# Living (and Dying) in a Deterministic World

Christl A. Donnelly and Neil M. Ferguson
*Department of Infectious Disease Epidemiology*
*Imperial College School of Medicine*
*St Mary's Campus*
*Norfolk Place*
*London W2 1PG, UK*
c.donnelly@ic.ac.uk

## 1. The Questions and the Models

Biostatistical models typically describe associations answering questions like

- Are people who smoke more likely to be diagnosed with lung cancer? or
- Do HIV-infected patients treated with drug A live longer (or in the shorter term have lower viral load measurements) than similar patients treated with drug B?

using cohort studies and clinical trials, respectively, to collect the data necessary to formulation public health recommendations. Biomathematicians tend to have different sorts of questions such as:

- How does a tumor grow and spread throughout a patient's body?
- Why do some diseases appear to have 2-year cycles?

The statistical questions as I have defined them here tend to have yes/no answers and then a measure of how much. The mathematical modeling questions cannot be generalized in such a manner.

Both biostatisticians and mathematical modelers use deterministic models – describing the mean behavior of a system or process. It is the form of models that most typically differentiates mathematical modelers, be they biologists or epidemiologists, from biostatisticians. Whilst biostatisticians talk in terms of generalized linear models, proportional hazard models or multi-level models, mathematical modelers tend to work with dynamical models that aim to describe the processes giving rise to the data.

## 2. The Problem of Non-linearity

Armed with a model and embarking on an analysis of data, applied biostatisticians and mathematical modelers both wish to fit the model to the data, estimating parameters (and corresponding measures of uncertainty) and judging model adequacy. A key factor that can complicate their tasks is non-linearity.

What does non-linearity mean exactly? Take a simple generalized linear model:

$$g(y_i) = \alpha + \beta \mathbf{x}_i + \varepsilon_i.$$

The derivative of the model with respect to each parameter:

$$\frac{dg(y_i)}{da} = 1 \quad \text{and} \quad \frac{dg(y_i)}{db} = x_i$$

are not parameter dependent, so the model is linear. The regularity conditions required for the asymptotic theory underlying likelihood ratio tests and confidence intervals are no problem. How does this go wrong for non-linear models? Considerable statistical attention has been focussed on the particular irregular problem of finite mixture models (see for example Titterington et al. 1985). Consider the deceptively simple mixture PDF:

$$g(y_i)= \gamma f(y_i,\theta_1)+(1-\gamma)f(y_i,\theta_2)$$

where $0\leq\gamma\leq1$ and $\theta_1\leq\theta_2$. An obvious hypothesis: $H_0: \gamma=0$ is equivalent to $\theta_1=\theta_2$. The non-linearity in the system

$$\frac{dg(y_i)}{dg} = f(y_i,q_1) - f(y_i,q_2), \quad \frac{dg(y_i)}{dq_1} = g\frac{df(y_i,q_1)}{dg} \quad \textbf{and} \quad \frac{dg(y_i)}{dq_2} = -g\frac{df(y_i,q_2)}{dg}$$

causes this to be an irregular problem (i.e. the likelihood ratio test for $H_0: \gamma=0$ is not the typical single degree of freedom chi-squared test). See Chen et al. (2001) for a full discussion and alternative test.

Even the simplest possible dynamical model for an epidemic, the so-called S-I model, referring to susceptible (S) and infected (I) individuals, has a complex non-linear relationship between the single parameter β and the observed data. If the infection rate depends upon the absolute number of infected individuals, then the model can be described by the simple differential equation:

$$\frac{dS}{dt} = -bSI\ ,$$

but solving this equation for a closed population (i.e. $S+I=N$ at all times) gives the equation:

$$S = N\left(1+\frac{(N-S_0)}{S_0}\exp(bNt)\right)^{-1}$$

quite a complicated link function if the data were such that at times $t_1$, $t_2$, …$t_M$ the number of infected individuals were counted. Clearly dS/d$\beta$ will depend on $\beta$. If additional complexities (including a latent (non-infectious) period, recovery from infection, births and deaths) were added, then the set of differential equations must be solved numerically making matters even more difficult.

## 3. The Results and Consequences

Likelihood-based estimation methods provide a well-defined structure for obtaining parameter estimation and corresponding confidence intervals for dynamical models. Such methods have been used to epidemic models to data on HIV/AIDS (Cox and Medley 1989), bovine spongiform encephalopathy (BSE or mad cow disease) (Anderson *et al.* 1996; Donnelly and Ferguson 1999) and most recently the current foot and mouth disease epidemic in Great Britain (Ferguson, Donnelly and Anderson, 2001).

What are the consequences of ignoring the irregular nature of the problems and proceeding with likelihood-based methods? The departure of the results from those predicted by standard asymptotic likelihood theory can depend on the particular parameter values as well as the model structure itself. For example, provided $0<\gamma<1$ and $\theta_1<\theta_2$, the model

$$g(y_i)= \gamma f(y_i,\theta_1)+ (1-\gamma)f(y_i,\theta_2)$$

is well defined with 3 distinct parameters. Thus, if 10 binomial samples were simulated from the probability

$$\gamma \exp(-\theta_1 t)+(1-\gamma)\exp(-\theta_2 t)$$

and N=100 for times t=1,2, …, 10 and maximum likelihood estimates were obtained for all three parameters, we would expect the standard goodness-of-fit test (2 times the difference in the maximized and saturated log likelihoods) to be chi-squared distributed with (10-p) degrees of freedom. However, the results obtained were quite different for extreme values of $\gamma$.

Simulations clearly demonstrate that when the ability of changes in a parameter to improve the fit of the model to the data is limited, fitting this parameter may not equate to a full degree of freedom. Although the simulation may seem artificial in form, epidemic models may present similar problems: for example, the rate of secondary transmission of infection (for example maternal transmission) may be of little consequence, and thus problematic to include as an estimated parameter, if primary infection is relatively uncommon. Unfortunately the complexity of multi-dimensional parameter estimation in such settings is sufficiently difficult numerically to make simulation studies of such problems infeasible.

Whilst the use of methods such as Markov chain Monte Carlo (MCMC) for stochastic epidemic models offers an alternative approach for some systems, the number of events would be so large in some settings (such as the BSE epidemic in Great Britain with roughly 900,000 infections over the course of the epidemic resulted in over 177,000 confirmed cases of disease) as to make such approaches impossible. While likelihood-based fitting of deterministic dynamical models provides a valuable framework for gaining insights into the processes underlying observed data, further development of the theoretical basis for such irregular applications of likelihood methods is a priority for future work.

## REFERENCE

Anderson RM, Donnelly CA, Ferguson NM, *et al.* (1996) Transmission dynamics and epidemiology of BSE in British cattle. *Nature* **382**, 779-788.

Chen H, Chen J, Kalbfleisch JD. (2001) A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society* B **63**, 19-29.

Donnelly CA and Ferguson NM. (1999) *Statistical Aspects of BSE and vCJD: Models for Epidemics,* Monographs on Statistics and Applied Probability, Chapman & Hall/CRC Press.

Ferguson NM, Donnelly CA, Anderson RM. (2001) The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, **292**, 1155-1160. Published online 12 April 2001; 10.1126/science.1061020.

Cox DR, Medley GF. (1989) A process of events with notification delay and the forecasting of AIDS. *Philosophical Transactions of the Royal Society London* B **325**, 135-145.

Titterington DM, Smith AFM, Makov UE (1985) *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

## RESUME

Vie (et mort) dans un monde déterministe

Biostatisticiens et modélisateurs (de modèles mathématiques) utilisent habituellement différentes approches pour ajuster leurs modèles. Bien que les problèmes non linéaires puissent survenir tant en statistique qu'avec les modèles mathématiques, les modèles utilisés pour décrire les épidémies sont presque toujours non linéaires. Les solutions numériques aux données observées sachant la moyenne déterministe, qui utilisent des modèles dynamiques dans le cadre d'une méthode du maximum de vraisemblance, facilitent l'estimation des paramètres et les tests d'hypothèse. Cependant, la théorie asymptotique qui permet de calculer les tests du rapport des vraissemblance exige souvent des conditions qui ne sont pas respectées par ces modèles non linéaires.