Model selection and inference

Key references

Harrell, Venables, Schielzeth 2010

```
require("reshape2")
library("ggplot2")
theme_set(theme_bw())
```

Model selection

- discrete set of possible models
- want to rank/evaluate relative quality (using *some* metric); need to adjust for model complexity somehow
  - adjusted R^2
  - cross-validation score
  - AIC: $-2L + 2k$
    - minimize expected Kullback-Leibler distance (average predictive error)
    - under some cases $\rightarrow$ CV score
    - inconsistent ...
    - ... but makes sense in the case of *tapering effects*
  - Bayes/Schwarz IC:
    - asymptotically equivalent to the *Bayes factor*, odds in favor of a specified model
    - penalty term is $(\log n)k$ (so, stricter than AIC when $n > 8$)
    - focus on *dimensionality* (true number of parameters of the model)
  - finite-size correction:
    - AICc (derived for linear models)
    - mixed models? *conditional AIC*, Vaida and Blanchard
  - Deviance information criterion: makes sense in principle (allows for priors, etc.) but tricky: "level of focus" problem

Model reduction

Why?

- Over- vs under-fitting (aka *bias-variance tradeoff*)

- Everything is interesting, but we don't have enough data to estimate everything reliably ...
- ... and trying to estimate everything messes up *all* of our estimates
- **REFRAIN FROM DATA-DRIVEN MODEL REDUCTION**, including stepwise and all-subsets approaches
- rule of thumb, 10-20 data points per parameter
- Harrell, p. 61: **type of response variable** | **limiting sample size** continuous | n binary | min(success,failure) ordinal (k categories) | $n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ survival time | number of failures
- (Harrell 2001) gives guidelines for different kinds of reduction:
  - 'domain knowledge'/common sense
  - PCA and related techniques (perhaps by groups?): clustering
  - penalized regression
- eschew the "minimal adequate model" approach (OK for getting rid of interactions, random effects???)
- same issues apply with collinearity: (Graham 2003)
- AIC-based 'multi-model averaging' is a form of penalized modeling, but probably

Correlated variables

- Correlation interferes with *interpretation*, not *fitting*
- ... except for fancy fitting methods (e.g. BUGS)
- ... and except for perfect correlation (unidentifiability) – missing factor combinations
- Centring takes care of correlation between continuous predictors and the *intercept*
- Test the simultaneous effect of *all* correlated predictors: e.g.

```
model1 <- lm(y~.,data=mydata)
model2 <- update(model1,.~.-corrvar1-corrvar2-corrvar3)
anova(model1,model2)
```

- If you must drop some correlated predictors:
  - Make an *a priori* decision (but know that your results are entirely conditional on it!)
  - use PCA or some other summarization tool

  Summarizing collinear variables using a summary score is more powerful and stable than arbitrary selection of one variable in a group of collinear variables (Harrell p. 65) # Model selection

OK, let's suppose you do have a good reason to do model selection (rather than model *testing*, which corresponds to hypothesis testing) * Model selection (comparison among choices) vs. model validation (unspecificed alternative) * Lots of good arguments vs stepwise/all-subsets approaches (Harrell, Whittington), although see Murtaugh 2009 * Algorithmic approaches: cross-validation, again (but again see)

Model parameterization/interactions

How you parameterize a model is important, because it frames how you ask your questions, sometimes in surprising ways.

Centering and scaling (Schielzeth 2010)

- *Centering* continuous input variables is a sensible way to make sure that interactions are interpretable, and that the estimate of the intercept is independent of the other estimates
- Centering to the mean is a reasonable default, but you might also consider, for interpretability, centering to some meaningful default value (e.g. if the mean temperature is 20.1, it might be better to set the baseline at 20)
- If everyone centers to the mean of their data set, it makes it harder to compare results (although not impossible, as long as the mean is actually reported)
- *Scaling* input variables (e.g. by 1 SD) doesn't change the statistical properties of the estimate at all, but allows meaningful comparison of the magnitude of the estimates
- May improve estimation slightly for complex fitting problems
- Same issues about choosing the scale, and across-study comparisons, apply
- `scale()` function in R

Marginality etc.

- Need to be very careful interpreting main effects in the presence of interactions: don't, or be very careful to set contrasts properly
- Type I, II, III ...
- be very careful of what `anova` does! may want to use `car::Anova`

Catch-22

Ideally, to test effects we would like to proceed with modeling in a way that (is):

- is interpretable

- incorporates all reasonable interactions
- is reasonably parsimonious
- is not based on variable selection (no snooping)
- respects marginality

???

Graham, Michael H. 2003. "Confronting multicollinearity in ecological multiple regression." *Ecology* 84 (11): 2809–2815. doi:10.1890/02-3114. http://www.esajournals.org/doi/abs/10.1890/02-3114.

Harrell, Frank. 2001. *Regression Modeling Strategies*. Springer.