Analyse Problems, Not Data

Peter J Diggle
(CHICAS, Lancaster University Medical School)

November 1, 2018

Abstract

The last fifty years or so have seen a transformational change in statistical methodology, from a discrete set of specific methods to a single, integrated paradigm. An early example is the seminal paper by Nelder and Wedderburn (1972) that introduced the unifying concept of the generalised linear model for independently replicated data. Later computational advances have stimulated a comparable unification for modelling data with various kinds of dependence, for example in time and/or in space. I argue that this transformation should encourage statistical scientists to change their focus from analysing data to solving problems. I give an example from an ongoing study of the acquisition of natural immunity to leptospirosis among slum-dwellers in northern Brazil.

Key words. data model; leptospirosis; process model; serial dilution assay

1 Introduction

The last fifty years or so have seen a steady trend towards de-compartmentalization of statistical methodology. Perhaps the best known example of this trend in the context of independently replicated data was the seminal paper of Nelder and Wedderburn (1972), which collected many apparently different methods for the analysis under the single umbrella of the generalized linear model (GLM). Over the subsequent 20 years, many authors considered extensions of the GLM to accommmodate particular types of dependence in data, including temporal (Stiratelli, Laird and Ware, 1984; Liang and Zeger, 1986) and spatial (Clayton and Kaldor, 1987; Besag, York and Mollié, 1991). A very wide class of models can be captured under the heading of the *generalized linear mixed model* (GLMM). A GLMM is a GLM whose linear predictor includes a (possibly high-dimensional) vector of unobserved random effects. Breslow and Clayton (1993) give a very good review of the literature on GLMMs up to that time, describe an approximate method for circumventing the general intractability of the associated likelihood and give several examples.

The intractability of the likelihood function was a major concern during the early development of spatial statistical methods in the 1970s. For example, two classic papers, Besag (1974) and Ripley (1977), greatly extended the availability of parametric models for spatial lattice data and spatial point process data, respectively, but were forced to use ad hoc methods for fitting the models to data. However, both of these authors recognised the potential for Monte Carlo methods of inference to get round the intractability problem. In the early 1990s, Monte Carlo methods for likelihood-based inference entered the statistical mainstream, following the influential papers by Gelfand and Smith (1990) and Geyer and Thomson (1992) who, respectively, considered Bayesian and non-Bayesian approaches.

These methodological advances, and the parallel development of more flexible software environments (notably R), have shifted the emphasis of parametric statistical inference away from developing methods tailored to particular kinds of *data* in favour of applying statistical method singular to an ever-wider range of scientific *problems*.

In this short note, I use an example from infectious disease epidemiology to suggest how a single data-set might be analysed in different ways, according to the scientific question of primary interest.

2 A longitudinal study of sub-clinical leptospirosis in a Brazilian slum community

2.1 Serial dilution assays

In infectious disease epidemiology, evidence of a person experiencing a sub-clinical infection event over a specified follow-up period is often gathered by periodically testing for presence, and if present the strength, of an antibody response using a serial dilution assay. In this procedure, a blood-sample is tested against a standardised challenge and if this gives a negative result the antibody response, W say, is recorded as "below detection limit." If the test gives a positive result, the sample is diluted by a known factor and the test is repeated, and so on. This generates an integer response, $K = 0, 1, 2, ...$, the number of dilutions required to return a negative result. The value of K corresponds to an interval-censored version of W, i.e. $K = k$ if and only if $ckd \leq W \leq c(k+1)d$, where c is the detection limit and d the dilution factor. Although this interpretation is rarely used explicitly, it is implicit in the standard practice of requiring an increase of at least two in successive values of K in order to declare the occurrence of an infection event at some time during the follow-up interval in question.

2.2 A leptospirosis cohort study

In a recent cohort study of sub-clinical leptospirosis infections in a Brazilian slum community, study-participants provided blood-samples at approximately yearly intervals. For each blood-sample a yes/no antibody response was recorded from a series of two-fold dilutions and an infection event was deemed to have occurred if, over the follow-up interval between successive blood-samples, either (a) the antibody response changes from non-detectable, i.e. $K = 0$, to positive, or (b) the number of dilutions, K, needed to return a negative result increases by at least two (Hagan et al, 2016).

2.3 Problem 1: understanding spatio-temporal variation in risk of infection

Hagan et al (2016) used the criteria (a) and (b) above to convert each sequence of five recorded values of K on a study-participant, say $k_{ij} j = 1, ..., 5$, to a binary sequence $Y_{ij} = 0/1 : j = 1, ..., 4$. They then fitted a generalized linear mixed model as follows. Let p_{ij} denote the probability that $Y_{ij} = 1$. Hagan et al (2016) modelled the p_{ij} as

$$\log\{p_{ij}(1 - p_{ij})\} = z'_{ij}\beta + S(x_i, j) + U_i. \tag{1}$$

In (2), z_{ij} is a vector of covariates with regression parameters β, $S(x, t)$ is a spatio-temporal Gaussian process and $U_i \sim \mathrm{N}(0, \nu^2)$ are uncorrelated individual-level random effects.

I would argue that this was a sensible strategy because the aims of the study were: to estimate covariate-effects on an individual's risk of infection; to map unexplained variation in risk due to unmeasured environmental risk-factors.

2.4 Problem 2: does a sub-clinical infection confer partial immunity to future infection?

The approach of Hagan et al (2016) now runs into some difficulty. A simple, and superficially attractive way to answer this question is to add to the covariate vector z_{ij} in (2) a person's antibody response at their previous follow-up. Partial immunity would then be indicated by a negative regression coefficient. But the reasoning behind this is flawed. Recall that the integer-valued response K can be considered as an interval-censored version of a real-valued antibody response, W. Let W_1 and W_2 be the values of X at two consecutive follow-up times t_1 and t_2. Then, the event I, a re-infection at some time between t_1 and t_2, is declared when (approximately, because of the interval censoring) $W_2 > 4W_1$. It follows that I and W_1 cannot be independent. Moreover, the circumstances in which $g(w) = \mathrm{P}(W_2 > 4W_1 | W_1 = w)$ could be other than a decreasing function of w would seem to be quite contrived.

An alternative, model-based solution to problem 2 is the following. To avoid unnecessary notational complications, we describe a version of the model for a single individual without covariates.

2.4.1 Process model

Let $W(t)$ denote the antibody response of an individual at time $t \geq 0$; for an infection-naive individual, $W_i(0) = 0$. Assume that $W(t)$ jumps by random iid amounts V_j at infection times

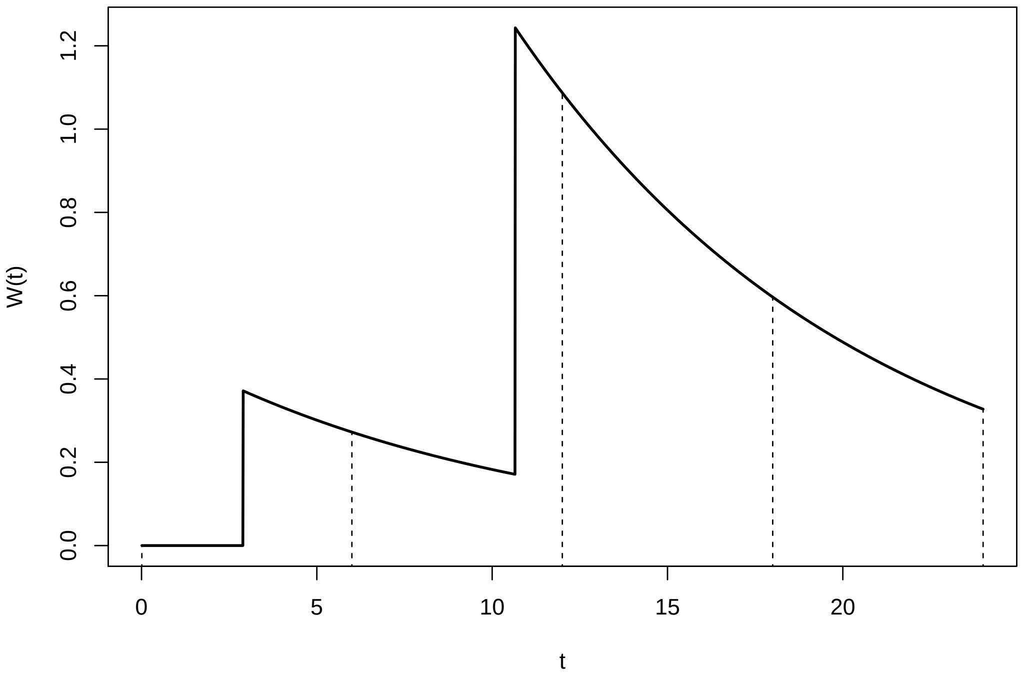

Figure 1: A simulated realisation of model (2) for the time-evolution of an individual's antibody response over 24 months, with follow-up times indicated at 0, 6, 12, 18 and 24 months.

t_j and decays exponentially at a rate ϕ between infection events. Finally, assume that infection events follow a Poisson process with intensity

$$\lambda(t) = \exp\{\alpha(t) + W(t)\beta\} \tag{2}$$

The parameter of interest here is β, with a negative value indicating that infection events confer partial immunity to future infections. Figure 1) shows a realisation of $W(t)$ under model (2), with constant $\alpha(t) = -1$, exponential decay factor 0.1 per month, $\beta = -3$ and iid V_j following exponential distributions with mean 1.

2.4.2 Data model

The observed response from an individual is the sequence of values of K_j at a set of pre-specified follow-times $f_j : j = 1, ..., n$. Each K_j is an interval-censored version of $\alpha W(t_j)$, where α is an unknown constant of proportionality; an observation $K_j = 0$ corresponds to $\alpha W(t_j) < c$, where c is the detection limit of the assay, whilst $K_j = k > 0$ corresponds to $c2^{k-1} < \alpha W(t_j) < c2^k$.

2.4.3 Fitting the model

Evaluation of the probability of an observed sequence $K_j : j = 1, ..., n$, and hence the likelihood, requires marginalisation with respect to the latent variables t_j and V_j. Whilst this is possible in principle, much more extensive data than were available for the analysis reported in Hagan et al (2016) would be needed to make likelihood-based inference a realistic proposition. For example, to capture the behaviour of the process $W(t)$ illustrated in Figure1 would need much more frequent follow-up than the six-month intervals shown. Alternatively, information could be added to the data either by using Bayesian inference with strong and well-informed priors on the model parameters or by incorporating data from other studies. For example, in an unpublished Yale University PhD Thesis, Dr Katharine Owers has used data from a single-source outbreak of leptospirosis reported in Lupidi et al (1991) to obtain a confidence interval for the exponential decay rate parameter ϕ.

3 Conclusion

Computationally intensive methods of inference developed over the last few decades have made it possible to fit a rich variety of models to data using principled, i.e. likelihood-based, methods of inference. An attendant danger is a temptation to fit over-complex models to sparse data; the virtues of parsimony in statistical modelling cannot be over-stated. Also, some questions, notably those addressed in late-phase randomised trials of novel drug-therapies, are better answered using design-based rather than model-based inference. With these qualifications, the freedom of the contemporary statistician to analyse a scientist's data in a way that answers *their* research question and can incorporate *their* subject-matter knowledge is refreshing. The papers that follow in this special issue of *Spatial Statistics* well-illustrate this freedom, ranging widely in both the scientific problems that they address and the specific methods that they use.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion).*Journal of the Royal Statistical Society* B **36**, 192–225.

Besag, J., York, J., and Mollié, A. (1991), Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Clayton, D., and Kaldor, J. (1987), Empirical Bayes estimates of age- standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *Journal of the Royal Statistical Society*, B **54**, 657–699.

Hagan, J.E., Moraga, P., Costa, F., Capian, N., Ribeiro, G.S., Wunder, E.A., Felzemburgh, R.D.M., Reis, R.B., Nery, N., Santana, F.S., Fraga, D., dos Santos, B.L., Santos, A.C., Queiroz, A.,Tassinari, W., Carvalho, M.A., Reis, M.G., Diggle, P.J. and Ko, A.I. (2016). Spatiotemporal determinants of urban leptospirosis transmission: Four-year prospective cohort study of slum residents in Brazil. *Public Library of Science: Neglected Tropical Diseases*, **10**, pp.e0004275.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lupidi, R., Cinco, M., Balanzin, D., Delprete, E. and Varaldo, P.E. (1991). Serological follow-up of patients involved in a localized outbreak of leptospirosis. *Journal of Clinical Microbiology*, **29**, 805–809.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, A **135**, 370–384.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society* B **39**, 172–212.

Stiratelli, R., Laird, N., and Ware, J. (1984). Random effects models for serial observations with binary responses. *Biometrics*, **40**, 961–971.