Analysis: the big picture

2013-07-01 15:49:13

What are we trying to do when we analyze (model) data?

- "learn from the data"; "answer scientific and management questions" (vague!)
- describe (are we ever *really* interested in pure description?)
- understand or explain something (slippery)
- test one or more hypotheses (why?)
- predict (this at least is fairly clear)
- make a decision (ditto)

[1][2]

In a slightly narrower framework where we can isolate particular effects that we are interested in . . . do we want to know

- if they are different from zero?
- whether they are positive or negative?
- what their magnitude (*point estimate*) is?
- how uncertain they are, or what their *interval estimate* is? (How do these more specific questions fit in with the bigger questions listed above?)

In this context Gelman (2000) talks about "type S" errors (we estimated the wrong sign) and "type M" errors (we estimated the wrong magnitude), in contrast to the classical "type I/type II" (we incorrectly rejected/failed to reject the null hypothesis).

All of the technical methods we often think about (model selection, parameter estimation, p-value estimation, statistical inference) . . . are all *means, not ends*.

The statistical Message Box

Can we develop an analogue of the Message Box for connecting scientific questions with analytical methods??

- *What is our scientific question?*

[1] Harrell has a nice (somewhat more concrete) discussion of analytical goals in the context of model reduction (Harrell 2001)

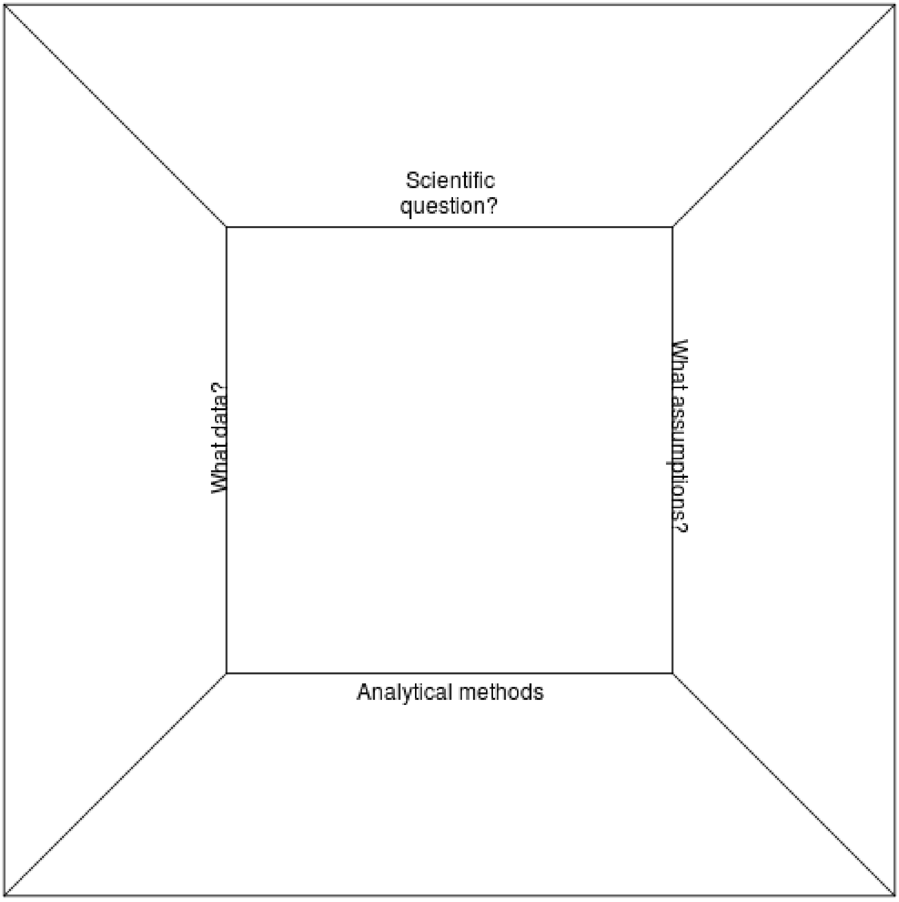
[2] The relative clarity of prediction and decision-making helps to explain scientific philosophies such as logical positivism (compare with the real-life messiness of Lakatos and Feyerabend . . .).

- The goal/reason we are doing the analysis in the first place
    - How concretely, precisely, and tersely can we state it?

- *What data (observables) are we using?*
    - the input to the analysis
    - What are we assuming about the relationship between our observables and the true (!!!) quantities and relationships of interest?

- *What quantities or relationships are we estimating?*
    - The (proximal) output of the analysis
    - Why? How do these quantities or relationships relate back to the scientific question?

- *What analytical methods are we using?*
    - What assumptions do they embody about the form of the data and/or the causal relationships among entities?



Limiting resources for analysis

- data
- signal
- brainpower (failure to simplify the question appropriately)
- analytical tools
- computational power
- conceptual framework (vague *or* misguided)
- independence (political pressure, academic or real-world; avoiding statistical machismo)
- conformity (unwillingness to use prior art)

Discussion

In groups:

- what is your big-picture question?
- what does your message box look like?
- what is your most limiting resource?

Paradigms

Associated with goals & techniques, but not defined by them)

algorithmic	**model-based**
descriptive/phenomenological	mechanistic
predictive	explanatory
exploratory	confirmatory
large data	small/medium data
robust	efficient
nonparametric	parametric
observational data	experimental data
computational	analytical
field	lab

Ripley (2004: `fortunes::fortune("machine learning")`)

> To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'

Breiman (2001)

> as data becomes more complex, the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature's mechanism.

Paradigm conflict across fields

- Platonists vs (?) Aristotelians; e.g. constructive empiricists
- Biology/ecology: Strong inference (Platt 1964); Peters (1991) *Critique for Ecology*
- linguistics: Norvig vs Chomsky
- arguments about microfoundations in economics; Big Data in econometrics
- Chris Anderson: "The End of Theory"

Methods

Models are *always* simplified versions of reality: otherwise they they don't help us understand, or predict, reality (Borges)

- constancy
- linearity
- independence
- smoothness
- discrete classes

Classical

- Linear models: mostly "classical", but:
  - least-squares/MVUE interpretation
  - very efficient for Big Data (large-scale linear algebra)
- extended linear models: GLMs, correlations, zero-inflation, etc.
  - more/different parametric assumptions in pursuit of efficiency & interpretability
- hierarchical/mixed models
  - ancestor (ANOVA) mostly used for hypothesis testing
  - relatively efficient way to do grouping

    - works well for large N, small n within clusters
    - computationally challenging
- classical (rank-based) nonparametrics [weak assumptions about conditional distributions]: mostly hypothesis-testing (provide *only* p-values)

Algorithmic

- modern nonparametrics
    - generalized additive models (technically still 'linear models', with attendant advantages)
    - kernel density estimators (*smoothing*)
    - quantile regression
    - great for description, but difficult for decomposing descriptions (interpretability)
    - interactions possible (tensor product splines, multidimensional KDEs) but comp. intensive
- classification and regression trees (plus extensions: random forests/bagging/boosting etc.)
    - mostly ignore interactions
- support vector machines
    - computationally powerful high-dimensional categorization
- penalized/regularized approaches (ridge regression, lasso, . . .)
    - mostly description-oriented; confidence intervals etc still difficult

Model building

Many tradeoffs (Levins 1966):

- Realism
- Computational feasibility (especially if resampling)
- Conformity with existing models
- Interpretability
- Flexibility

etc. etc. etc. . . .

Deciding on a model?

- no free lunch
- bias-variance tradeoff = under/overfitting
- **BE VERY, VERY CAREFUL WHEN USING THE DATA TO DECIDE ON A MODEL**, especially if doing hypothesis testing (*data snooping*)
- in- vs out-of-sample prediction
    - bad in-sample prediction → bad model
    - good in-sample prediction: maybe overfitted?

Model checking and diagnostics

- Graphical tools
- Goodness-of-fit measures (*avoid hypothesis testing!*)
    - Compare to saturated and null model
- Explore residuals
- Posterior predictive sampling
- Assessment of predictive skill:
    - hold-out data
    - cross-validation: this document points to `boot::cv.glm`; `rms::validate.*` (*But* see Wenger and Olden (2012))
- Fit to simulated data
    - Simulated from estimation model (= positive/negative controls)
    - Simulated from a different model (robustness)

References

Gelman, Andrew, and Francis Tuerlinckx. 2000. "Type S error rates for classical and Bayesian single and multiple comparison procedures." *Computational Statistics* 15: 373–390.

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

Levins, R. 1966. "The Strategy of Model Building in Population Biology." *American Scientist* 54: 421–431.

Peters, R. H. 1991. *A Critique for Ecology.* Cambridge, UK: Cambridge University Press.

Platt, John R. 1964. "Strong Inference." *Science* 146 (oct): 347–353. http://links.jstor.org/sici?sici=0036-8075%2819641016%293%3A146%3A3642%3C347%3ASI%3E2.0.CO%3B2-K.

Wenger, Seth J., and Julian D. Olden. 2012. “Assessing transferability of ecological models: an underappreciated aspect of statistical validation.” *Methods in Ecology and Evolution* 3 (2) (apr): 260–267. doi:10.1111/j.2041-210X.2011.00170.x. http://doi.wiley.com/10.1111/j.2041-210X.2011.00170.x.