

Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression

DAVID FLETCHER,^{1,2,*} DARRYL MACKENZIE² and
EDUARDO VILLOUTA³

¹*Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand*

E-mail: dfletcher@maths.otago.ac.nz


²*Proteus Wildlife Research Consultants, P.O. Box 5193, Dunedin, New Zealand*

³*Department of Conservation, Wellington, New Zealand*

Received July 2003; Revised September 2004

We discuss a method for analyzing data that are positively skewed and contain a substantial proportion of zeros. Such data commonly arise in ecological applications, when the focus is on the abundance of a species. The form of the distribution is then due to the patchy nature of the environment and/or the inherent heterogeneity of the species. The method can be used whenever we wish to model the data as a response variable in terms of one or more explanatory variables. The analysis consists of three stages. The first involves creating two sets of data from the original: one shows whether or not the species is present; the other indicates the logarithm of the abundance when it is present. These are referred to as the ‘presence data’ and the ‘log-abundance’ data, respectively. The second stage involves modelling the presence data using logistic regression, and separately modelling the log-abundance data using ordinary regression. Finally, the third stage involves combining the two models in order to estimate the expected abundance for a specific set of values of the explanatory variables. A common approach to analyzing this sort of data is to use a $\ln(y+c)$ transformation, where c is some constant (usually one). The method we use here avoids the need for an arbitrary choice of the value of c , and allows the modelling to be carried out in a natural and straightforward manner, using well-known regression techniques. The approach we put forward is not original, having been used in both conservation biology and fisheries. Our objectives in this paper are to (a) promote the application of this approach in a wide range of settings and (b) suggest that parametric bootstrapping be used to provide confidence limits for the estimate of expected abundance.

Keywords: abundance, bootstrap, conditional model, evechinus, ecklonia

1352-8505 © 2005  Springer Science+Business Media, Inc.

*Corresponding author

1352-8505 © 2005  Springer Science+Business Media, Inc.

Introduction

In many ecological research studies, abundance data often exhibit two features: a substantial proportion of the values are zero, and the remainder has a skewed distribution. Both these attributes reflect the patchiness of the environment and/or the inherent heterogeneity of the species concerned. Suppose we wish to model the abundances in terms of one or more covariates. A common approach would be to use a general linear model, in conjunction with a $\ln(y + c)$ transformation, where y is the response and c is some constant (usually $c = 1$). The aim of this transformation is to better satisfy the assumption that the errors are normal and have constant variance. An obvious disadvantage of this approach is that the choice of c is arbitrary and yet may influence the results of the analysis. Use of a square-root transformation avoids this problem, but may not always lead to normality of the errors. For both types of transformation, the presence of a substantial proportion of zero values will often make the assumption of constant error variance invalid.

A number of alternative approaches have been suggested for the analysis of this kind of data:

1. Fit a generalized linear model, in which the response is modelled as a random variable with a Poisson or negative binomial distribution. Both of these approaches suffer from the handicap that the proportion of zero values must necessarily be linked to the distribution of the positive values, often leading to a poor fit to ecological data (Welsh *et al.*, 1996).
2. Modify the approach in (1) by assuming that the response has a *mixture* distribution. With probability p it is equal to zero, and with probability $1-p$ it has a Poisson or negative binomial distribution (Lambert, 1992).
3. Separately model (a) the occurrence of a zero value (as a Bernoulli random variable) and (b) the positive abundances. This has two major advantages. First, we can model these two aspects of the data separately, and gain insight into whether they are being influenced by the covariates in different ways. Second, the analysis is simpler than with the mixture model approach as the parameters for the two models can be estimated and interpreted independently (see Welsh *et al.*, 1996 for details).

The last of these approaches has been used in both ecology and fisheries. An early reference is Lachenbruch (1976), who focussed on the simple case of comparing two groups, and considered exponential, lognormal and truncated Poisson distributions for the positive data. In an ecological setting, Welsh *et al.* (1996) called this a *conditional* model, and suggested using a truncated Poisson or negative binomial distribution for the positive abundances. Dobbie and Welsh (2001) extended the model to deal with serial dependence in repeated measurements. The general idea behind this approach is mentioned by Manly, McDonald and Thomas (1993, Section 10.4), in the context of modelling the amount of resource used by a population of animals. In a fisheries context, Stefansson (1996) suggested using either a gamma or a lognormal distribution for the positive values. Both Welsh *et al.* (1996) and Stefansson (1996) give expressions for the likelihood associated with this type of

model, and show that it contains two distinct components corresponding to the two models being fitted. Welsh *et al.* (1996) also pointed out that the covariance matrix for the full set of parameters can be obtained from the covariance matrices obtained by fitting the two component-models. Lo, Jacobson and Squire (1992) used a lognormal distribution for the positive values, but used only a normal approximation to model the proportion of zero values. In meteorology, Coe and Stern (1982) used this type of approach to model rainfall data, and chose a gamma distribution to model amount of rainfall. In principle, we can choose any model for the positive abundances that we think is appropriate. Note that a related topic of current interest in ecology is the relationship between occupancy and abundance (Gaston *et al.*, 2000).

The purpose of this paper is to present an example of the use of a special case of the conditional model, in which the model of the positive abundances is assumed to have errors that are lognormally distributed. The advantage of this special case is that we can use well-established techniques, ordinary and logistic regression, to provide estimates of the parameters. We also suggest the use of parametric bootstrapping to provide confidence intervals, rather than analytical expressions (c.f. Stefansson, 1996).

Our motivation for this work came from a study carried out by one of us (EV) into the potential ecological impacts of a sea urchin (*Evechinus chloroticus*) fishery in Fiordland, New Zealand. The data from this study are summarized in the next section.

2. Data

Data were collected in Dusky Sound (45° 45' S; 166° 35' E), in Fiordland, New Zealand, as part of a study by the Ministry of Fisheries and Agriculture to assess the relationship, if any, between algal abundance and that of the sea urchin *Evechinus chloroticus* (hereafter *Evechinus*: McShane *et al.*, 1993). Sites were 100-m sections of coastline, randomly selected from within the study area on each of four sampling dates between 1993 and 1995. Thirty-two of the sites were visited on more than one date: for the simplicity of presentation we have omitted these from our analysis, giving a total of 103 sites, each visited once. At each site, the research vessel was stopped at an offshore-point suitable for diving. Two divers were sent down from the research vessel, to a randomly chosen depth between 6.5 and 12 m. Densities of *Evechinus* and the seaweed *Ecklonia radiata* (hereafter *Ecklonia*) were then measured using a 25 m × 1 m quadrat, formed by 'rolling' a 1 m² quadrat 25 times in a randomly-chosen direction (McShane *et al.*, 1993). The data are available from the first author (DF).

The sites chosen differed in their exposure to water motion. We assumed that in the study area water movement is a function of oceanic swell, local-generated waves, and tidal currents. Difficulties in making direct measurements of these three factors led to a search for one or more related indices for each site. The *distance to the mouth* of the fiord entrance was chosen as an index for exposure to swell. Broken and hilly topography results in complex local-generated wind patterns. Due to the difficulties that would be involved in measuring the direction and force of the wind at each site

during a representative period of time, we developed an index of potential *fetch*. This was defined as the sum of a set of radial distances as follows. Radii were drawn at 10 degree intervals, and the distance to the point of first intersection with land was measured. For open stretches of water, we arbitrarily set the radial distance to be 10 km. Tidal currents were assumed to be stronger when passing through narrow passages than in open water. We therefore used the *distance across* to the nearest land as an index of tidal current speed.

The subset of the data that we consider here consists of six measurements for each site. The response variable is the mean density of *Ecklonia* (individual plants/m²). The five explanatory variables are:

<i>K</i>	Logarithm of mean abundance of <i>Evechinus</i> (individuals/m ²)
<i>D</i>	Date
<i>M</i>	Distance to mouth (km)
<i>A</i>	Distance across (km)
<i>F</i>	Fetch (km)

We used a $\ln(x+1)$ transformation for *K*, in order to allow for zero mean abundances of *Evechinus*. As *K* is not a response variable, the arbitrary nature of the choice of constant is not crucial.

The objectives of the analysis were to predict the mean density of *Ecklonia* that would be observed with the explanatory variables set at specific values. In particular, we were interested in predicting the density of *Ecklonia* that would result from a change in the density of *Evechinus* caused by introducing a fishery. Assessing the importance of physical factors (oceanic swell, local waves and tidal currents) on the relationship between *Ecklonia* and *Evechinus* has important management implications. If the effect of *Evechinus* harvest on *Ecklonia* varies according to the level of these physical factors, it is harder to achieve a uniformly small impact throughout the fiord. We might then need different harvesting limits in different parts of the fiord, in order to achieve this objective.

3. Analysis

We first created two datasets: one indicating whether *Ecklonia* was present or not at each site, the other showing the log-transformed abundance for those sites where *Ecklonia* was present. These two data sets are referred to here as the ‘presence data’ and the ‘log-abundance data’, respectively. Note that the log-abundance data contained fewer observations than the presence data, as it excluded those sites where *Ecklonia* was absent.

We then modelled both the presence data and the log-abundance data in terms of the explanatory variables, using logistic and ordinary regression, respectively. Date was regarded as a fixed blocking factor. We therefore considered models containing date (*D*) plus the main effects and all possible interactions between the other four explanatory variables (*K*, *M*, *A* and *F*). For both types of model, we used a stepwise selection procedure, with *D* always being kept in the model. Details of the selection procedures are not given here, as they are not relevant to the approach. We checked model

adequacy for the ordinary regression by inspection of the residuals. These indicated no obvious problems with the model. For the logistic regression, a reliable lack-of-fit test was not possible, as continuous covariates were present (Hosmer *et al.*, 1997).

The final models for the presence and log-abundance data were combined to predict the expected density of *Ecklonia* as follows. Let $Y(k, d, m, a, f)$ be the density of *Ecklonia* when $K = k, D = d, M = m, A = a$ and $F = f$. Also, let $Z(k, d, m, a, f)$ be a binary variable, equal to one when *Ecklonia* is present and zero otherwise. The expected value of Y is given by:

$$\begin{aligned} E(Y) &= \Pr(Z = 1)E(Y|Z = 1) + \Pr(Z = 0)E(Y|Z = 0), \\ &= \Pr(Z = 1)E(Y|Z = 1), \\ &= \pi\mu, \end{aligned}$$

where $\pi = \Pr(Z = 1)$ and $\mu = E(Y|Z = 1)$. A natural estimate of the expected density of *Ecklonia* is given by (Stefansson, 1996; Welsh *et al.*, 1996):

$$\hat{E}(Y) = \hat{\pi}\hat{\mu}, \quad (1)$$

where

$$\hat{\pi} = \exp(\mathbf{x}'\hat{\beta}) / \{1 + \exp(\mathbf{x}'\hat{\beta})\} \quad (2)$$

and

$$\hat{\mu} = \exp(\mathbf{w}'\hat{\theta} + \hat{\sigma}^2/2) \quad (3)$$

are the estimates of π and μ obtained from the two regression models. Thus, $\hat{\beta}$ is the vector of estimates of the coefficients in the logistic regression model for the presence data, and x is the corresponding vector of explanatory variables. Similarly, $\hat{\theta}$ is the vector of estimates, w the vector of explanatory variables, and $\hat{\sigma}^2$ the residual mean square in the regression model for the log-abundance data (Crow and Shimizu, 1988).

A confidence interval for the estimate in equation (1) was obtained using parametric bootstrapping (Davison and Hinkley, 1997). This involved randomly generating alternative values of $\hat{E}(Y)$ by resampling β , θ and σ^2 and then using equations (2) and (3). In the resampling, it is simpler and more direct to resample $x'\beta$ and $w'\theta$ than β and θ . Thus, $x'\beta$ was selected from a normal distribution with mean $x'\hat{\beta}$ and variance $x'\hat{\Sigma}_{\hat{\beta}}x$, where $\hat{\Sigma}_{\hat{\beta}}$ is the covariance matrix for $\hat{\beta}$, obtained from the logistic regression analysis. Likewise, $w'\theta$ was selected from a normal distribution with mean $w'\hat{\theta}$ and variance $w'\hat{\Sigma}_{\hat{\theta}}w$, where $\hat{\Sigma}_{\hat{\theta}}$ is the covariance matrix for $\hat{\theta}$, obtained from the ordinary regression analysis. The values of $\hat{\sigma}^2$ were generated by selecting $k(\sigma^2/\hat{\sigma}^2)$ from a χ^2 distribution with k degrees of freedom, where k is the number of residual degrees of freedom in the ordinary regression analysis. Note that the assumption of normality for the errors in the ordinary regression model implies independence of $\hat{\theta}$ and $\hat{\sigma}^2$, allowing them to be selected independently in the bootstrap sample. There are a number of options for using the resulting bootstrap sample of values of $\hat{E}(Y)$ to produce a confidence interval (Davison and Hinkley, 1997; Manly, 1997): we chose to take the 2.5th and 97.5th percentiles from 9999 bootstrap samples.

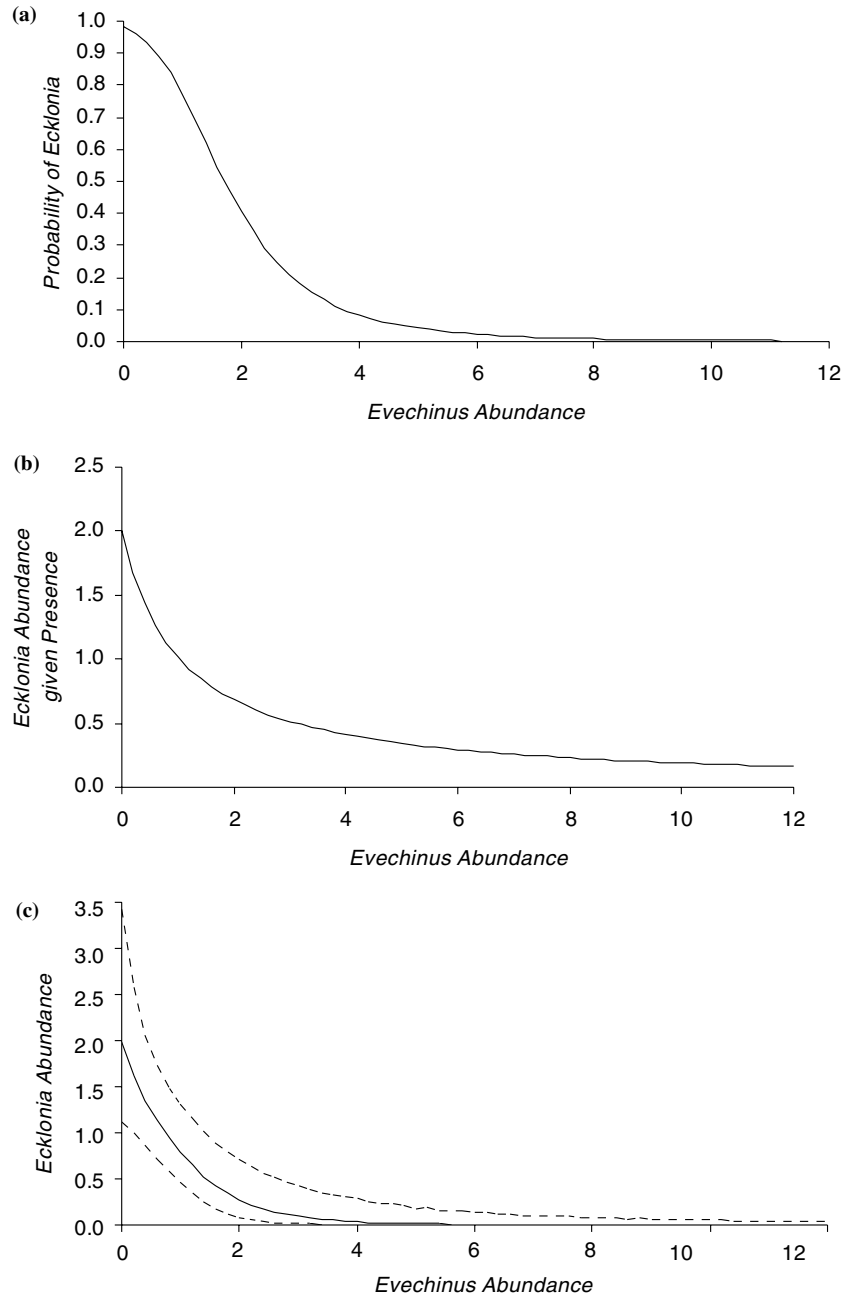


Figure 1. Estimates of (a) probability of presence, (b) expected abundance given presence and (c) expected abundance of *Ecklonia* (dashed lines are 95% confidence limits), plotted against abundance of *Evechinus*. The predictions are for an average site (M, A and F set to 9.22 km, 0.72 km and 32.68 km, respectively). Abundance is measured in individuals per m².

4. Results

Table 1 shows the parameter estimates for the logistic regression and ordinary regression models for the explanatory variables of interest. Note that for computation purposes, the variables M, A and F were ‘studentized’ before analysis by subtracting the corresponding mean and then dividing by the standard deviation. For the presence data, the selected logistic regression model included a 3-way interaction between A, F and K. For the log-abundance data, the selected model contained a two-way interaction between M and K.

Figure 1 shows how the results from the two models are combined to give $\hat{E}(Y)$, the expected *Ecklonia* abundance at an average site, together with bootstrap-based confidence limits. This average site has the values of M, A and F set to 9.22 km, 0.72 km and 32.68 km, respectively, the means observed in the study.

There is clear evidence that increasing *Evechinus* density is associated with a decrease in both the presence of *Ecklonia* and its abundance given that it is present. There is also evidence that these relationships depend on A, F and M. As *distance across* increases or as *fetch* decreases, a change in *Evechinus* density is associated with a larger decrease in the probability of *Ecklonia* being present. As *distance to mouth* increases, a change in *Evechinus* density is associated with a larger decrease in the abundance of *Ecklonia* (given that it is present). Using the estimates and standard errors given in Table 1, we can quantify the overall relationship between *Evechinus* and *Ecklonia* for an average site (Figure 1) as follows. If *Evechinus* density doubles, the odds of *Ecklonia* being present are reduced by 94% (95% CI: 67–99%). Second, when *Ecklonia* is present, the same doubling of *Evechinus* density leads to *Ecklonia* abundance being reduced by 50% (95% CI: 6–73%). Note that since a $\ln(x+1)$

Table 1. Estimates and standard errors of the coefficients for the explanatory variables of interest in the final logistic and ordinary regression models, chosen using stepwise selection. K = Logarithm of (mean abundance + 1) of *Evechinus*; M = Distance to mouth; A = Distance across; F = Fetch; $\hat{\sigma}^2$ = residual mean square in the regression for the log-abundance data. Estimates of the coefficients for the blocking factor *Date* are not given, as they are not of direct interest in the analysis.

	Logistic regression		Ordinary regression	
	Estimate	SE	Estimate	SE
Constant	3.7669	0.9452	0.0183	0.2715
K	-3.9669	1.1772	-0.9860	0.4475
M	–	–	0.9125	0.2378
A	1.1430	1.1819	-0.2080	0.3022
F	-0.5105	0.8422	-0.3191	0.2154
MK	–	–	-1.5462	0.3955
AK	-2.3432	1.6391	–	–
FK	1.9401	1.1256	–	–
AF	-2.5735	1.1143	0.3407	0.1297
AFK	3.3222	1.4912	–	–
$\hat{\sigma}^2$	–	–	1.2977 (69 d.f.)	–

transformation was used for K, a doubling of *Evechinus* density here strictly means a doubling of (*Evechinus* density + 1).

Using the bootstrap samples that were generated to calculate 95% CI, we estimated the bias in expected abundance of *Ecklonia* for the range of *Evechinus* abundances shown in Figure 1. The maximum bias occurred when *Evechinus* abundance was zero, and was less than 0.07 individuals per m².

5. Discussion

The analysis we have presented has allowed us to assess the effect of a number of biological and physical factors on the abundance of a patchily-distributed organism. By separately modelling the probability of presence and the abundance given presence, we have learned more about the system than we would have using a single model for abundance. In addition we have been able to simply combine the results from the two analyses to provide predictions that will be useful for management.

We suggest that the approach we have used can be applied in a wide range of settings. The idea of using a conditional model for positively skewed data that contain a large proportion of zeros is not new. It has been used in both an ecological context (Welsh *et al.*, 1996) and a marine setting (Lo, Jacobson and Squire, 1992; Stefansson, 1992). It is also akin to the methods put forward by Coe and Stern (1982) in meteorology, and by Lambert (1992) in manufacturing process control. Our aim has been to present this approach as one which could be used widely in the biological sciences. If the presence data and the log-abundance data can be adequately modelled using logistic and ordinary regression respectively, the estimation part of the analysis simply involves separate application of these two well-known methods. Indeed, in this case, the use of the conditional model can be seen as a way of extending standard general linear model analyses to better cope with this type of data.

There are three issues to mention regarding use of the methods we have put forward here. First, the log-abundance data may be highly unbalanced, compared to the original dataset. This will mean that more care is needed in the resulting analysis, but such an analysis may be more reliable than one based on a $\ln(y + c)$ transformation. Second, if the predicted probability of presence approaches zero or one for values of the explanatory variables that are of interest, the use of parametric bootstrapping may not provide a reliable estimate of precision (Efron and Tibshirani, 1993). Third, if the sample size is small, the logistic regression may not provide sufficient power to detect effects that would be of interest. A related issue is that we may prefer to use a non-parametric bootstrap procedure if we are unsure of the validity of the assumptions inherent in the parametric bootstrap, i.e. that both $x'\beta$ and $w'\hat{\theta}$ are normal, and that $k(\sigma^2/\hat{\sigma}^2)$ has a χ^2 distribution. Given these caveats, we feel that the approach we have put forward is well worth considering when analyzing skewed data with a large proportion of zeros.

It is worth noting here the similarities and differences between the conditional model and the mixture-model approach (Lambert, 1992). A key point is that the mixture-model approach allows zeros to be part of either component. This may lead to a better fit to the data when some of the zeros arise as a consequence of mea-

surement error. The associated disadvantage is that both estimation and interpretation are not as straightforward as for the conditional model. Both approaches allow different sets of explanatory variables to be used to model the two components, thereby leading to a better understanding of the system under study.

We have advocated use of the parametric bootstrap to provide confidence intervals for the expected response. We would expect this approach to be more reliable than one based on large-sample analytical results. Such results are often expressed in terms of an approximate standard error. Any confidence interval based on this would be symmetrical, and thereby deny the skewness inherent in the data.

In the context of marine abundance surveys, Pennington (1983) suggested use of the so-called Δ -distribution, which was discussed by Aitchison and Brown (1957). This distribution is the same as the one we have assumed in our example. Myers and Pepin (1990) has argued against uncritical use of this distribution, in situations where the positive abundances are not lognormally distributed. Clearly the use of the conditional model does not do away with the need to check model assumptions. With the logistic model for the probability of a zero there is the possibility of overdispersion (McCullagh and Nelder, 2000). With the model for the positive values, it can be useful to check the relationship between the mean and variance, in order to distinguish between a lognormal, gamma (Stefansson, 1992), truncated negative-binomial/Poisson (Welsh et al., 1996) or Ades (Perry and Taylor, 1985) distribution.

Acknowledgments

We are grateful to two anonymous referees, whose comments helped improve the manuscript, and to Guy Pardon for his suggestions.

References

- Aitchison, J. and Brown, J.A.C. (1957) *The Lognormal Distribution*, Cambridge University Press, Cambridge, UK.
- Coe, R. and Stern, R.D. (1982) Fitting models to daily rainfall data. *Journal of Applied Meteorology*, **21**, 1024–1031.
- Crow, E.L. and Shimizu, K. (eds.) (1988) *Lognormal Distributions: Theory and Applications*, Dekker, New York, USA.
- Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK.
- Dobbie, M.J. and Welsh, A.H. (2001) Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics*, **43**, 431–444.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- Gaston, K.J., Blackburn, T.M., Greenwood, J.D., Gregory, R.D., Quinn, R.M., and Lawton, J.H. (2000) Abundance-occupancy relationships. *Journal of Applied Ecology*, **37**(Suppl. 1), 39–59.
- Hosmer, D.W., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997) A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, **16**, 965–980.

- Lachenbruch, P.A. (1976) Analysis of data with clumping at zero. *Biometrical Journal*, **18**, 351–356.
- Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.
- Lo, N.C.H., Jacobson, L.D., and Squire, J.L. (1992) Indices of relative abundance from fish spotter data based on delta-lognormal models. *Canadian Journal of Fisheries and Aquatic Science*, **49**, 2515–2526.
- Manly, B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman and Hall, London, UK.
- Manly, B.F.J., McDonald, L.L., and Thomas, D.L. (1993) *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*, Chapman and Hall, London, UK.
- McCullagh, P. and Nelder, J.A. (2000) *Generalized Linear Models*, Chapman and Hall, London, UK (2nd Edition).
- McShane, P.E., Naylor, J.R., Anderson, O., Gerring, P., and Stewart, R. (1993) Pre-fishing surveys of kina (*Evechinus chloroticus*) in Dusky Sound, Southwest New Zealand. New Zealand Fisheries Assessment Research Document 93/11.
- Myers, R.A. and Pepin, P. (1990) The robustness of lognormal-based estimators of abundance. *Biometrics*, **46**, 1185–1192.
- Pennington, M. (1983) Efficient estimators of abundance, for fish and plankton surveys. *Biometrics*, **39**, 281–286.
- Perry, J.N. and Taylor, L.R. (1985) Ades: new ecological families of species-specific frequency distributions that describe repeated spatial samples with an intrinsic power-law variance-mean property. *Journal of Animal Ecology*, **54**, 931–953.
- Stefansson, G. (1996) Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*, **53**, 577–588.
- Welsh, A.H., Cunningham, R.B., Donnelly, C.F., and Lindenmayer, D.B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

Biographical sketches

David Fletcher is a Senior Lecturer in the Department of Mathematics and Statistics at the University of Otago, Dunedin, New Zealand. He is also a director of Proteus Wildlife Research Consultants, which specialises in statistical ecology. He has collaborated extensively with zoologists, both at Otago University and in the New Zealand Department of Conservation.

Darryl MacKenzie is a biometrician for and director of Proteus Wildlife Research Consultants. While studying at the University of Otago, New Zealand, Darryl became interested in ecological statistics, particularly in the application of computer intensive methods and mark-recapture studies. He has collaborated extensively with scientists at the Patuxent Wildlife Research Center, most recently on methods for estimating occupancy probabilities from presence/absence data.

Eduardo Villouta is a Marine Ecologist in the Science & Research Unit at the Department of Conservation, Wellington, New Zealand. He is involved in research on the population ecology of introduced macroalgae, recruitment and movement of fish, and the design of marine reserves.