Models Versus Full Reality

True Models

A fundamental paradigm in FW663 is that there is no “true model" that generates the biological data we observe (see, for example Bancroft and Han 1977). We believe that “truth" (full reality) in the biological sciences has essentially infinite dimension and hence full reality cannot be revealed with only finite samples of data and a “model" of the information in the data. It is generally a mistake to believe that there is a simple, “true model" in the biological sciences and that, during data analysis, this model can be uncovered and its parameters estimated. Instead, biological systems are complex with many small effects, interactions, individual heterogeneity, and individual and environmental covariates (most being unknown to us); we can only hope to identify a model that provides a good *approximation* to the data available. The words “true model" represent an oxymoron, except in the case of Monte Carlo studies whereby a model is used to generate “data" using pseudo-random numbers (we will use the term “generating model" for such computer-based studies).

Taub (1993) suggests that unproductive debate concerning true models can be avoided by simply recognizing that a model is not truth by definition. A model is a simplification or approximation of reality and, hence will not reflect all of reality. Full truth (reality) is elusive (see de Leeuw 1988). Box (1976) noted that “all models are wrong, but some are useful." While a model can never be “truth," a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless. Model selection methods try to rank models in the candidate set relative to each other; whether any of the models is actually "good" depends primarily on the quality of the science and *a priori* thinking that went into the modeling. Proper modeling and data analysis tells what inferences the data support, not what full reality might be (White et al. 1982:14-15, Lindley 1986). Models, used cautiously, tell us “what effects are supported by the (finite) data available." Increased sample size (information) and measures of relevant covariates allow us to chase full reality, but never quite catch it.

Tapering Effect Sizes

We believe that there are tapering effect sizes in many biological systems; that is there are often several large, important effects, followed by many smaller effects and, finally, followed by a myriad of yet smaller effects. These effects may be sequentially unveiled as sample size increases. The main, dominant effects might be relatively easy to identify and support, even using fairly poor analysis methods, while the second order effects (e.g., a chronic treatment effect or an interaction term) might be more difficult to detect. The still smaller effects can only be detected with very large sample sizes (cf. Kareiva 1994 and related papers), while the smallest effects have little chance of being detected, even with very large samples. Rare events, that have large effects, may be very important but quite difficult to study. Approximating models must be related to the amount of data and information available; small data sets will appropriately support only simple models with few parameters, while more comprehensive data sets will support, if necessary, more complex models.

This tapering in “effect size" and high dimensionality in biological systems might be quite different from some physical systems where a small dimensioned model with relatively few parameters might accurately represent full truth or reality. Biologists should not believe that a simple, “true model" exists that generates the data observed, although some biological questions might be of relatively low dimension and could be well approximated using a fairly simple model. The issue of a range of tapering effects has been realized in epidemiology where Michael Thun notes “. . . you can tell a little thing from a big thing. What's very hard to do is to tell a little thing from nothing at all" (Taubes 1995). Full reality will always remain elusive.

Parsimony in Understanding

Often the investigator wants to simplify some representation of reality in order to achieve an understanding of the dominant aspects of the system under study. If we were given a nonlinear formula with 200 parameter values we might make correct predictions, but it would be difficult to *understand* the main dynamics of the system without some further simplification or analysis. Thus, one might tolerate some inexactness (an inflated error term) to facilitate a more simple and useful understanding of the phenomenon. We will provide examples where some inferences are best made using a model that is not the K-L best model.

Literature Cited

Bancroft, T. A., and Han, C-P. (1977). Inference based on conditional specification: a note and a bibliography. *International Statistical Review* **45**, 117-127.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association* **71**, 791-799.

de Leeuw, J. (1988). Model selection in multinomial experiments. Pages 118-138 *in* T. K. Dijkstra (Ed.) *On model uncertainty and its statistical implications.* Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, NY.

Kareiva, P. (1994). Special feature: higher order interactions as a foil to reductionist ecology. *Ecology* **75**, 1527-1559.

Lindley, D. V. (1986). The relationship between the number of factors and size of an experiment. Pages 459-470 *in* P. K. Goel, and A. Zellner (Eds.) *Bayesian inference and decision techniques.* Elsevier Science Publishers, New York, NY.
fitting. *Suri-Kagaku* (Mathematic Sciences) **153**, 12-18. (In Japanese).

Taub, F. B. (1993). Book review: Estimating ecological risks. *Ecology* **74**, 1290-1291.

Taubes, G. (1995). Epidemiology faces its limits. *Science* **269**, 164-169.

White, G. C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982). *Capture-recapture and removal methods for sampling closed populations*. Los Alamos National Laboratory, LA-8787-NERP, Los Alamos, NM, USA. 235pp.