Generalized Linear Models (GLMs)

With GLMs, you can handle data distributions that are not Gaussian (normal), instead of trying transformations. Yay! We still need to evaluate homogeneity of variance the same way you did for multiple regressions – by evaluating residuals. And we can evaluate assumptions of normality vs. other distributions using AICs, as long as we compare the same predictors.

Here we work with a professor rating data set, using multiple regressions in a GLMs framework.

1. Load and then attach a professor grading data set. The data set is at: https://www.openintro.org/stat/data/evals.csv

   More info on the data set is available at https://htmlpreview.github.io/?https://github.com/andrewpbray/oiLabs-base-R/blob/master/multiple_regression/multiple_regression.html

2. Student evaluation scores of professors are shown, along with information on the professors, including "beauty" scores (bty_), which are combined as a bty_avg. The following linear model tests the hypothesis that gender, age, ethnicity, seniority, and "beauty" affect student evaluations of teachers:

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

What can you interpret from this model?
How well does this model represent the variation in scores among teachers?

3. Now use the same basic model, but use **glm** (instead of lm), where you can also try different underlying distribution families: gaussian, poisson, Gamma, quasipoisson.

```
glm.model1 <- glm(score ~ scale(age) + scale(bty_avg) + rank + gender + language,
   family=gaussian)
summary(glm.model1)
plot(glm.model1)
```

   How does the model above differ from lin.model?

4. Now read about family in the Help window to see what other options exist for your modeling, and

5. Try a few other distribution families that might capture variation better, such as gamma. Note: binomial (and quasibinomial) apply to [0 / 1] responses – we will use binomial when we do Logistic Regressions. But all others may apply here.

6. Now try one more family – a *negative* binomial, which is *not* about 0/1 data, but is like a Poisson but more stretched/skewed. To do so requires that you first load the MASS package, and then you do everything as above except:

- instead of typing `glm`, you type `glm.nb`
- you omit the "`family = …`" phrase - because this is only for negative binomials.

7. Warnings pop up for non-integer data (e.g., `score`) when you try *Poisson* or *negative binomial* distributions because those *assume integer data*. People disagree on how severe this problem is.

8. So how would you select among distribution options? Three options occur to me:
   a) Look at residuals of your model. That works but is subjective.
   b) Run an AICctab using the bbmle package to see which distribution is most plausible. *If you keep predictors the same in all models you are then testing only which distribution most plausibly matches the data.* But this alone ignores the warnings.
   c) Use gaussian or *gamma* distributions for continuous data, and poisson or negative binomial for integer data.

4. For the model above, do we need to sweat other distributions, or is gaussian OK?

5. Now that you have found a "best" distribution to use, now try to find the most plausible model to predict teacher evaluation scores, using AICc.

6. Finally, should you use `glm` (or `glm.nb`) by default for regressions (instead of `lm`)?