Generalized Linear Mixed Models (GLMMs)

With GLMs, you can handle data distributions that are not Gaussian (normal). With GLMMs you can also include random effects – factors you should account for, but that are not the planned-*a-priori*, designed, and controlled players in your hypotheses. Last class we played with mixed-effect ANOVAs, with categorical treatments (the fixed effects of interest). Here we continue to mess about with GLMMs, but we stop worrying about distinctions of quantitative and categorical predictors because *this is a subtle distinction when all output is regression-style*. And we address distributions.

We work here 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

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 model would be:

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

Notice that quantitative covariates (age, bty_avg) were listed before categorical factors, as we should list them **in lm**.

All these variables represent the teachers and are the planned, hypothesis-related factors of interest. These are the fixed effects - controlled by teacher inclusion and the main intent. How well does this model represent the variation among teachers for score?

3. Now use the same (contents) of this model, but use glm, where you can also work with different underlying distribution families: gaussian, poisson, gamma. Run an AICctab to see which distribution is most plausible, but where that model remains the same.

4. Is the lm model the same as the `glm gaussian` model? Should it be?

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

6. Repeat the `glm gaussian` model but use `glmmadmb` – do we get the same answer? If for some reason you get an error like this: "invalid type (closure) for variable 'rank'", it is because R thinks rank is numeric. Just make a new factor like this: `frank <- factor(rank)` and use `frank` as a variable instead.

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

Now turn your attention to other, unplanned, uncontrolled effects. For example, the study could not control the percent of the class that completed an evaluation (cls_perc_eval).
[Because students show up. Sometimes. Or not. Ahem. Back to analyses...]

8. Let's add cls_perc_eval as a random factor:

```
glmm.gau2 <- glmmadmb(score ~ age + bty_avg + gender + frank + language,
random = ~1|cls_perc_eval, family="gaussian")
```

9. What happened? A quantitative variable that may also explain patterns is a *covariate*. We list covariates in the main equation, like you did earlier. BUT! If that covariate depends on a category, then it also gets listed as depending on a categorical factor. Like this:

```
glmm.gau2 <- glmmadmb(score ~ age + bty_avg + gender + frank + language +
cls_perc_eval, random = ~ cls_perc_eval | cls_level, family="gaussian")
```

Where this says we expect the percent of a class to complete the survey to depend on the level of the class (upper division [yr 1, 2] or lower division [yr 3, 4]).

10. How much more plausibly does this model work? Use AICctab to find out.

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). So:

```
glmm.gau2 <- glmmadmb(score ~ age + bty_avg + gender + frank + language +
cls_perc_eval, random = ~ (cls_perc_eval | cls_level) + (bty_avg | pic_color),
family="gaussian")
```

Notice that >1 random factor must be in ( ).

Also, you should know this:

- A *random slope* like `(cls_perc_eval | cls_level)` says that the slope of the effect of `cls_perc_eval` on score also depends on `cls_level`.

- That differs from a *random intercept* effect of `cls_level`, which would be written as `(1 | cls_level)`, and says only the intercept of a regression depends on cls_level.

- You might even question if a fixed effect must remain thus. For example, if the intent of rank and age was *really* to get information on seniority, would the term be (age | rank) ? Or maybe rank alone could be an random intercept term: (1 | rank)?

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