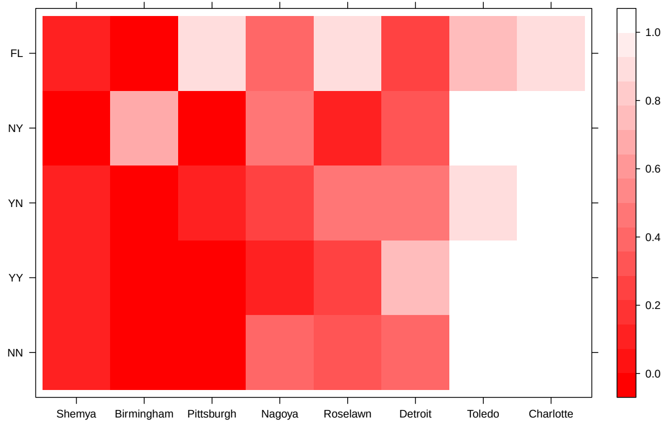
- The pilots in the study were in one of five groups, depending on their training: “YY” means they received both airplane upset training and aerobatic flight training, “YN” means upset training but no aerobatic training, and so on (the fifth category, “FL”, refers to pilots who received in-flight traning)
- There were also eight different upset scenarios, each taking place near a different (simulated) airport

Two-way ANOVA

- One way to think about this data is that, if we let y_{jk} denote the recovery proportion in scenario j for group k, then we have $8 \times 5 = 40$ observations, one for each combination of scenario and group
- This is referred to as a *two-way ANOVA without replication*, since we only have a sample size of 1 per combination
- In reality, we have multiple observations per condition, one for each pilot; we'll take another look at this data using a hierarchical logistic regression model later in the lecture

Descriptive statistics

Red indicates a recovery proportion of 0, white a recovery proportion of 1



Model

- A reasonable model for the data is the following:

$$\begin{aligned} y_{jk} &\sim \mathrm{N}(\mu + \alpha_j + \beta_k, \sigma_y^2) \\ \alpha_j &\sim \mathrm{N}(0, \sigma_\alpha^2) \\ \beta_k &\sim \mathrm{N}(0, \sigma_\beta^2) \end{aligned}$$

- Note that we cannot introduce, say, a μ_α parameter, as that would render the model non-identifiable

Comparison with traditional ANOVA

- It is worth comparing this model to a traditional, "independent parameters" ANOVA approach, in which, without replication, it is not possible to simultaneously estimate σ_y^2, σ_α^2, and σ_β^2
- We avoid that problem here by assuming that the scenarios are related to one another – *i.e.*, that knowing outcomes in 7 scenarios tell you something about the 8th – as are the groups
- This assumption, formally known as the assumption of "exchangeability", keeps the problem identifiable and enables us to estimate all three variances

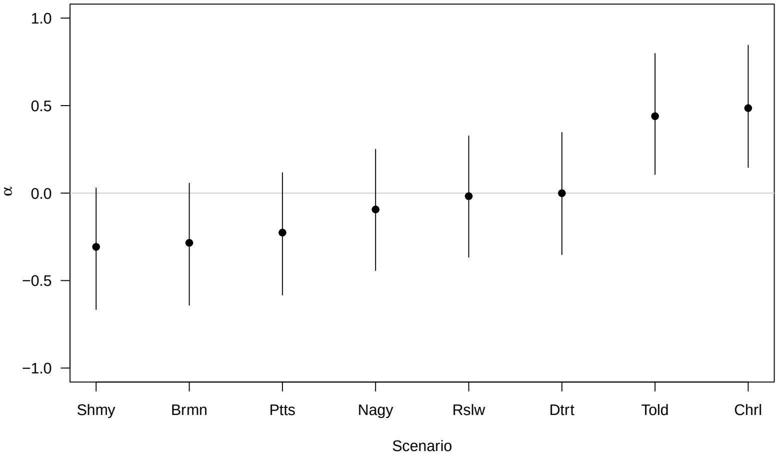
Posterior: μ

- The posterior for μ has median 0.442, with 95% interval (0.131, 0.764)
- Note that this has the same center (0.442) as a simple normal-theory interval (ignoring scenarios and groups), but is considerably wider; the 95% t-interval is (0.322, 0.561)
- This is appropriate: although we can be fairly confident that the recovery proportion for these groups and these scenarios is between 35% and 55%, we would have to expand that interval if we considered possible recovery proportions for new training groups and new scenarios

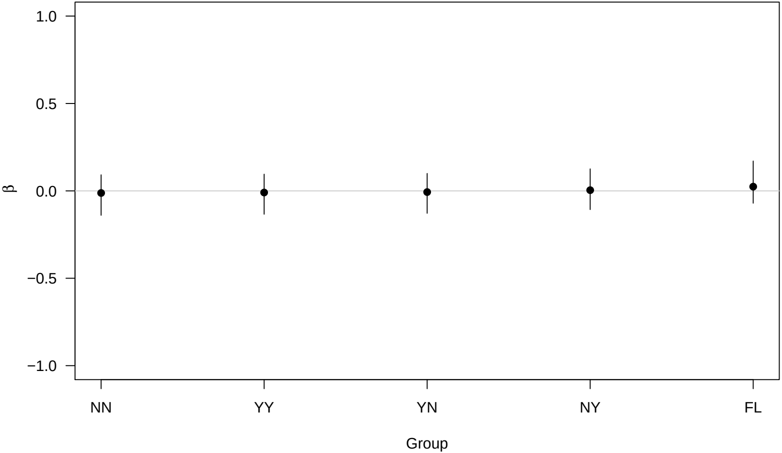
Posterior: Variances

- Our posterior means for the standard deviation parameters are: $\sigma_y = 0.23$, $\sigma_\alpha = 0.39$, and $\sigma_\beta = 0.06$
- Thus, the variability between the scenarios is very large – larger, even, than the variability among individual measurements – but very little variation among groups
- To put it another way, 75% of the variability among recovery rates is due to the scenarios, 22% results from inherent variability in the measurements, and just 3% is due to the training groups

Posterior: α



Posterior: β



Logistic regression model

Modifying this model into a hierarchical logistic regression model is straightforward, as the only change is for the likelihood portion of the model – the priors and hyperpriors remain the same:

$$\begin{aligned} y_i &\sim \mathrm{Binom}(\theta_i, 1) \\ \log\left(\frac{\theta_i}{1-\theta_i}\right) &= \mu + \alpha_{j[i]} + \beta_{k[i]} \\ \alpha_j &\sim \mathrm{N}(0, \sigma_\alpha^2) \\ \beta_k &\sim \mathrm{N}(0, \sigma_\beta^2), \end{aligned}$$

where $\alpha_{j[i]}$ and $\beta_{k[i]}$ refer to the scenario and group, respectively, that observation i belongs to

Comparison

- In this particular example, with a roughly balanced design (nearly identical sample sizes in each group), the qualitative conclusions of the logistic regression model are quite similar to the ANOVA approach
- However, the two models are not identical: consider the estimate of the posterior mean recovery for the Toledo scenario
- The ANOVA approach has a posterior median of 0.88, with a the somewhat nonsensical 95% interval of (0.66, 1.10)
- The logistic regression approach yields a posterior median of 0.90, with a more reasonable 95% interval of (0.76, 0.97)

Earnings and height

- One additional model/example for the day: let's consider modeling the relationship between income and height, while allowing varying slopes and intercepts that may depend on both ethnicity and age
- Obviously, height is not the dominant factor that influences income; however, studies consistently show positive correlations between them
- Various explanations have been proposed, ranging from discrimination against short people to the notion that taller people, used to having others "look up" to them, have more experience in leadership roles

Age and ethnic group

- The data (from a 1994 survey of American adults) separates individuals into $J = 4$ ethnic groups (white/black/hispanic/other)
- Following Gelman & Hill's approach, we will consider categorizing age into three groups: 18-34, 35-49, and 50-64
- In addition, we will allow age and ethnicity to have interactions as well as main effects on earnings
- Finally, because incomes are considerably right-skewed, we will model the log of earnings rather than income directly

Model, version 1

We can write our model as follows:

$$\begin{aligned} y_i &\sim \mathrm{N}(\alpha_{j,k} + \beta_{j,k} z_i, \sigma_y^2) \\ \boldsymbol{\theta}_{j,k} &= \boldsymbol{\mu} + \boldsymbol{\gamma}_j + \boldsymbol{\delta}_k + \boldsymbol{\lambda}_{jk} \\ \boldsymbol{\gamma}_j &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma) \\ \boldsymbol{\delta}_k &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\delta) \\ \boldsymbol{\lambda}_{j,k} &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\lambda), \end{aligned}$$

where $\theta_{j,k} = (\alpha_{j,k}, \beta_{j,k})$, the $\boldsymbol{\Sigma}$ terms may be given Wishart/scaled Wishart priors, and $\boldsymbol{\mu}$ is given a reference prior

Model, version 2

An equivalent formulation is to express the interactions as correlations:

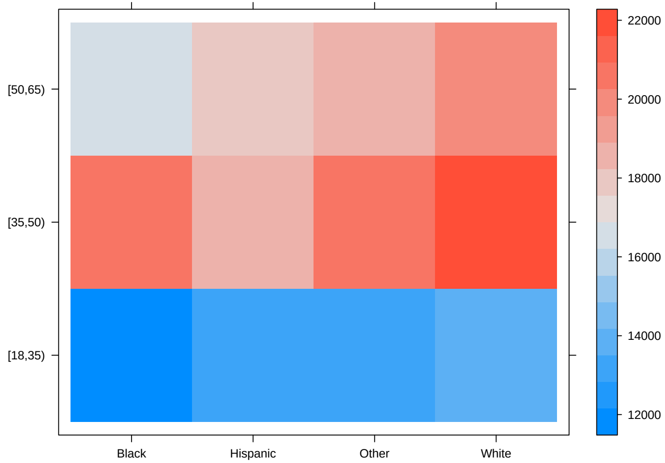
$$\begin{aligned} y_i &\sim \mathrm{N}(\alpha_{j,k} + \beta_{j,k} z_i, \sigma_y^2) \\ \boldsymbol{\theta}_{j,k} &\sim \mathrm{N}(\boldsymbol{\mu} + \boldsymbol{\gamma}_j + \boldsymbol{\delta}_k, \boldsymbol{\Sigma}_\theta) \\ \boldsymbol{\gamma}_j &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma) \\ \boldsymbol{\delta}_k &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\delta) \end{aligned}$$

where $\boldsymbol{\Sigma}_\theta$ is equivalent to $\boldsymbol{\Sigma}_\lambda$ in the previous slide

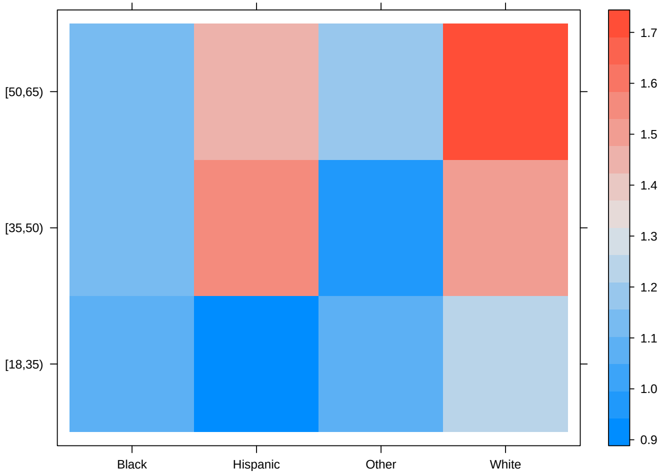
Centering

- It is a good idea here to center height (subtract off its mean) before fitting the model
- Failing to do so results in α estimating an intercept for a person with a height of zero inches
- Not only would this render α virtually meaningless, but also all of the σ_α terms would be impossible to interpret
- Furthermore, $\{\alpha_{j,k}\}$ and $\{\beta_{j,k}\}$ would be highly correlated in the un-centered model, potentially resulting in slower mixing

Posterior: α



Posterior: β (for a 5-inch difference)



Variance components: α

- The variance components for the intercept are as follows:

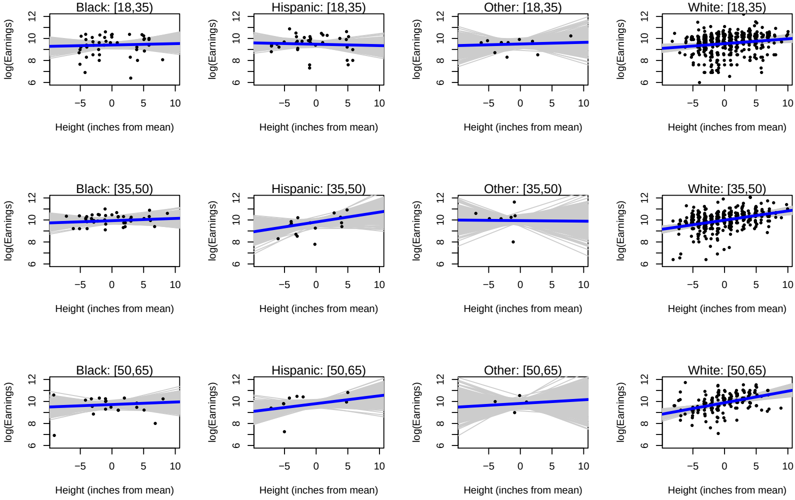
Error	Age	Ethnicity	Interaction
0.87	0.10	0.02	0.01

- Among the factors considered, age certainly plays a larger role than the others
- It is worth noting, however, that the vast majority of variation in income cannot be explained by this model

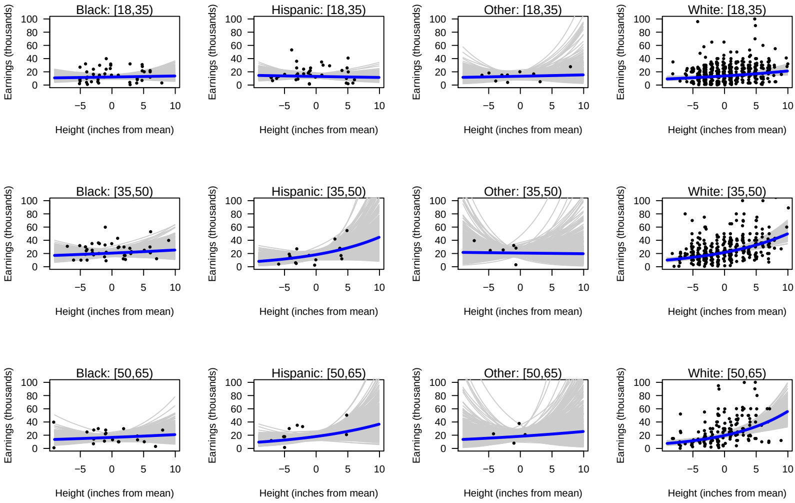
Further commentary on the data-level variance

- Indeed, the posterior mean for σ_y^2 is 0.87, implying that the model can only predict income to within a factor of about $e^{0.87} = 2.4$
- In other words, we might predict that an individual will make \$20,000, but they could easily make \$48,000 or just \$8,333
- This should not come as a huge surprise, given that ethnicity and age are the only variables in the model

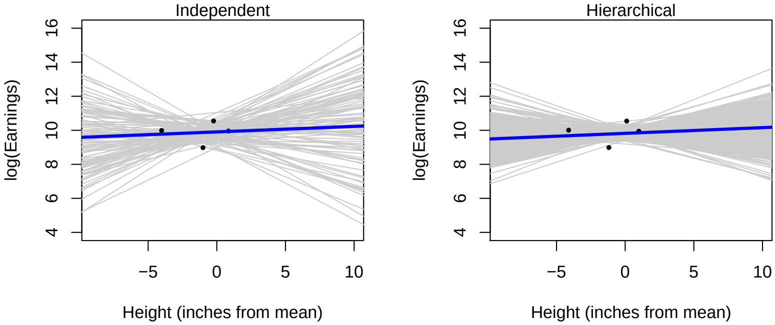
Posterior: Regression lines



Posterior: Regression lines on original scale



Hierarchical vs. independent: Other, Age 50-64



Independent: $SD_p(\alpha) : 0.45$, $SD_p(\beta) : 0.21$
Hierarchical: $SD_p(\alpha) : 0.18$, $SD_p(\beta) : 0.09$