**Data exploration and analysis workflow**

**Haldre Rogers**

**EEOB590A**

1. Start with a tidy dataset
2. Check structure of data to make sure in right class. Fix if necessary.
3. Subset data to the level at which you are going to do the analysis (if relevant)
4. Decide which variables are your response variables, and which are predictors.
5. Examine data: What type of variation occurs within my continuous variables?
   1. Outliers
      1. plot response and predictors to check for outliers (only with continuous data)
         1. Use histogram
         2. Use identify function in plot to id outliers
      2. If have outliers, decide what to do about them (keep or remove). If you keep, add a note to your script to keep track of this.
   2. Zero inflation Y (for count data)
      1. What proportion of response values = zero?
      2. If >25%, may have zero-inflated data. Make a note to keep an eye on overdispersion.
6. Examine data: What type of covariation occurs between my variables?
   1. Collinearity X: correlation between covariates
      1. Plot each predictor and random effect against each other
         1. As long as at least one is continuous, use ggplot (use facet\_grid for multiple categorical variables) or coplot
         2. Use summarize or table for two categorical (ftable for more than 2 categories)
      2. Check Variance Inflation Factor (function = vif())- should be less than 10 which is equal to a pairwise correlation = ca 0.95 (the cutoff VIF is debated, some say 3 or 5).
      3. Check correlation coefficient- should be less than 0.95. Note that correlation and collinearity are not the same: collinearity means linearly related, whereas data with varying amounts of linear relatedness can have the same correlation coefficient. Nevertheless, high absolute correlation coefficients usually indicated high linear relatedness.
      4. If variables are collinear, have choices
         1. drop one, based on biological knowledge. Dropping variables that are redundant is ok, when the choice is supported by subject-matter knowledge.
         2. Combine them (e.g. by averaging two positively correlated measures of the same concept, or by computing principal component scores) is better than dropping.
   2. Linearity and homogeneity - Look at relationships of Y vs X’s:
      1. Determine whether the relationship is linear or not
      2. Plot response against each predictor and random effect.
         1. Use plot or xyplot or ggplot or Myxyplot for two continuous variables
         2. Use boxplot or ggplot or coplot for continuous response and categorical predictor(s)
      3. Determine whether the variance at each X value is homogenous. This is the quick and dirty method- look at residuals below for more rigorous method.
   3. Independence Y
      1. Examine spatial/temporal aspects of sampling design
         1. Is there a pattern across time or space that is not incorporated in the model?
         2. Use Variograms or ACF (autocorrelation function) on the residuals (not the Y’s)
      2. Are there other potential factors that reduce independence of Y’s?
   4. Sufficient data?
      1. Is there sufficient data across all levels? Ideally the study design leads to adequate samples at each level, but ecology is messy.
         1. Use tables or ddply then ggplot to see if data is missing or insufficient for any level
         2. If not….
            1. Make sure adequate sample size for each level.
            2. Consider dropping levels if inadequate #’s, combining levels if it makes biological sense, or simulating to make sure it’s not influencing results.
      2. Examine interactions
         1. Is the quality of the data good enough to include them? (i.e. do we have enough samples for each level of the interaction?)
         2. Use ggplot to examine interactions
7. Fix up dataframe, if necessary
   1. Remove missing values (NA’s)
   2. Standardize continuous predictors
   3. Deal with any of the issues that you uncovered above
8. Run model
9. Model validation (use ggresidpanel package for lm, glm, glmer)
   1. Test for overdispersion (if poisson, binomial with proportions, or nb)
   2. Look at homogeneity: plot fitted values vs residuals
   3. Look at influential values: Cook
   4. Look at independence:
      1. Plot residuals vs each covariate in the model
      2. Plot residuals vs each covariate not in the model
      3. Common sense
   5. Look at normality: histogram
   6. If model validation checks out continue. If not, adjust model and run again
10. Model interpretation
    1. Graph fitted values
       1. Create grid for covariates
       2. Extract parameters
       3. Calculate fitted values on grid
       4. For confidence intervals for predicted values: Calculate the square root of the diagonal elements of of X \* cov(beta) \* X'
       5. Plot it all

References:

<http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x/abstract>

<https://dynamicecology.wordpress.com/2013/10/16/in-praise-of-exploratory-statistics/>

<https://dynamicecology.wordpress.com/2013/10/01/does-peer-review-increase-researcher-degrees-of-freedom/>

<https://smallpondscience.com/2013/06/04/pretending-you-planned-to-test-that-hypothesis-the-whole-time/>

<http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html>

Zuur 2009: http://onlinelibrary.wiley.com/enhanced/doi/10.1111/j.2041-210X.2009.00001.x/

Freedman’s paradox: <http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1983.10482729>