

Topics in Applied Regression

J.H. Maindonald*

January 8, 2007

‘A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.’ M.J.Moroney

The following topics will be treated at various levels of detail, depending on time available.

- Review the theory of linear models, emphasizing the use of regression splines, and noting (but not discussing in detail) the further extension into generalized additive models.
[c.f., Ch. 1 of Wood (2006)]
- Introduce the theory of Generalized Linear Models, with logistic regression and Poisson regression as special cases.
[Maindonald & Braun (2007, Sections 8.1 – 8.5); Wood (2006, Sections 2.1 – 2.3)]
- Provide an introductory account of supervised learning (classification). Several alternative methodologies will be discussed and illustrated; probably tree-based methods, Breiman’s random forests, and linear discriminant analysis as implemented in R’s function `lda()`. Note that this account will be introductory, with very limited discussion of the theory. These methods are covered because they will be used to provide constructed variables for use in regression on constructed variables – known in this context as propensities.
[Maindonald & Braun (2007, Ch. 11, Section 12.2)]
- Note issues in the use of linear and generalized linear models, and of other regression models. This will be, largely, an exercise in awareness raising.

*Centre for Mathematics & Its Applications, Australian National University, Canberra ACT 0200, AUSTRALIA. <mailto:john.maindonald@anu.edu.au>

- Preliminary data exploration; [Maindonald & Braun (2007, ;Chapters 2 & 6)]
 - Transformation of explanatory variables, where possible, to ensure linearly related predictors; [Maindonald & Braun (2007, Ch. 6); Cook and Weisberg (1999)]
 - Diagnostic plots; uses of simulation; [Maindonald & Braun (2007, Ch. 6)]
 - Missing variables; noting striking examples that arise in multi-way tables, perhaps modeled using logistic regression; [Maindonald & Braun (2007, Subsections 2.2.1, 3.4.5, 6.8.3 & Section 8.3)]
 - Observational versus experimental data – implications for interpretation and inference; Maindonald & Braun (2007, Ch. 6); Rosenbaum (2002)]
 - Variable selection, noting the use of resampling methods to obtain realistic “error” estimates; [Maindonald & Braun (2007, Ch. 6)]
 - Errors in explanatory variables; implications of classical measurement error for inference; [Maindonald & Braun (2007, Ch. 6); Carroll (2006, Ch. 1)]
 - Regression on constructed variables – principal components, partial least squares, and propensity scores; [Maindonald & Braun (2007, Ch. 13)]
- Worked examples, some using data that have been a basis for published research, will be used as a basis for a more extended discussion of selected issues from the list given above;
 - Laboratory exercises, 2 hours per week, using the R system for the computations. These will adapt and extend relevant material from the set of laboratory exercises used for a 2006 Mathematics Department course at Australian National University.
[<http://www.maths.anu.edu.au/~johnm/courses/dm/statminers> see especially sub-directory **labs**]

1 References and Further Reading

References

Carroll, R., Ruppert, D. and Stefanski, L. A. 2006. *Measurement Error in NonLinear Models*. 2nd edition, Chapman and Hall.

[This is the definitive text on errors in linear and nonlinear models.]

Cook, R. D. and Weisberg, S., 1999. *Applied Regression Including Computing and Graphics*. Wiley.

[This emphasizes geometric insights, linear predictors (transformation of predictors, if possible so that relationships between them are linear), and dimension reduction.]

Faraway, J.J. 2006. *Extending the Linear Model with R*. Chapman & Hall/CRC.

Faraway, J.J. 2005. *Linear Models with R*. Chapman & Hall/CRC.

Maindonald, J.H. and Braun, W.J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach*. 2nd edition, Cambridge University Press.

[URL: http://wwwmaths.anu.edu.au/~johnm/r-book.html](http://wwwmaths.anu.edu.au/~johnm/r-book.html)

[As the title says, this is example-based, drawing attention to theoretical as they arise in the context of specific examples. There is extensive use of graphs that may provide insight on data and on fitted models. There is detailed advice on practical data analysis issues, with careful critiques of practical data analysis issues.]

Rosenbaum, P. R., 2002. *Observational Studies*. 2nd edition, Springer-Verlag.

[This is required reading for anyone who works with observational data.]

Wood, S. N., 2006. *Generalized Additive Models. An Introduction with R*. Chapman & Hall/CRC.

[This is an elegant treatment of linear models and generalized linear models, as a lead-up to generalized additive models.]