Fitting Statistical Models in Julia

Douglas Bates

06/26/2014

Current state of play in Statistical Computing

- Some older style systems (SAS, SPSS) are still used but primarily because of inertia.
- R has converted the discipline to Open Source, and to function-based computing in a REPL.
- The CRAN archive has democratized contributions to statistical software
  - theses, etc. frequently include production of an R package
  - good news: reference implementations of methods are available
  - bad news: some of these implementations are not examples of best practices
- Design decisions of R (and, before it, S) are exacting a toll.
  - S was originally an interface language. Good R performance often requires compiled code.
  - Functional semantics are a mixed blessing.

Popularity of R is remarkable

- Estimated to have many millions of users
- Probably over 10,000 packages in various archives by now.
- Coursera MOOCs by Roger Peng and others at Johns Hopkins have themselves attracted more than a million enrollees.
- Used in many commercial settings.
- Support from RStudio (IDE, reproducible research). Commercialization at Revolution.
- Artfully used it can lessen the pain of an intro stats course.
- Hundreds of books on " with R", web resources, etc.
- Some elements of its design are very effective
  - Dataframes, Missing Data, Factors (JMW, next up)
  - Formula language

Is Julia the "R of the future"?

- I am comfortable recommending to those using compute-intensive methods or working with large data sets that they learn Julia
- Julia provides much greater flexibilty, speed and power than R does or can, without radical changes.
- Presently Julia is suitable for early adopters and geeks.
- Most statistical analysis is done by researchers, not statisticians, who are not ready for Julia.
- Will or should they be ready for Julia and Julia ready for them?
- Perhaps as Rick said in Casablanca - "Maybe not today, maybe not tomorrow, but some day and for the rest of your life."

Model representation as a formula

A taxonomy of basic statistical models

- $\mathbf{X}$ a model matrix of size $m \times n$ (statisticians write this as $n \times p$; number of observations by number of parameters). Typically $m > n$.
- $\mathbf{y}$, the m-dimensional vector of responses, the realization of the random variable $\mathcal{Y}$.
- $\eta = \mathbf{X}\beta$, the m-dimensional *linear predictor* depending on the n-dimensional *coefficient vector*, β.
- For a *linear model* we assume

$$\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I_m})$$

  That is, $\mathbb{E}(\mathcal{Y}) = \mu = \eta$.
- A *generalized linear model* (GLM) incorporates a component-wise mapping $\mathbf{g}$ called the *link* function such that $\mathbf{X}\beta = \eta = \mathbf{g}(\mu)$

"distinct" techniques are often special cases

- Linear models encompass
  - simple linear regression
  - multiple linear regression
  - polynomial regression
  - analysis of variance
  - t-tests
- GLMs encompass
  - logistic regression (Bernoulli distribution and logit link)
  - Poisson regression (Poisson distribution and log link)

Analysis and interpretation requires more than $\hat{\beta}$

- There is more to fitting such models than just evaluating $\hat{\beta}$.
- Model fit diagnostics, measures of precision (confidence intervals) and/or hypothesis tests are part of the analysis.
- Often groups of coefficients and groups of columns in $\mathbf{X}$ have important interpretations.
- The formula language is a high-level description of a model from which the model matrix, $\mathbf{X}$, is derived
- Post-fit analyses: analysis of variance, analysis of deviance, various other hypothesis tests, ... use information from the terms in the formula.

Terms in a formula

- We will write a generic response as `y`, categorical covariates as `f` and `g`, and continuous covariates as `u` and `v`. An intercept column (column of 1's in $\mathbf{X}$) is implicit.
- Some examples

```
y ~ u         # simple linear regression, X is m by 2
y ~ 1 + u     # simple linear reg. w/ explicit intercept
y ~ 0 + u     # reg. through origin, suppressed intercept
y ~ 1 + u + v # multiple linear regression
y ~ f         # one-way analysis of variance
y ~ f + g     # two-way analysis of variance
y ~ f + g + f&g # two-way anova with interaction
y ~ f * g     # expands to y ~ f + g + f&g
```

The "there is only one formula" phenomenon

- Statistics is often taught as rote application of formulas from some text.
- The concept that there is a model behind the formula is often never mentioned.
- This leads to the conviction that there is only one possible way of evaluating the result.
- It is well-known that the only possible way to evaluate regression coefficients is

$$\widehat{\beta} = (X'X)^{-1} X'y$$