Generalized Linear Mixed Model (GLMM) Regressions

With GLMs, you can handle data distributions that are not Gaussian (normal). With GLM***M***s you also include random effects – factors you should account for, but that are not the planned, *a priori*, designed, and controlled drivers in your hypotheses. Last class we played with mixed-effect ANOVAs, with categorical, fixed-effect treatments. Here we continue GLMMs, but with a regression focus. And we address distributions. And we assume you remember past coding...

We work here again with the professor rating data set again – you already did multiple regressions with these data. Today we add in the focus on random effects and distributions.

1. Load and attach the professor grading data set. The data set is at:
   http://www.openintro.org/stat/data/evals.RData
   click on that link to download, and then open directly in to RStudio.
2. We pick up about where we left off last time. The hypothesis was that gender, age, ethnicity, seniority, and "beauty" affect student evaluations of teachers. Thus a simple linear model would have been something like:

```
lin.model <- lm(score ~ age + bty_avg + rank + gender + language)
summary(lin.model)
```

Remember that we listed quantitative covariates (age, bty_avg) before categorical factors, as we should ***in lm***. *How well did this simple model represent the "story"?* Let's try to improve that by analyzing better.

The variables above represent the teachers and are the planned, hypothesis-related factors of interest - i.e., the fixed effects - controlled by teacher inclusion and the main intent.

3. Now use the same (contents) of this model, but use `glm`, where you can also work with different underlying distribution families: `poisson, gamma`. And run `glm.nb` (in MASS) for a negative binomial distribution. No need to run a `glm` with `family=gaussian`; that is identical to a `lm`.

4. Run an `AICctab` to see which distribution is most plausible, but keeping that model the same.

Do we need to sweat other distributions, or is gaussian OK?

5. And what if you scramble the order of quantitative and categorical predictors in a `glm`? Does it matter anymore?

6. Let's also adjust for the percent of the class that completed an evaluation (cls_perc_eval). [Because students show up. Sometimes. Or not. *Ahem*.] Include cls_perc_eval as a covariate in the model and run again. Did this improve the fit?

Now we turn to other, unplanned, uncontrolled effects (i.e., random effects).

7. First we try a random intercept term, using `lmer` in the `lme4` package. This model assumes intercepts vary among random effects (but slopes do not). Include the term `+ (1|cls_level)` in the same equation as used above to organize results by class levels (i.e., lower [yr 1, 2] and upper [yr 3, 4]).

8. How much of the "story" was due to random effects vs. fixed effects? Hint: try r.squaredGLMM in MuMIn to find out. Also compare random effects Std. Dev. to fixed effect coefficients – *comparable values mean comparable effects*.

9. Now try a random slopes model – which assumes intercepts *and* slopes vary among the random effects. Include the term `+ (cls_perc_eval | cls_level)` in the model. This says that percent of class evaluating the teachers affects scores, and that those percent evaluations vary among class levels (e.g., senioritis).

10. How much more plausibly does this model work? Use AICctab to find out, and use random effects Std. Dev. and r.squaredGLMM again to evaluate random effects.

11. Keep building a better model: For example `bty-avg` might also depend on whether a photo was color or not (`pic_color`; because "beauty" was judged later by others, using teacher photos). Notice that each new random factor must be in separate (`_|_`).

Time for you to explore: As incentive, whomever obtains the model to predict teachers' scores with **the *lowest AICc and "good" residuals* wins 2% extra credit on the final!**