Stat 414 - Day 15
Shinkrage Estimates (4.8)

Last Time: Fixed vs. Random Effects

- If categories themselves aren't so much of interest, but want to consider the "grouping units" as a random sample from a larger population, can treat as random effects
  - "effect" = deviation from overall mean
- $E(Y_{ij}) = \beta_0 + u_j + \epsilon_{ij}$ where we are assuming $\epsilon_{ij} \sim N(0, \sigma^2)$ and $u_j \sim N(0, \tau^2)$
- Benefits include fewer parameters to estimate and generalizabiity to larger population of units. Also induces a non-zero correlation between two observations from the same Level 2 units (this allows us to model dependence within the groups)

Even though we say we are not all that interested in the individual u_j and that they aren't really parameters but "unobservable latent effects," we do still get estimates for them that are used to estimate τ and it might still be interesting to explore those estimates (e.g., do they appear to be normally distributed?) But how are they estimated differently?

Example: Suppose we have batting averages for 6 players over several seasons.

(a) What is the overall mean batting average for these 43 observations?

(b) Record the sample sizes, means, and standard deviations for each player:

	Anderson	Jones	Mitchell	Rodriguez	Smith	Suarez
n_j						
$\bar{y}_j$						
s_j						

(c) Do you expect a high or low ICC value? Explain.

(d) Opinion: Do you really think Rodriguez and Suarez are that much better than everyone else? What else could be going on? Which of these averages do you find the most/least "trustworthy"? Why?

Treat player as a fixed effect (with effect coding) and fit a linear model.

(e) What is the estimated overall mean batting average across all players? Is this the same as in (a)? Why not? Where does it come from? (*Hint*: Why is it larger?) ICC?

(f) What does this model estimate for the "effect" of each player? (R lists them alphabetically)

1.Anderson	2.Jones	3.Mitchell	4.Rodriguez	5.Smith	6.Suarez

(g) What does this model estimate for Jones' population mean batting average? (*Hint*: overall mean + Jones' effect. Does this number look familiar?)

Now treat player as a random effect.

(h) What is the estimate of the overall mean?

(i) What is the intraclass correlation coefficient?

The output no longer gives us the estimated effects for the players, but R does store them for us.

```
ranef(model2)
```

(j) How do these player estimated effects compare to model 1?

We can also convert the effects to the predicted values for each player

```
fits=predict(model2); fits
tapply(bball$Batting, bball$Player, mean)
tapply(fits, bball$Player, mean)
```

(k) Whose estimates (Jones or Suarez) changed more? Why does that make sense for these data?

Definitions: One way to estimate a player's effect is to ignore all the other players, call this *no pooling*. Another way is to ignore the player to player differences and use the overall mean, call this *complete pooling*. Treating the player as a random effect creates *partial pooling*. We can think of each random effect estimate (predicted group mean) as being a weighted average of the group mean and the overall mean: $w(group\ mean) + (1-w)(overall\ mean)$ where the weight for group j (w_j), depends on the relative sizes of the variance components and on the group size, $w_j = \tau^2/\left(\tau^2 + \sigma^2/n_j\right)$. The weights reflect the "reliability" of the group.

(l) Calculate the weights for Jones and Suarez. Why are the weights pretty large? Which is larger? Why?

(m) Verify the estimated group means for Jones and Suarez using these weights. Which changes (from the group mean) more? Why?

To think about:

(n) Explain why some say “no pooling” overestimates the player to player variation.

(o) Summarize what you learn from the graph.