ELSEVIER

Mixed models: getting the best use of parasitological data

Steve Paterson and Joanne Lello

Department of Biological Sciences, University of Stirling, UK, FK9 4LA

Statistical analysis of parasitological data provides a powerful method for understanding the biological processes underlying parasite infection. However, robust and reliable analysis of parasitological data from natural and experimental infections is often difficult where: (1) the distribution of parasites between hosts is aggregated; (2) multiple measurements are made on the same individual host in longitudinal studies; or (3) data are from 'noisy' natural systems. Mixed models, which allow multiple error terms, provide an excellent opportunity to overcome these problems, and their application to the analysis of various types of parasitological data are reviewed here.

Statistical models provide powerful tools to investigate the biological processes underlying parasite infection and disease, including exposure and susceptibility of hosts, and infectivity and virulence of parasites [1–5].

Statistical models in parasitology

Statistical models in parasitology generally have two objectives: (1) to test hypotheses regarding the significance of different biological factors on the duration, intensity or prevalence of a parasitic infection; and (2) to quantify the effect of these biological factors on a parasitic infection. For example, does treatment with drug A significantly reduce the duration or intensity of a parasitic infection relative to untreated controls and, if so, by how much? Is the presence of host species B associated with higher parasite burdens in host species A and, if so, by how much?

A common problem with parasitological data (and biological data in general) is PSEUDOREPLICATION (see Glossary), where observations are correlated either in space or through time [6–8]. These correlations arise, for example, if samples are taken from the same geographical location or are taken at repeated time points from the same individual. Because these observations are not independent of each other, testing the significance of treatment effects or quantifying the size of these effects is confounded by the difficulty in separating treatment effects from spatial or temporal correlations [6,9]. Thus, a group of observations associated with repeated measures on an individual or with multiple samples from the same site may be consistently higher or lower than other such groups of observations owing to innate differences between individuals or sites [7,8,10]. In addition, observations within a group might exhibit autocorrelation, i.e. the degree of correlation between a pair of observations might be higher if closely related in space or time [8,9]. For example, in a longitudinal study, one observation might depend on the previous observation.

How then should pseudoreplication in parasitological data be dealt with? One approach is to design the experiment or sampling regime such that only independent data are analysed [9]. Suppose, for example, one wants to follow the progress of a parasitic infection through time by inoculating hosts with parasites and monitoring the infection at successive time points. By only using different hosts at each time point, one can avoid pseudoreplication entirely. For some types of observation, such as gut counts of helminths, this will be the natural design of the experiment if only destructive sampling can be used. Analogously, it may be possible to design sampling regimes such that only a single sample is taken from each site. In some studies, one may take several measures from

Glossary

Blocks : similar observations can be assigned to blocks, where blocks are assumed to vary randomly from each other.
Error term : random departure of the response variable from that predicted by either the fixed or random effects.
Fixed effects : the mean predicted value of the response variable over hypothetical repetitions, independent of the group sampled.
Geometric decrease : a progression that decreases proportionally towards an asymptote of zero (e.g. 1, 1/2, 1/4, 1/8, etc).
GLM (generalized linear model) : a common type of statistical model used to analyse independent data with a non-normal error distribution.
GLMM (generalized linear mixed model) : an extension of a GLM that allows random effects to be modelled.
Linear model : a basic regression model specification appropriate for independent data with a normal distribution.
Mixed model : a model that incorporates random effects between groups of observations.
Normal distribution : a symmetrical distribution ('a bell curve') widely applicable to describe the probability distribution of continuous variables.
Overdispersion : where the variance of the response variable is greater than expected for a particular distribution; for example, for count data, one predicts the variance should be equal to the mean but the variance of the observed distribution of helminths between hosts often exceeds the mean.
Poisson distribution : an asymmetric distribution describing the probability of a certain number of events occurring in a specified space or period of time.
Poisson process : a stochastic process where the accumulation of events occurs according to an underlying, probabilistic rate.
Pseudoreplication : non-independence of data arising from spatial or temporal correlations between observations.
Random effects : the consistent departure of the response variable for a group of observations from that predicted from the fixed effects and are drawn from a random distribution.
REML (restricted estimate maximum likelihood, also sometimes called residual maximum likelihood) : a computer algorithm often used to fit the parameters of mixed models.

Corresponding author: Steve Paterson (steve.paterson@stir.ac.uk).

http://parasites.trends.com 1471-4922/03/$ - see front matter doi:10.1016/S1471-4922(03)00149-1

an individual or a sampling site but only be interested in one general feature of this group of observations, such as peak parasitaemia in an individual or prevalence of infection in a village. Using a single measure per group of observations ensures that measures are independent of each other and so can be analysed by standard methods such as LINEAR MODELS, *t* tests, non-parametric tests and so on [8,9].

An alternative approach to deal with pseudoreplication is to use nested analysis of variance (ANOVA) [9,11]. This method is derived from agricultural experiments where fields are divided into smaller BLOCKS and the crop within each block is given the same treatment combination, while at the same time it is assumed that there is some innate variability between blocks. By extension to parasitology, observations from individuals or samples that are similar to each other and that receive the same treatment combination can be assigned to the same block. Nested ANOVAs include an extra ERROR TERM to describe the innate variation between blocks before determining whether there is a significant effect of treatment, thereby separating treatment effects from block effects. A major consideration of nested ANOVAs is that, in general, for these tests to be carried out the data must be both orthogonal and balanced (i.e. all treatment combinations must be present and all treatment combinations must have equal numbers of observations). These restrictions can sometimes be relaxed, for example by randomly removing data or by bootstrapping data, but usually at the expense of inflated standard errors [11]. Nested ANOVAs are therefore ideal in many types of experimental design, prospective studies or in certain sampling regimes. For many other situations, where the researcher has less control over the form of the data to be analysed, regression methods might be more appropriate in dealing with unbalanced, non-orthogonal data.

The remainder of this review will concentrate on mixed regression models as a general approach to deal with spatial and temporal correlations in parasitological data. This is not to imply that they represent *a priori* a better method to deal with all forms of parasitological data, but these models certainly provide powerful tools to analyse many types of data for which the methods above are inappropriate.

Mixed models of parasite infection

MIXED MODELS are a group of models designed to analyse observations structured into groups. These models are defined by FIXED EFFECTS, which are common to any group of observations one could make, and RANDOM EFFECTS, which are specific to a particular group of observations (Box 1). These random effects are coefficients drawn from a random distribution, generally a NORMAL DISTRIBUTION [8]. Parameter estimation in these models involves fitting coefficients for the fixed terms and for the variance component describing the distribution of the random effects. Mixed models can therefore be usefully applied to some types of parasitological data and can help to overcome problems of pseudoreplication. Where groups of observations are made, either on a single individual through time or on a group of samples from the same site, random effects are assigned to each group of observations. In the simplest case, a single random term is used to describe consistent deviations between groups of observations, for example due to innate differences between the susceptibility of individuals in a longitudinal study or innate differences in rates of exposure at different sampling sites. Fitting such deviations between groups of observations as random effects has the advantage that relatively few parameters are required to describe the variance of the distribution from which the random effects are drawn, which loses relatively few degrees of freedom from the model regardless of the number of different groups of observations in the analysis. More-complex formulations are also possible, with more than one random effect assigned to each group of observations or to have random effects interacting with each other or be nested within each other.

The final part of a mixed model is the error term, which defines the deviation of the response variable from that predicted by either the fixed or random effects. In the simplest case, the error term associated with each observation is simply drawn from a normal distribution implying that the correlations present within the data are accounted for by the inclusion of random effects. However, spatial and temporal data can also generate autocorrelations between observations within groups, for example where one observation depends on the last, or observations exhibit correlations that depend on their temporal or spatial proximity. Given a group of observations, and where the errors within groups are correlated to each other, these error terms can be drawn from a multivariate normal distribution, with covariances that specify the correlation between the errors associated with pairs of observations. Some common forms of error structure include: equicorrelated errors, where the measures within a group are correlated equally to each other; autoregressive errors in longitudinal data, where an error term associated with an observation may be correlated to that of the previous observation; and Markov errors, where the correlations between errors show a GEOMETRIC DECREASE with time or space (Box 1).

Another use of the random terms in mixed models, which is of particular use in field data, is to 'clean up' noisy data. Field data tends to have a large amount of random noise associated with it, which can cloud the true effects of fixed model terms. In the example of a study that uses a large number of sites with multiple samples taken from each site, fitting random effects for the sampling sites controls for pseudoreplication within these sites and, in doing so, cleans up the noise associated with variation between these sites and allows the significance of factors such as geographic location, climate, and so on, to be tested within the fixed component of the model [12]. Similarly, in a long-term study, the year of sampling could cause variation over and above known seasonal effects, accounted for in the fixed model. Moreover, including these terms as random effects, rather than as fixed effects, uses only a single degree of freedom. In our field example, if we had 20 years, fitting the year in the fixed model might have accounted for the variation but would have used 19 degrees of freedom instead of one; thus, accounting for

Box 1. A pocket guide to mixed models

Linear models (LMs) versus linear mixed models (LMMs)

Consider a linear model having a regression equation:

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + e_i, \quad \text{(Eqn I)}$$

where y_i is an observation (i.e. the response variable) on the ith of n individual hosts, $\mathbf{X}_i$ is a row vector of q explanatory variables, $\boldsymbol{\beta}$ is a column vector of q coefficients and e_i is the error component associated with the ith individual. A mixed model is a straightforward extension of a linear model to include random terms associated with groups of observation [8]. For example, if p_i observations are made on the ith individual host to give a vector of observations $\boldsymbol{y}_i = (y_{i1}, ..., y_{ij}, ..., y_{ip_i})^{\mathrm{T}}$, a linear mixed model has the regression equation:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad \text{(Eqn II)}$$

where the fixed component is defined by $\mathbf{X}_i[p_i \times q]$, which is a matrix of q explanatory variables for each observation and $\boldsymbol{\beta}$, which is a column vector of q fitted coefficients for these explanatory variables as before. The random component of the model is defined by $\mathbf{Z}_i[p_i \times r]$, which is a matrix of r explanatory variables for each observation and $\mathbf{b}_i$, which is a column vector of r coefficients for these explanatory variables specific to each group of observations. $\mathbf{b}_i$ has a multivariate normal distribution, $\mathbf{b}_i \sim N_p(0, \mathbf{B})$, where $\mathbf{B}[r \times r]$ is the covariance matrix of $\mathbf{b}_i$ over the population of n individuals. In the simplest case, only the mean value of each group of observations is assumed to vary as a random effect with variance σ_b^2 over the population of individual hosts, $b_i \sim N_p(0, \sigma_b^2 \mathrm{I}_p)$. In more complex cases, $\mathbf{b}_i$ can be constructed to be dependent on time or other variables.

$\mathbf{e}_i[p_i \times 1]$ denotes the errors remaining after fitting the fixed and random components to the model. In the simplest case, these will be distributed according to a normal distribution with variance σ_e^2, $e_{ij} \sim N(0, \sigma_e^2)$. However, other error structures are also possible. In particular, errors might be autocorrelated through time, such that an observation at time t_j is correlated to the observation at time t_k. The most obvious treatment is to give the errors a Markov correlation structure, $\mathrm{e}_i \sim N_p(0, \mathrm{E}_i)$, which decreases geometrically through time such that

$$E_{jk} = \sigma^2 \rho^{|t_j - t_k|}, \quad \text{with } 0 \leq \rho \leq 1. \quad \text{(Eqn III)}$$

The correlations $\rho^{|t_j - t_k|}$ clearly decrease (geometrically) as the separation $|t_j - t_k|$ increases.

The major hurdle to using mixed models is the difficulty in parameter estimation. Reformulating Eqns I and II illustrates the problem. A linear model can be given by:

$$y_i \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma_e^2), \quad \text{(Eqn IV)}$$

where σ_e^2 is the variance of y_i around the fixed component. A mixed model is given by:

$$y_i \sim N_p(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{V}_i) \quad \text{(Eqn V)}$$

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{B}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{E}_i$$

where $\mathbf{V}_i$ is the covariance matrix of $\mathbf{y}_i$ around the fixed component. Thus, estimation of σ_e^2 in Eqn IV is a straightforward estimation of a single parameter. However, estimation of $\mathbf{V}_i$ in Eqn V can involve simultaneous estimation of many parameters in both $\mathbf{B}$ and $\mathbf{E}_i$. Elegant, analytical methods to estimate these parameters are not always available and so brute force, computer-intensive methods such as restricted maximum likelihood (REML), simplex algorithms and Monte Carlo Markov Chain (MCMC) are often used instead [15,34,35].

Generalized linear models (GLMs) versus generalized linear mixed models (GLMMs)

GLMs are used to model data that are not necessarily normally distributed and take the general form:

$$\mathrm{E}[y_i] = \mu_i \quad \text{(Eqn VI)}$$

$$\mathrm{g}(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$$

where g(.) is a link function used to relate a vector of explanatory variables, $\mathbf{X}_i$, associated with observation y_i to the expected value of y_i, μ_i. Random effects can be added to GLMs to give the analogous mixed models, GLMMs [15,16]:

$$\mathrm{E}[y_i|\mathrm{b}] = \mu_i \quad \text{(Eqn VII)}$$

$$\mathrm{g}(\mu_i) = \mathbf{x}_i\boldsymbol{\beta} + \mathrm{z}_i\mathrm{b}$$

Mixed models of aggregation

Its worth noting the properties of a mixed model with Poisson-distributed errors. Here, the standard GLM has the properties [36]:

$$\mathrm{E}[y_i] = \mu_i \quad \text{(Eqn VIII)}$$

$$\mathrm{var}(y_i) = \mathrm{E}(\mu_i)$$

Note that the variance is equal to the mean. The analogous GLMM has the properties [15]:

$$\mathrm{E}[y_i|\mathbf{b}] = \mu_i \quad \text{(Eqn IX)}$$

$$\mathrm{var}(y_i) = \mathrm{E}(\mu_i) + \mathrm{var}(\mu_i)$$

In this case the variance is greater than the mean, a characteristic of aggregated parasite distributions. Note that as the variation between hosts in μ_i increases so too does the level of aggregation. Where the μ_i are distributed according to a gamma distribution with mean $\mu = \mathrm{E}(\mu_i)$ and index k, the y_i will be distributed according to a negative binomial distribution [36] with:

$$\mathrm{var}(y_i) = \mu + \mu^2/k \quad \text{(Eqn X)}$$

where k is the overdispersion parameter and is inversely related to the degree of aggregation [1].

random variance in this manner provides the model with more explanatory power [6].

A common feature of parasitological data is that they seldom exhibit a normal distribution (e.g. proportion data from parasitaemia measures in malaria or aggregated distributions of helminths between hosts) [13,14]. A common method to deal with such data is provided by GENERALIZED LINEAR MODELS (GLMS), which are extensions of standard linear models and which can accommodate various non-normal error distributions (reviewed in Ref. [1]). However, GLMs assume that the data are independent and so, as discussed previously, are not suitable for all types of parasitological data. Box 1 shows how non-normal distributions can be handled in mixed models using GENERALIZED LINEAR MIXED MODELS (GLMMS), which are analogous to GLMs but incorporate random terms [15,16]. GLMMs therefore provide a powerful method to analyse parasitological data that does not conform to a normal distribution and, at the same time, controls for correlations between measures that arise from grouped observations. For example, Diggle *et al.* [12] have recently used this approach to model the prevalence of malaria in children sampled from 65 villages in the Gambia. They found a large amount of variation between villages, which they were able to control for before testing the significance of various fixed effects. This confirmed the importance of age and bednet use for malaria prevalence within villages but did not indicate support for satellite-based measures, such as the greenness of surrounding vegetation, to predict malaria prevalence between villages. Further analysis of the error component of the model showed that variation in prevalence between

Table 1. Computer packages available for mixed models[a]

Package	Supplier	Mixed model methods	Pros	Cons
SAS	SAS Institute Inc. (Cary, NC) http://www.sas.com	Nested ANOVA, LMM, GLMM	Widely used statistics package with plenty of support and add-in modules	Expensive
S-plus	Insightful Inc. (Seattle, WA) http://www.insightful.com	Nested ANOVA, LMM, GLMM	Widely used statistics package with plenty of support and add-in modules; powerful object-orientated programming	Expensive, fitting of GLMMs is often slow
GenStat	VSN International Ltd (Oxford, UK) http://www.vsn-intl.com/genstat/	Nested ANOVA, LMM, GLMM	Widely used statistics package with plenty of support and add-in modules; fast and robust fitting of GLMMs	Expensive
R	http://www.r-project.org/	Nested ANOVA, LMM, GLMM	Free, uses most of the functions in S-plus	Command-driven and difficult to use
Stata	http://www.stata.com	Nested ANOVA, LMM	Relatively cheap and easy to use	No GLMM methods as yet
BUGS	MRC Biostatistics Unit (Cambridge, UK) http://www.mrc-bsu.cam.ac.uk/bugs/	Nested ANOVA, LMM, GLMM	Free, graphical construction of models; virtually unlimited model specification allowed	Very difficult to compare models and select a minimal model

[a]Abbreviations: ANOVA, analysis of variance; GLMM, generalized linear mixed model; LMM, linear mixed model.

villages was spatially structured, possibly owing to some, as-yet unidentified, environmental variable. As this example shows, the flexibility of GLMMs allows several hypotheses to be tested within one class of analysis.

Learning from your errors

It would be wrong to think that the only reason for using a mixed model is to overcome the nuisance of pseudoreplication. It is true that including an extra error term in the model allows unbiased estimates of the magnitude and significance of the fixed effects to be made. However, the random and error terms (i.e. the variation between repeated measures or between individuals) are intrinsically interesting themselves. An important application is in macroparasite infections, which often show a characteristic aggregated distribution [13,17–20]. Epidemiological models stress the potential consequences of this aggregation to stabilize the dynamics of host and parasite populations [18,21,22]. In this respect, understanding the causes of aggregation is crucial to link processes of infection at the individual level to the population-level consequences of these processes [20,23,24]. Mixed models might provide a useful technique to understand the relative contribution of different processes to the observed patterns of aggregation in field studies. For example, consider a population of hosts infected with macroparasites and showing an aggregated distribution of parasites. One can model the causes of this aggregation by analysing the rate at which parasites infect hosts. If all hosts pick up infections at the same rate, and if infection can be described by a stochastic, POISSON PROCESS, the distribution of adult parasites between hosts will follow a POISSON DISTRIBUTION (i.e. there will be no aggregation). However, where different hosts will pick up infections according to Poisson processes with different rates, the distribution of parasites across all hosts will be aggregated if the factors that underlie the infection rate combine multiplicatively [2,13,18,23]; for example, if infection rate is proportional to parasite fecundity × larval survival [18].

Since mixed models allow the variation between hosts to be modelled explicitly, the relative contribution of different factors to this variation can also be tested. For example, Elston *et al.* [25] used this approach to model the factors affecting the aggregation of ticks (*Ixodes ricinus*) in red grouse chicks (*Lagopus lagopus scoticus*). Using a GLMM with Poisson errors [15], they analysed the variation in tick infection intensity between individuals and between broods, and found that more than a half of the variation between randomly selected individuals could be explained using random terms corresponding to temporal (year) and spatial (altitude and location) effects.

Another area where the variation between hosts is of particular interest is in animal breeding. Here, one wishes to know how much of the variation between hosts is due to genetic effects (e.g. variation in resistance) and environmental effects (e.g. variation in exposure) [26]. Relatives, which share some proportion of their genes, will be correlated with each other for a particular trait, such as parasite resistance. Incorporating different error terms allows variation between hosts in their parasite burden to be partitioned into genetic and environmental components of variation [27]. For example, this approach has been used to analyse *Teladorsagia circumcincta* infection in lambs and shows that host genetics appears to have little detectable effect on the number of nematodes in the gut but a very strong effect on the size and fecundity of these worms [28–30]. In related work, these analyses also show a strong effect of host genetics on immunological measures associated with reduced worm length and fecundity, such as eosinophilia and IgA [31,32]. This approach highlights the potential for mixed models as a tool to quantify the genetic basis of different immunological pathways underlying resistance to parasite infection and to understand the linkages between these pathways.

Applying mixed models

It is crucial that the high quality of data generated by modern parasitological research is coupled with appropriate statistical analysis. Unfortunately, this is often not the case and statistical methods are used that either do not

make full use of available data or that are inappropriate. A literature survey of papers over a 12-month period in the *International Journal of Parasitology* (October 2001 to September 2002) shows the degree to which (in our view) the appropriate statistics were used: 10% of all papers published presented no statistics at all but in our view should have done so; 4% of papers used GLMs but a further 25% would have benefited from their use; 2% of papers used mixed models and a further 12% would have benefited from their use. (We note that these findings are likely to be general across the parasitology literature and that it is not our aim to single out one journal for particular opprobrium, indeed the journal works hard at providing guidance on how to improve statistical analysis in parasitology research publications [33].) Our brief analysis demonstrates a clear need for parasitologists to consider carefully whether the statistical techniques that they use make the best of their data.

How easy is it then for a parasitologist to apply mixed models to their data? Three points are relevant to answering this question. First, you need to know what you are modelling before you can think about how to model it [6]. In this, there is no substitute for a good biological knowledge of the system being modelled. For example, what are the factors within the host or the external environment that are likely to influence the observed patterns of parasite abundance or prevalence? From these questions, defined hypotheses can be made and an appropriate statistical model constructed to test these hypotheses. Second, understanding how to apply a mixed model requires an understanding of its structure and how it relates to the hypothesis being tested. Spending the effort to understand the basic algebra, such as outlined in Box 1, is probably unavoidable, but more complex topics, such as how the computer derives the parameters of this model, should not be required for most applications. However, one does need to be aware that parameter estimation of a large number of variables in complex models might vary according to the precise procedures used by different computer packages. Third, mixed models are finding their way into widely available computer packages (Table 1), making them more accessible to laboratory and field parasitologists and less a preserve of the dedicated statistician.

Given these points, applying mixed models remains challenging but is now an achievable task for a scientist trained in biology rather than one trained in statistics. Hopefully, the application of mixed models will continue to become easier as more biologists become familiar with their use and as computer packages become more user friendly.

Parasitologists have a good record in being among the first to use new techniques in genetics, immunology and biochemistry to help to explore the interaction between hosts and parasites. Techniques in statistics, such as mixed models, should be seen in this light and it is hoped that they will bring significant advances in understanding a wide range of micro- and macroparasite infections. Mixed models offer a particularly promising method of understanding the causes and consequences of variability between hosts in their susceptibility or exposure to infection. New statistical techniques that help to understand these processes of parasite infection should form an integral part of our attempts to understand and control the diseases that parasites cause.

References

1 Wilson, K. and Grenfell, B.T. (1997) Generalized linear modelling for parasitologists. *Parasitol. Today* 13, 33–38
2 Pacala, S.W. and Dobson, A.P. (1988) The relation between the number of parasites and host age: population causes and maximum likelihood estimation. *Parasitology* 96, 197–210
3 Paterson, S. (2001) The use of repeated measure linear modelling to analyze longitudinal data from experimental parasite infections. *J. Parasitol.* 87, 969–971
4 Wilson, K. *et al.* (1996) Analysis of aggregated parasite distributions: a comparison of methods. *Funct. Ecol.* 10, 592–601
5 Zinsstag, J. *et al.* (2000) Effect of strategic gastrointestinal nematode control on faecal egg count in traditional West African cattle. *Vet. Res.* 31, 259–266
6 Crawley, M.J. (1993) *GLIM for Ecologists*, Blackwell Scientific Publications
7 Lindsey, J.K. (1999) *Models for Repeated Measures*, Oxford University Press
8 Crowder, M.J. and Hand, D.J. (1990) *Analysis of Repeated Measures*, Chapman & Hall
9 Grafen, A. and Hails, R.S. (2002) *Modern Statistics for the Life Sciences*, Oxford University Press
10 Smith, J.A. *et al.* (1999) Heritable variation in resistance to gastrointestinal nematodes in an unmanaged mammal population. *Proc. R. Soc. Lond. B Biol. Sci.* 266, 1283–1290
11 Sokal, R.R. and Rohlf, F.J. (1995) *Biometry*, Freeman
12 Diggle, P. *et al.* (2002) Childhood malaria in the Gambia: a case-study in model-based geostatistics. *J. R. Stat. Soc.* 51, 493–506
13 Shaw, D.J. and Dobson, A.P. (1996) Patterns of macroparasitic abundance and aggregation in wildlife populations: a quantitative review. *Parasitology* 111, S111–S133
14 Hudson, P.J. and Dobson, A.P. (1995) Macroparasites: observed patterns. In *Ecology of Infectious Diseases in Natural Populations* (Grenfell, B.T. and Dobson, A.P., eds), pp. 144–176, Cambridge University Press
15 McCulloch, C.E. and Searle, S.R. (2001) *Generalized, Linear and Mixed Models*, Wiley
16 Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalised linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25
17 Crompton, D.W.T. *et al.* (1984) Investigating over-dispersion; *Moniformis* (Acanthocephala) and rats. *Parasitology* 88, 317–331
18 Anderson, R.M. and May, R.M. (1992) *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press
19 Bliss, C.I. and Fisher, R.A. (1953) Fitting the negative binomial distribution to biological data. *Biometrics* 9, 176–200
20 Wilson, K. *et al.* (2002) Heterogeneities in macroparasite infections: patterns and processes. In *The Ecology of Wildlife Diseases* (Hudson, P.J. *et al.*, eds), pp. 6–44, Oxford University Press
21 Anderson, R.M. and May, R.M. (1978) Regulation and stability of host–parasite population interactions; regulatory processes. *J. Anim. Ecol.* 47, 219–247
22 May, R.M. and Anderson, R.M. (1978) Regulation and stability of host–parasite population interactions; destabilising processes. *J. Anim. Ecol.* 47, 249–267
23 Grenfell, B.T. *et al.* (1995) Modelling patterns of parasite aggregation in natural populations: trichostrongylid nematode–ruminant interactions as a case study. *Parasitology* 111, S135–S151
24 Smith, G. *et al.* (1995) Macroparasite group report. In *Ecology of Infectious Diseases in Natural Populations* (Grenfell, B.T. and Dobson, A.P., eds), pp. 209–229, Cambridge University Press
25 Elston, D.A. *et al.* (2001) Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology* 122, 563–569
26 Falconer, D.S. (1989) *Introduction to Quantitative Genetics*, Longman Scientific and Technical
27 Tempelman, R.J. (1998) Generalized linear mixed models in dairy cattle breeding. *J. Dairy Sci.* 81, 1428–1444

28 Stear, M.J. *et al.* (1997) The genetic basis of resistance to *Ostertagia circumcincta* in lambs. *Vet. J.* 154, 111–119
29 Stear, M.J. *et al.* (1995) Regulation of egg-production, worm burden, worm length and worm fecundity by host responses in sheep infected with *Ostertagia circumcincta*. *Parasite Immunol.* 17, 643–652
30 Stear, M.J. *et al.* (1998) The processes influencing the distribution of parasitic nematodes among naturally infected lambs. *Parasitology* 117, 165–171
31 Strain, S.A.J. *et al.* (2002) The genetic control of IgA activity against *Teladorsagia circumcincta* and its association with parasite resistance in naturally infected sheep. *Parasitology* 124, 545–552
32 Stear, M.J. *et al.* (2002) Eosinophilia as a marker of resistance to *Teladorsagia circumcincta* in Scottish Blackface lambs. *Parasitology* 124, 553–560
33 Morrison, D.A. (2002) How to improve statistical analysis in parasitology research publications. *Int. J. Parasitol.* 32, 1065–1070
34 Gamerman, D. (1997) *Markov Chain Monte Carlo*, Chapman & Hall
35 Lagarias, J.C. *et al.* (1998) Convergence properties of the nelder-mead simplex algorithm in low dimensions. *SIAM J. Optimization* 9, 112–147
36 McCullagh, P. and Nelder, J.A. (1989) *Generalised Linear Models*, Chapman & Hall