Data analysis practice

Ben Bolker

23:57 01 July 2015

Principles

- goals? confirm, predict (decide), explore ...
- avoid snooping: decide on (at least primary) analyses first – self-registration
- biology first!
- workflow:
  - self-register
  - data viz
  - fit models/run tests
  - diagnostics
  - data viz+predictions

Considerations

- problem size
  - observations
  - predictors/parameters
  - clusters/individuals
- number of complications
  - missing data (imputation); mechanistic models; observation vs process error; phylogenetic/pedigree structure; zero-inflation; big data; large number of predictors ($p \gg n$); compositional data; ordinal data; multivariate/multitype responses; circular data; spatial/temporal corr.; causal networks ...
- choose how to spend data, human, computational effort

*Choosing a model (*a priori*)*

- Harrell (2001)
- 'spending' parameters on effects
  - 1 parameter per linear term
  - $n - 1$ parameters per categorical variable
  - spline terms (flexible)
  - interactions!
- confirmatory studies: *one* parameter per 10-20 data points
- dimension reduction:

  - *a priori* choice
  - PCA (possibly by group)
  - informative Bayesian priors

post hoc *model choice (don't do it!)*

- stepwise
- minimal adequate model
  - everyone (except Harrell) does it to some extent: Bates, Venables
- OK for prediction (Murtaugh 2009)

variable importance

- scaled parameter estimates (Schielzeth 2010)
- summed AIC weights: problematic! Cade (2015):

  > The associated sums of AIC model weights recommended to assess relative importance of individual predictors are a measure of relative importance of models, with little information about contributions by individual predictors.

- random forest approaches
- variance explained (tricky)
- what does variable importance mean anyway?

Interactions

- if non-sig: remove? not sure ...
- if sig:
  - interpret main effects **carefully**
  - subgroup analysis (avoid comparing significance across groups)

Goodness of fit

- maximal model should fit "adequately"
- important but hard
- R^2 has many definitions: GLMs (UCLA stats site, nonlinearity, multilevel models (Nakagawa and Schielzeth 2013)
- predictions/prediction CI/visualization
- cross-validation

Model averaging

- sensible for *prediction*
- need to average terms (predictions) that are *consistent across models*

- not for hypothesis testing
- shrinkage/penalized estimation
- be careful of collinearity; interactions; nonlinearity
- consider explicit penalized regression (lasso, ridge regression) `rms` (`ols`, `lrm`), `glmnet` (Hastie, Tibshirani, and Friedman 2009)

Multiple comparisons

- Holm better than Bonferroni (default of `p.adjust`)
- FDR (Benjamini-Hochberg)
- all-pairwise-comparisons (`multcomp`): OK, but why bother?
- possibly needed for a large set of predictors

References

Cade, Brian S. 2015. "Model Averaging and Muddled Multimodel Inference." *Ecology*, March. doi:10.1890/14-1639.1.

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

Hastie, Trevor, Robert Tibshirani, and J. H Friedman. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York: Springer. `http://public.eblib.com/EBLPublic/PublicView.do?ptiID=437866`.

Murtaugh, Paul A. 2009. "Performance of Several Variable-Selection Methods Applied to Real Ecological Data." *Ecology Letters* 12 (10): 1061–68. doi:10.1111/j.1461-0248.2009.01361.x.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models." *Methods in Ecology and Evolution* 4 (2): 133–42. doi:10.1111/j.2041-210x.2012.00261.x.

Schielzeth, Holger. 2010. "Simple Means to Improve the Interpretability of Regression Coefficients." *Methods in Ecology and Evolution* 1: 103–13. doi:10.1111/j.2041-210X.2010.00012.x.