Putting it all together – what to do when

The following key will help you decide what type of response variable(s) and error distribution(s) you have.

1. What is the smallest unit of observation?
2. What variables were measured on this observational unit?
3. Which variable(s) is (are) the response variable(s)?
4. How many response variables are there?
    a. One (1) – go to 5
    b. > One (>1) – Multivariate Analysis, get professional help
5. Is the relationship between the link function of the expected value of the response and the covariate(s) linear? (Plot the response variable or suitable transformations of the response variable against each of the covariates.)
    a. No, use an additive model to explore non-linear responses; replace lm() and glm() with gam().
    b. Yes, use linear or generalized linear models.
6. Is the smallest unit of observation *nested* inside a larger unit (e.g. plots within fields)?
    a. Yes, consider a mixed model, replace lm() with lmer() and glm() with glmer(), gam() with gamm()
    b. No, continue with models with single error distributions.
7. Is the response variable continuous (a real number) or discrete (an integer)?
    a. Continuous, go to 8
    b. Discrete, go to 11
8. Is the continuous response *bounded* on the bottom (e.g. at zero).
    a. No, use a model with a normal error distribution (e.g. lm())
    b. Yes, go to 9
9. Does the response variable sometimes take a value of exactly zero?
    a. No, consider log transforming the response prior to using a model with a normal error distribution (e.g. lm()), or use a gamma error distribution (e.g. glm(…,family=gamma).
    b. Yes. Go to 10.
10. Choose from the following two options:
    a. Add a small (≤1) constant to all values of the response, log transform and use a model with a normal error distribution.
    b. Split the analysis into a binomial presence/absence model, and a normal error model of the log transformed observations > 0.
11. Is the discrete response bounded on the bottom (e.g. are negative values possible)?
    a. No, consider a model with a normal error distribution, but check carefully for heteroscedasticity.
    b. Yes, go to 12.

12. Does the discrete response have a maximum value (upper bound)? The value may differ for each observation.
    a. No, consider the response to be Poisson and go to 15.
    b. Yes, go to 13.
13. How many different discrete outcomes are possible for each response?
    a. Two (yes/no, present/absent, true/false). Consider the response to be binomial (*n* successes in *m* trials; use glm(…,family=binomial)). Check for overdispersion if *m*>1.
    b. More than two. Go to 14.
14. Are all the covariates categorical?
    a. Yes, use contingency tables and related techniques (get expert help)
    b. No, consider multinomial regression (get expert help) or convert to numeric scores and treat as normal (get expert help here too – psychometrics does this alot).
15. Does each observational unit represent the same amount of time or space?
    a. No, use an offset in the formula to account for variation in sampling effort between observations, and proceed with using glm(. ~ . + offset(log(effort.variable)), …,family=poisson). Be sure to check for overdispersion.
    b. Yes, proceed with using glm(…,family=poisson). Be sure to check for overdispersion.

2012-09-26

Model Selection According to Drew

The following series of steps summarize my approach to model selection. This works for any model or set of models fitted using lm() and derivatives.

1. Is your study
    a. a *manipulative, controlled experiment?* Go to 2.
    b. an observational study or mensurative experiment, or a manipulated experiment with many additional observed covariates? Go to 3.
2. Fit a model with all experimental variables included, with at least all pairwise interactions; consider including higher order interactions if they are biologically sensible hypotheses. If necessary check for over- or under-dispersion. Check the residuals and plots of the fitted values vs. the observations to detect systematic lack of fit. If necessary, add additional variables, polynomials or non-linear functions of variables. Use backwards selection with F-tests or Likelihood Ratio tests as appropriate to simplify the model as far as possible.
3. Now the fun begins. Write down a list of all the variables that you think affect the response variable, along with the reasons WHY they affect your response variable.
4. Construct a set of models that will allow you to tease apart any distinct hypotheses you have. Bound this set with a null model (g(E(Y))~1) and a 'global' model that includes all the variables and interactions that you want to consider. Consider including univariate models and combinations of hypotheses. Fewer models are better, but a large set can be useful for developing new hypotheses although it will not be confirmatory.
5. Fit the global model. If necessary check for over- or under-dispersion. Check the residuals and plots of the fitted values vs. the observations to detect systematic lack of fit. If necessary, return to step 3 to add additional variables, polynomials or non-linear functions of variables.
6. Ensure that all of the coefficients are estimable (are there NA's in the table of coefficients?). If not, return to 4 and adjust the set of models to eliminate problematic interactions.
7. Fit your models using the appropriate function.
8. Construct an AIC_c table, or $QAIC_c$ if you detected overdispersion in step 5.
9. Interpret your table. Look for evidence of nested pretending variables (similar Log-likelihood values). Refer back to the list of hypotheses posited in (3).
10. Make plots of the response against the variable with the strongest effect. Add predictions and confidence intervals from the top model, single most parsimonious model (in case of egregious nesting), or model averaged predictions.
11. If > 1 model with weight > 0.1, and they are not nested, consider constructing an exploratory model to examine new combinations of hypotheses including interactions. Do not add this model to the original set. Draw inference from this model independently of the model set, and be clear it was constructed using the results from the model set. Interesting plots can often be made from predictions of this model, but be clear that these are hypotheses to be tested with new data.