Advanced Modeling in R

Non-linear, Bayesian, and mixed effect methods

R. Condit*
Smithsonian Tropical Research Institute, 7-9 May 2012

1 General organization

The course will cover several advanced statistical modeling methods using the programming language R, including maximum-likelihood, non-linear, Bayesian, and multi-level (hierarchical) methods as well as techniques for using data simulation to test models. The R function *lmer*, an accessible yet complex tool for advanced modeling, will be covered in detail. To establish a base for understanding multi-level models, some review of standard regression will be included, plus a session on fitting non-linear models with maximum likelihood.

During the first half of each session, I will explain methods and present examples of their use; in the second half, students will work on assignments using the same methods. Datasets will be provided, but students are encouraged to bring their own data as well. A course web site will provide sample code, data, and a list of key R functions. Students should be familiar with R: manipulating dataframes, graphing, and linear regression.

1.1 Applying

- Apply: Contact Liliana Londoño, Center for Paleobiology, STRI

1.2 Schedule

- When: Three sessions, 8:30-4:30, 7-9 May 2012
- Where: Tupper Training Room (Next to Small Meeting Room, below cafeteria)

2 Software requirements

I assume you will have laptops running R, that you know how to manipulate dataframes in R, and have some experience with graphing and simple summary statistics. I suspect you have

*CTFS & SIGEO

already used the functions *lm* and (perhaps) *glm*, but in case you haven't, you will quickly learn them. The course will begin with those functions as a baseline for moving off into more advanced methods for fitting models. Please have the packages listed below installed and running beforehand, and I encourage you to get programming editor already installed before we start.

- R base package
- R contributed packages *lme4*, *arm*, *coda*, *mvtnorm*, *date*, available at `http://cran.r-project.org/`
- RStudio, or a programming editor such as Geany or equivalent (Notepad++; NOT Wordpad nor Notepad)
- CTFSRPackage from `http://ctfs.arnarb.harvard.edu/Public/CTFSRPackage`

3 Course web site

- `http://ctfs.arnarb.harvard.edu/Public/Workshops/Tupper2012`

  outline.pdf and `http://ctfs.arnarb.harvard.edu/Public/Workshops/Tupper2012/outline.html`
  assignments.pdf and `http://ctfs.arnarb.harvard.edu/Public/Workshops/Tupper2012/assignments.html`
  sample R datasets `http://ctfs.arnarb.harvard.edu/Public/Workshops/Tupper2012/data`
  R scripts `http://ctfs.arnarb.harvard.edu/Public/Workshops/Tupper2012/source`
  history of commands I enter `http://ctfs.arnarb.harvard.edu/Public/Workshops/Tupper2012/history`
- All will be updated regularly

4 Sources

- Bates' online book `http://lme4.r-forge.r-project.org/`
- Random vs. fixed effects `http://andrewgelman.com/2005/01/why_i_dont_use/`
- Gelman text (Amazon: `http://www.amazon.com/Analysis-Regression-Multilevel-Hierarch`
  `dp/052168689X`)
- Kruschke: `http://www.amazon.com/Doing-Bayesian-Data-Analysis-Tutorial/dp/0123814855/ref=pd_sim_b_2`
- Carlin: `http://www.amazon.com/Bayesian-Methods-Analysis-Edition-Statistical/dp/1584886978/ref=pd_sim_b_4`
- Albert: `http://www.amazon.com/Bayesian-Computation-R-Use/dp/0387922970/ref=pd_sim_b_5`
- Robert: `http://www.amazon.com/Introducing-Monte-Carlo-Methods-Use/dp/1441915753/ref=pd_sim_b_2`

5 Contents

- Modeling with standard regression and maximum likelihood [morning 1]
  1. Linear regression with lm (review)
     - Gaussian error
     - Residuals and statistics (coef, summary)
     - Data treemass: log(agb) vs. log(dbh)
     - Centering x in linear regression!
       Use xCenter = x - mean(x)
  2. Numerical estimation with optim
     - maximize likelihood vs. minimize sum of squares
     - alternate methods in optim (Nelder-Mead etc.)
     - comparing models with AIC
     - Non-linear models
- Bayesian methods [afternoon 1, day 2]
  1. Bayes rule and the posterior distribution
  2. Metropolis, the Gibbs sampler (MCMC)
     a) Another method for fitting parameters
     b) Automatically provides fully accurate confidence
     c) Much more flexible modeling options (ie, non-linear with many parameters)
     d) Any error distribution
     e) Latent states or latent data
  3. Hierarchical modeling
  4. Limitations: long run time, complicated program
  5. Keys to your own program
     a) Getting the correct likelihood functions, and this can be difficult in complex models
     b) Preparing data structures to save all the data and likelihood
     c) Looping through all the parameters and hyperparameters
     d) Returning results
  6. Details
     a) Parameter correlation, autocorrelation and poor convergence
     b) Diagnostics (see coda package)
     c) Fitting the covariance
     d) Special cases where Metropolis not needed
- Data simulation [not covered]
  1. Two purposes of simulation
     - Understand connection from Process –> Data
     - Test whether models work
  2. R's probability distribution functions
  3. Regression with error

4. Multi-level regression
5. Extra: Survival

- Multi-level models (mixed effect, hierarchical, random vs. fixed effects) [day 3]
  1. Why multi-level modeling?
  2. Limitation: linear (or transformed linear) with normal error
  3. Multi-level vs. standard regression
     Bates Chap 4, Section 4.4; Gelman & Hill pp. 251-259
  4. Regression with one group using lmer
     - output of display
     - graphs using the coefficients
     - variable intercept, slope, or both
  5. Regression with two groups or two predictors x using lmer
     - output of display
     - models with or without covariance
     - group level predictor (see Gelman&Hill p. 265)
     - graphs using the coefficients
  6. Random for fixed?
     - Traditional
       * Random: nuisance effects, unrepeatable (batch, plot)
       * Fixed: permanent group, repeatable (sex
       * Gray area: year? site?
     - Recent issues favoring multi-level approach
       (ie, Gelman, who replaces 'random' with 'grouping')
       * Is group-level variation an explicit research topic?
       * Can different groups be thought of as similar?
       * Can information on one group support other groups?
       * Are some groups rare and thus needing support?
       * Are there enough groups? (too few -> little evidence on group-level variation)

6 Error functions

- dnorm is the standard
- dbinom is the standard for survival or occurrence (or similar)
- dlnorm
  - for abundances, whether integer or not (but usually not used in favor of log-transformation
  - good match for tree growth rates
  - but cannot handle zeroes
- dgamma is similar to log=normal
- dpois including zeroes (but does not handle much ecological data well)
  - for integer abundances
  - handles zeroes
  - however, close to Gaussian so not appropriate for much ecological data
- dnbinom
  - for integer abundances that are highly skewed
  - very common in ecology
  - R: prob=dnbinom(count,size=k,mu=mu)
  - size is 'clumping parameter'; mu is mean

7 R functions

- Data extraction
  1. subset
  2. apply
  3. tapply
  4. cut
  5. dim
  6. str
  7. names
  8. ifelse [R base package]
  9. IfElse [CTFSRPackage version]
- Graphics
  1. hist
  2. plot
  3. points
  4. line
  5. curve
  6. abline
  7. box

8. axis
9. X11
10. dev.set

- Modeling
  1. summary
     - mean
     - median
     - sd
     - var
     - cor
     - CI [CTFSRPackage]
  2. model
     - lm
     - glm
     - lmer [lme4 package]
     - coef
     - summary
     - fixef [arm package]
     - ranef [arm package]
     - display [arm package]
     - dotplot [lattice package]
     - xyplot [lattice package]
- Probability distributions
  1. PDFs
     - dnorm, rnorm, pnorm, qnorm
     - dbinom, rbinom, pbinom, qbinom
     - dlnorm etc.
     - dnbinom etc.
- Likelihood
  1. optimize
  2. optim
  3. metrop1step [in CTFSRPackage]